CHAPTER 19

Evolution of Glycan Diversity

Ajit Varki, Hudson H. Freeze, and Pascal Gagneux

THIS CHAPTER PROVIDES A BRIEF COMPARATIVE OVERVIEW of the patterns of glycosylation in various taxa of living organisms and discusses the complexity and diversity of these glycans from an evolutionary perspective. Because much of the currently available information concerns vertebrates, this chapter focuses on comparisons between the glycans of vertebrates and those of other taxa. The evolutionary processes that likely determine the generation of glycan diversity are briefly considered, including intrinsic host glycan-binding protein functions and interactions of hosts with extrinsic pathogens or symbionts.

RELATIVELY LITTLE IS KNOWN ABOUT GLYCAN DIVERSITY IN NATURE

The genetic code is essentially the same in all known living organisms, and several core functions such as gene transcription and energy generation tend to be conserved across various taxa. Although glycans are also found in all organisms, considerable diversity of their struc-

ture and expression exists in nature, both within and between evolutionary lineages. Partly because of the inherent difficulties in studying their structures, relatively little is known about the details of this glycan diversity, and there are few comprehensive data sets on this subject. For many taxa, essentially no information is available on their glycan profiles. Sufficient data are available, however, to indicate that there is no universal "glycan structure code" akin to the genetic code. Indeed, the glycans expressed by most free-living Eubacteria and Archaea (formerly grouped as prokaryotes) (see Chapter 20) have relatively little in common structurally with those of eukaryotes (an exception occurs when bacterial pathogens mimic eukaryotic host structures). On the other hand, most major glycan classes identified in animal cells seem to be represented in some related form among other eukaryotes, and sometimes in Archaea. Figure 19.1 outlines the major branches of the eukaryotic tree of life, with an emphasis on the protostome-deuterostome split and the phylogeny of "model" organisms, for which whole-

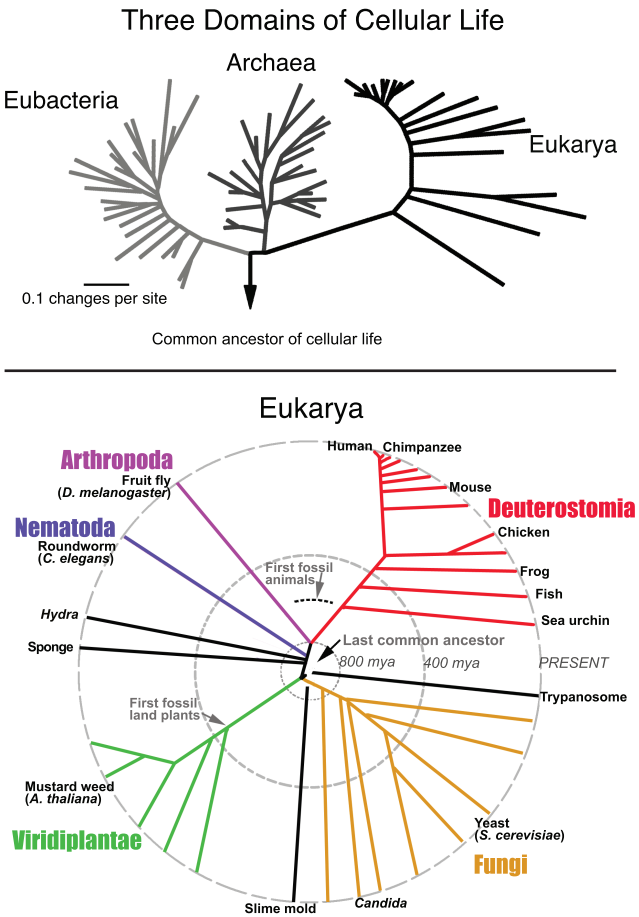

FIGURE 19.1. Phylogenetic trees of the three domains of cellular life (*upper panel*) and of the multicellular Eukarya (*lower panel*). The universal tree of life (*upper panel*) is inferred from maximum likelihood analysis of 1620 homologous nucleotide positions of small-subunit ribosomal RNA sequences from each organism. (The tree is redrawn, with permission, from Barns S.M. et al. 1996. *Proc. Natl. Acad. Sci.* **93:** 9188–9193, © National Academy of Sciences, U.S.A. The eukaryotic phylogeny is redrawn and modified, with permission, from Pollard T.D. et al. 2007. *Cell Biology*, 2nd Edition. Saunders, New York, © Elsevier.) Common eukaryotic "model" organisms are indicated. Except for the sponge, all indicated species have had their genomes sequenced. (*Gray dotted rings*) Approximate time before present (mya = millions of years ago). Major groups are indicated by different colors and refer to specific chapters (see text for discussion). The unicellular alveolates (e.g., trypanosomes) and slime mold diverged more than 1 billion years ago. Thus, their branching points are not shown.

genome sequence data have been generated. In contrast, far fewer organisms have been the subject of in-depth glycan structural analyses. The high levels of diversity encountered in the best-studied vertebrate species are a predictor of similar diversity in other groups of organisms. The existing information on the distribution of glycan types points to complicated patterns. On the one hand, glycan patterns can form "trends" and characterize entire phylogenetic lineages where one encounters further biochemical variation with subsets unique to certain sublineages. On the other hand, many glycans show rather discontinuous distribution across the tree of life and distantly related organisms can express surprisingly similar glycans.

EVOLUTIONARY VARIATIONS IN GLYCANS

O-Glycans

Homologs of the UDP-GalNAc:polypeptide *N*-acetylgalactosaminyltransferases (ppGalNAcTs) that initiate synthesis of the most common O-glycan class in vertebrates have been found throughout the animal kingdom (see Chapter 9). Multiple isoforms of ppGalNAcTs exist in most species. The common core-1 Galβ1-3GalNAcα1-O-Ser/Thr structure of vertebrates is present in insects, where it also forms part of a mucin-like protective layer in the gut. In contrast, plants do not appear to have O-linked GalNAc. Instead, they express arabinose O-linked to hydroxyproline and galactose O-linked to serine and threonine (see Chapter 22). Far less is known about bacterial O-glycosylation, although it is clear that novel O-glycans can be found within bacterial "S" layers, for example, a Galβ1-O-Tyr core (see Chapter 20).

Glycosphingolipids

Glucosylceramide is found in both plants and animals (see Chapter 10). However, the commonest core structure of vertebrate glycosphingolipids (Galβ1-4Glc-Cer) is varied in other organisms, for example, Manβ1-4Glc-Cer and GlcNAcβ1-4Glc-Cer in certain invertebrates. Other variations are inositol-1-O-phosphorylceramide, for example, mannosyldiinositolphosphorylceramide, which is the most abundant sphingolipid of yeast, and GlcNAcα1-4GlcAα1-2-*myo*-inositol-1-O-phosphorylceramide, which is found in tobacco leaves. Galactosylceramide and its derivatives seem to be limited to the nervous system of the deuterostome lineage of "higher" animals (see Figure 19.1). In contrast, all protostome nerves contain mainly glucocerebrosides. An evolutionary trend is suggested: A transition from gluco- to galactocerebrosides corresponds with changes in the nervous system from loosely structured to highly structured myelin. With regard to the complex gangliosides of the deuterostome nervous system, some general trends are seen in comparing reptiles to fish to mammals: an increase in sialic acid content, a decrease in the complexity of ganglioside composition, and a decrease in "alkali-labile" molecules (bearing O-acetylated sialic acids). A general rule has also been suggested: the lower the environmental temperature, the more polar the composition of brain gangliosides. Thus, poikilothermic (cold-blooded) animals tend to express many polysialylated gangliosides in the brain.

N-Glycans

Perhaps the broadest base of evolutionary information concerns asparagine–N-linked glycans (see Chapter 8). All plants and animals studied to date seem to share the same early stages of the classic N-glycan processing pathway (see Chapter 8), including the generation and transfer of $Glc_3Man_9GlcNAc_2$ from a dolichol-linked precursor to asparagine residues on newly synthesized proteins. Such an extreme degree of conservation is understandable,

given the critical role of this glycan structure in modulating the folding and maturation of newly synthesized glycoproteins in the endoplasmic reticulum (ER) (see Chapter 36). However, some parasitic protists can transfer truncated forms of otherwise similar lipid-linked oligosaccharides, sometimes even just the core $GlcNAc_2$ sequence. The intracellular localization of these structures also makes them less likely to be involved in rapid evolutionary arms races due to exploitation by parasites and pathogens (see below). The trimming and extension steps that occur thereafter along the vertebrate N-glycan processing pathway are recapitulated to varying extents in other eukaryotic taxa (Figure 19.2). Yeasts and vegetative slime molds do not appear to complete the trimming of mannose residues, and are thus unable to generate typical "complex-type" N-glycans. Yeast often further specialize their high-mannose glycans by extending them into large mannans. In contrast, developing slime molds trim down the high-mannose forms to some extent but then do not extend them. In insects, mannose trimming appears to be generally completed, as in mammals, down to a $Man_3GlcNAc_2$ structure. The subsequent addition of GlcNAc residues is frequently followed by removal of these residues by a very active β-hexosaminidase (see Chapter 24). Thus, the final structure found in insects often has only the three core mannose residues (Figure 19.2). Prior to the removal of the GlcNAc residues in insect cells, an α1-3-linked fucose unit is often added to the core GlcNAc residue (frequently in addition to the α1-6-linked fucose typically found on the core GlcNAc of vertebrate N-glycans). Plants follow a pathway similar to that in vertebrates in the initial stages, but then often add a bisecting β1-2-linked xylose residue on the β-linked mannose residue (see Chapter 22 and Figure 19.2). The latter structure is also present in some invertebrates, but it appears to be immunogenic in vertebrates. In keeping with the above findings, the early-processing α-mannosidases of the N-glycan pathway have a wide evolutionary distribution. In contrast, the endo-α-D-mannosidase processing enzyme that provides an "alternate deglucosylating pathway" for N-glycans (see Chapter 8) appears to be limited to members of the

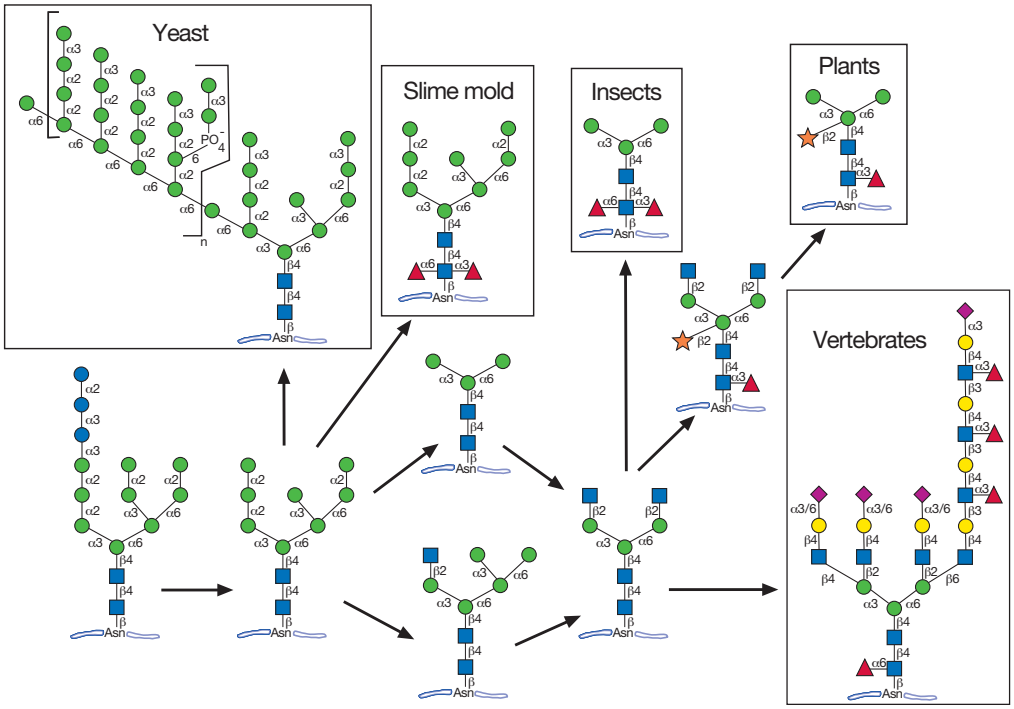

FIGURE 19.2. Dominant pathways of N-glycan processing among different taxa. See text for discussion.

chordate phylum, with the exception of the *Mollusca*, where it was detected in three distinct classes. The absence of this enzyme in other invertebrates examined, as well as in yeasts, various protozoa, and higher plants, suggests that the need for an alternate deglucosylation route paralleled the development of complex N-glycans in higher animals.

Overall, it is clear that the N-glycan pathway is evolutionarily ancient and found throughout the eukaryotes. However, limited information exists regarding the presence and distribution of most N-glycan pathway enzymes and genes in most animals and plants. Thus, there is still insufficient information to paint a clear picture of exactly how it has been diversified and specialized during evolution. It was once thought that only eukaryotes express N-glycans. However, it is now clear that Archaea express GalNAc-Asn and Glc-Asn linkages, and recent data indicate that a few bacteria such as *Campylobacter jejuni* express a very similar N-glycan pathway, but with a novel linkage unit, involving bacillosamine (2,4-diacetamido-2,4,6-trideoxy-D-glucose or $BacAc_2$) linked to asparagine.

Shared Outer Chains of Glycans

Outer terminal sequences are often shared among N- and O-glycans and glycosphingolipids (Chapter 13). The common outer-chain Galβ1-4GlcNAcβ1- (*N*-acetyllactosamine or "LacNAc") structure of vertebrates (see Chapter 13) is also found in plants. Some plants even add outer-chain Fucα1-3 residues to the GlcNAc residues of LacNAc units, generating $Lewis^x$-like structures identical to those found in animal cells. In some taxa, such as mollusks, an outer GalNAcβ1-4GlcNAcβ1- structure (the so-called LacDiNAc or LDN unit) tends to dominate, in place of the typical LacNAc structure more commonly seen in vertebrates. The SO_4-4-GalNAcβ1-4GlcNAcβ1- terminal units of pituitary glycoprotein hormones (see Chapter 13) have been conserved throughout vertebrate evolution, suggesting that they are critical for biological activity. Controversy exists as to whether or not further extensions and terminations of N-glycan antennae typical of vertebrates occur in insect or plant cells (see discussion regarding sialic acids below). The genetic elimination of the common vertebrate terminal sequence Galα1-3Galβ1-4GlcNAcβ1- in Old World primates and variations in sialic acids are discussed below.

Sialic Acids

Sialic acids are prominently expressed at the outer termini of N-glycans, O-glycans, and glycosphingolipids of the deuterostome lineage of animals (see Figure 19.1 and Chapter 14). It was once thought that sialic acids were an evolutionary innovation unique to this lineage that originated during the Cambrian Expansion, and that all other reports of sialic acids in a few scattered taxa reflected lateral gene transfer and/or convergent evolution (i.e., independent evolution of sialic acid synthesis in these taxa). However, although such lateral transfer mechanisms exist and may explain the presence of sialic acids in some bacterial taxa, sialic acids are also reported in some fungi and mollusks. Together with evidence for a limited set of genes for sialic acid production and addition in some protostomes (e.g., in insects such as *Drosophila*), the situation is indicative of an earlier evolutionary origin for sialic acids. *Caenorhabditis elegans*, the free-living nematode, does not contain genes for synthesizing or metabolizing sialic acid. In addition, prior claims of the presence of sialic acids in plants are probably due to environmental contamination and/or incorrect identification of the chemically related sugar Kdo (3-deoxy-octulosonic acid). However, recent studies have found that the sialic acid biosynthetic genes of some insect and bacterial species share homology with those of vertebrates. Overall, it appears likely that sialic acids were an ancient invention

derived from genes of the related pathway for Kdo synthesis. In this scenario, sialic acids were differentially exploited during evolution, becoming prominent only in the deuterostome lineage, while being abandoned or substantially reduced in complexity and/or biological importance in other animal and fungal taxa. Meanwhile, a variety of bacteria synthesize other sialic-acid-like molecules using a very similar biosynthetic pathway (see Chapter 14).

It is curious that the most sialic acid diversity tends to be found in invertebrate deuterostomes, such as echinoderms (sea urchins and starfish), and the simplest profiles are found in humans. Likewise, whereas complex substituted polysialic acids are found in echinoderms and fish eggs, simpler polysialic acids are found in humans. Thus, sialic acids are not subject to incremental sophistication along recently evolved lineages. Rather, they seem to have evolved in many possible directions, disappearing altogether, or complicating or simplifying their structures. Although there is a tendency for some types of sialic acids to be dominant in certain mammalian species (e.g., *N*-glycolylneuraminic acid [Neu5Gc] in pigs and 4-O-acetylated sialic acids in horses), careful investigation reveals the presence of lower quantities of such sialic acids in most other species. Polysialosyl groups and sialic acid O-acetylation in gangliosides seem to be particularly enriched in poikilothermic (cold-blooded) animals. An interesting finding is that humans are "knockout" primates for the enzyme CMP–Neu5Ac hydroxylase (CMAH), because they contain a mutated and inactive *CMAH* gene. Thus, unlike the closely related great apes, humans are deficient in expression of Neu5Gc acid (see Chapter 14). As chickens also make an immune reaction against Neu5Gc, it remains to be determined whether birds represent another lineage that lost this sialic acid.

Glycosaminoglycans

Structures thought to be typical of "higher" animal heparan and chondroitin sulfate chains have been found in many invertebrates, including insects (see Chapter 24) and mollusks. The most widely distributed and evolutionarily ancient class appears to be chondroitin chains, which are not always sulfated (e.g., in *C. elegans*) (see Chapter 23). The more highly sulfated and epimerized forms of heparin and dermatan sulfate tend to be found primarily in "higher" animal species of the deuterostome lineage. The same is true of hyaluronan. Echinoderms such as the sea cucumber make typical chondroitin chains, but some glucuronic acids have branches containing fucose sulfate. Simpler multicellular animals such as sponges can have novel glycosaminoglycans that include uronic acids, but they do not have the typical repeat units of chondroitin sulfate and heparan sulfate. Plants do not have typical animal glycosaminoglycans. Instead, they have acidic pectin polysaccharides, characterized by the presence of galacturonic acid and its methyl ester derivative (see Chapter 22). Bacteria have completely distinct polysaccharides (Chapter 20), although certain pathogenic strains can mimic mammalian glycosaminoglycan chains (see below).

Glycosylphosphatidylinositol Anchors

Glycosylphosphatidylinositol (GPI)-anchored proteins and lipids (see Chapter 11) that share the "core" motif Manα1-4GlcNα1-6-*myo*-inositol-1-P lipid are distributed ubiquitously in eukaryotes. In some species (e.g., yeasts and slime molds), the lipid tail can be a ceramide instead of a phosphatidylinositol (see Chapter 21). GPI-anchored lipids and proteins can constitute the major components of the highly variable outer membranes of some parasitic protozoans, such as *Leishmania* and *Trypanosoma* (see Chapter 40). GPI anchors are generally thought to be absent in prokaryotes. However, at least one archaeal organism has been reported to have a GPI-anchored protein.

Nuclear and Cytoplasmic Glycans

The O-β-GlcNAc modification commonly found on cytoplasmic and nuclear proteins (see Chapter 18) is widely expressed in "higher" animals and in plants. Conserved homologs of the O-GlcNAc transferase that is responsible for synthesizing this structure have been found in many eukaryotic taxa. No clear homolog is evident in the yeast genome. Although the structure has been claimed in *Dictyostelium*, this is actually a distinct α-linked O-GlcNAc. There is currently no evidence that bacteria or Archaea can generate this modification.

VIRUSES ACQUIRE GLYCOSYLATION FROM THEIR HOSTS

Viruses often carry minimalist genomes that typically do not direct glycosylation of their own glycoproteins but instead utilize host-cell machinery. Thus, the glycosylation of viruses reflects that of host cells from which they emerge. However, there are some exceptions to this rule, with some viruses and especially bacteriophages containing genes encoding unusual glycosyltransferases. For example, the chlorella virus generates a glycoprotein termed PBCV-1 that is modified by a "Fringe-type" glycosyltransferase encoded in the viral genome. Some of the phage viral glycosyltransferases can modify their surface antigens to change the serotype of their host bacteria or glycosylate their own DNA to block it from degradation by restriction enzymes. Baculoviruses also encode their own glycosyltransferases to glycosylate host ecdysteroids, allowing them to block molting of the insect host.

Utilization of host glycosylation machinery is particularly prominent in the case of enveloped viruses. Most viral envelope glycoproteins are glycosylated (mostly with N-glycans) during the passage of these proteins through the host Golgi apparatus. This glycosylation is typically quite extensive and appears to protect the virus from host immune reactions directed against the underlying viral polypeptide. In this regard, it has been suggested that the relatively common occurrence of the heterozygous state for congenital disorders of glycosylation in humans (see Chapter 42) may reflect selection for heterozygous individuals whose genomes interfere with viral replication by preventing complete glycosylation of proteins of invading viruses. In other instances, host lectins may be "hijacked" by the glycans on viral surface glycoproteins, aiding attachment and/or entry into target host cells.

VAST DIVERSITY IN BACTERIAL AND ARCHAEAL GLYCOSYLATION

Despite the enormous potential for structural diversity built into monosaccharides, a rather limited subset of all possible monosaccharides and their possible linkages and modifications are found in eukaryotic cells. Why one encounters only such a limited subset of the possible glycan structures is one of the puzzling questions of glycobiology. On the other hand, this limited subset has allowed extensive elucidation of the structure of eukaryotic glycans. In contrast, Bacteriae and Archaea have had several billion additional years to experiment with glycan variation. These organisms also have short generation times and are capable of exchanging genetic material across vast phylogenetic distances, via plasmid-mediated horizontal gene flow. They also inhabit a much wider range of ecological niches with innumerable physicochemical and biological conditions, ranging from the deep litho- and hydrosphere to the stratosphere. Thus, it should not come as a surprise that bacteria and Archaea express a much greater diversity in glycosylation, both in terms of range of monosaccharides that they utilize or synthesize and with regard to their types of linkages and modifications. Some discussion of such structures is presented in Chapter 20. However, much of the work to date has focused on the glycans of pathogens, and it is safe to say that we have barely scratched the surface of this diversity.

TABLE 19.1. *Examples of molecular mimicry of animal glycans by pathogenic bacteria*

Organism	Animal glycans synthesized
E. coli K1, *Meningococcus* group B	polysialic acid
E. coli K5	heparosan (heparan sulfate backbone)
Group A *Streptococcus*	hyaluronan
Group B *Streptococcus*	sialylated *N*-acetyllactosamines
Campylobacter jejuni	sialylated ganglioside-like glycans

MOLECULAR MIMICRY OF HOST GLYCANS BY PATHOGENS

It is evident that great differences exist between the pathways generating the glycan structures of bacteria and those of vertebrates. Despite this, occasional microbial surface structures are found to be strikingly similar to those of mammalian cells. Interestingly, most examples of this type of "molecular mimicry" occur in pathogenic microorganisms, presumably adapting them for better survival in the host by avoiding, reducing, or manipulating host immunity. A few examples are listed in Table 19.1. The initial hope of scientists trying to clone vertebrate glycosyltransferases was that most of the responsible microbial genes arose from lateral gene transfer and that these would provide a backdoor approach to isolating the corresponding ones from eukaryotes. However, in most instances where full genetic information has become available, the evidence points toward convergent evolution rather than gene transfer as the dominant mechanism. For example, the genes involved in synthesizing sialic acids in bacteria seem to have been mainly derived from the preexisting bacterial pathways for the biosynthesis and transfer of Kdo, a bacterial sugar with a structural resemblance to sialic acids. Meanwhile, bacterial sialyltransferases bear little resemblance to those of eukaryotes, and the vast sequence differences between different bacterial sialyltransferases indicate that these have even been reinvented on several separate occasions. On the other hand, lateral gene transfer appears to have been quite common among the bacteria and Archaea themselves, facilitating rapid phylogenetic dissemination of such enzymatic "inventions."

INTERSPECIES AND INTRASPECIES DIFFERENCES IN GLYCOSYLATION

Why do closely related species differ with regard to the presence or absence of certain glycans? Does the same glycoprotein have the same type of glycosylation in different but related species? Relatively little data are available concerning these issues, but examples of both extreme conservation and extreme diversification can be found. A reasonable explanation is that conservation of glycan structure is only required when there are very specific functions for the glycans in question. In other instances, considerable drift in the details of glycan structure might be tolerated, as long as the underlying protein is able to carry out its primary functions (changes with no consequences for survival or reproduction, i.e., selectively neutral).

Even in the absence of important functions within an organism, glycans can have important roles in the mediation of interactions with symbionts and pathogens. The evolution of diversity and microheterogeneity (across tissues and cell types) in glycosylation could well be of value to the organisms in evading pathogens that use glycans as signposts for attachment and entry. Glycans can also have important roles in attracting the important symbiont microbial communities needed for gastrointestinal functions and in accommodating or restricting these to particular areas of the host.

It is also clear that there can be significant variation in glycosylation among members of the same species, particularly with regard to terminal glycan sequences. The classic example

is that of the ABH(O) blood group system (see Chapter 13), a glycan-defined polymorphism found in all human populations, which has also persisted for tens of millions of years of primate evolution and has even been independently rederived in some instances. Somewhat surprisingly, despite its great clinical importance for blood transfusion, this polymorphism appears to cause no major differences to the intrinsic biology of individuals of the species (see Chapter 13). Like other blood groups, the ABO polymorphism is accompanied by the production of antibodies against the other variants. It has been suggested that these antibodies are protective, by causing complement-mediated lysis of enveloped viruses generated within other individuals who can express the target structure for the antibody. Thus, an enveloped virus generated in a B blood group individual might bear this structure and be susceptible to lysis upon contact with an A or O blood group individual, who would express anti-B antibodies. Recent experimental evidence is supportive of such a mechanism. However, this mechanism alone should strongly favor O individuals, as these form antibodies against the A and B variants and should lead to higher frequencies of O type than are observed.

Another possible explanation for interspecies diversity is the selection exerted by pathogens that recognize glycans as targets for attachment and entry into cells. This mechanism is likely operative in generating the diversity of sialic acid types and linkages (see above). However, it should result in selection of ABO subtypes and result in approximately even frequencies of each phenotype, not what is observed. Recent analyses have tried to combine the two mechanisms: the antibody-mediated protection from intracellular viruses and possible frequency-dependent protection from glycan-exploiting extracellular pathogens, such as Noroviruses and *Plasmodium falciparum* malaria. Modeling approaches have successfully generated observed frequencies of ABO by incorporating these two simultaneous selection pressures. It is fair to say that the evolutionary persistence of the ABO system needs further explanation.

Another unexplained phenomenon is the genetic inactivation in Old World primates of the ability to synthesize the otherwise very common terminal Galα1-3Galβ1-4GlcNAc-R structure. This glycan variation system is also associated with spontaneously appearing and persistently circulating antibodies against the missing glycan determinant, thus forming a kind of interspecies "blood group." It has been proposed that this glycan difference is protective for the primate lineage which lost "α Gal" and has a high-titer circulating antibody, as it is now better protected against infection by viruses emanating from other mammals.

Regardless of the precise underlying purposes of these types of polymorphic systems, such intra- and interspecies diversity might also provide for "herd immunity," a phenomenon whereby one glycan-variant-resistant individual can effectively protect other susceptible individuals by limiting the spread of a pathogen through the population. It is also important to emphasize that these proposed protective functions of glycan diversity are only apparent at the level of populations and not the individual. This complicates their study in model organisms, where the focus is classically on the individual.

Future studies will have to test precisely how much of interspecies and intraspecies glycan variation is directly driven by such host–pathogen interactions. Despite a lack of comprehensive studies of such phenomena, it is becoming clear that glycan variation forms an important determinant of host susceptibility and must be considered when trying to understand disease, especially epidemics or zoonotics involving different host species and their interactions, for example, influenza A (see Chapter 14).

"MODEL" ORGANISMS FOR STUDYING GLYCAN DIVERSITY

Details of glycan expression patterns in various popular "model" organisms can be found in other chapters in this volume. In recent years, there has been increasing definition of the

structures of bacterial inner cell wall peptidoglycans and the outer membranes that are composed of lipooligosaccharides and lipopolysaccharides, particularly those of *Escherichia coli* (see Chapter 20). Chapter 21 provides some details about various genera and species of fungi and protists, including *Saccharomyces cerevisiae* and *Dictyostelium discoideum*. Pathogenic protists such as trypanosomes and leishmanial parasites express very high densities of surface GPI anchors and are discussed in Chapters 11 and 40. Chapter 23 presents an overview of the roundworm *C. elegans*, its development, glycan structures, and expression of glycosyltransferases. The functional insights derived from this organism cover all classes of glycans. For some details about glycobiology of *Drosophila*, see Chapter 24. Like *C. elegans*, this workhorse of genetics has made important functional contributions to the field during the last 5 years. It has been especially important in understanding how O-fucosylation modifies Notch signaling (Chapter 12) and how heparan sulfate proteoglycans determine morphogen and growth factor gradients. Chapter 22 discusses the glycans of plants, including those of the model organism, *Arabidopsis*. Various aspects of sea urchin glycobiology, including the acrosome reaction, egg/sperm interactions, and the role of proteoglycans and lectins, are covered in Chapter 25, which also discusses aspects of *Xenopus* glycobiology, including the synthesis of chitin oligosaccharides, the role of proteoglycans in determining left–right asymmetry, and lectins that help in fertilization and the innate immune system. The same chapter also discusses aspects of zebrafish glycobiology, including glycoproteins, proteoglycans, and lectins.

The recent discovery that rodents are the closest evolutionary cousins to primates has provided added justification for the use of rats and mice as model organisms to understand the mechanisms of human disease (see Chapter 25). Last but not least, Nobel Laureate Sydney Brenner has suggested that we now have enough information about humans to consider ourselves to be the "next model organism." Indeed, there is an increasing tendency to focus tractable questions about glycans and their biology directly on humans and on naturally occurring human mutants. Most recently, there has also been interest in studying the great apes (our closest evolutionary cousins) and an independent realization of recent hominid evolution, particularly with regard to several differences in sialic acid biology (see Chapter 14).

WHY DO WIDELY EXPRESSED GLYCOSYLTRANSFERASES SOMETIMES HAVE LIMITED INTRINSIC FUNCTIONS?

Prior to the generation of glycosyltransferase-deficient mice, it was popular to suggest that every single glycan on every single cell type must have a critical intrinsic host function. Analysis of available gene disruption data indicates that this is not the case. For example, the ST6Gal-I α2-6 sialyltransferase is the main enzyme that produces Siaα2-6Galβ1-4GlcNAcβ1- termini on vertebrate glycans. Although this sequence serves as a specific ligand for the B-cell regulatory molecule CD22 (Siglec-2; see Chapter 32), it is also found on many other cell types, as well as on many soluble secreted glycoproteins. Furthermore, the ST6Gal-I mRNA varies markedly among cell types, and its transcription is regulated by several cell-type-specific promoters, which are in turn modulated by hormones and cytokines. Despite all these data suggesting very diverse and complex roles for this enzyme and its products, the prominent functional consequences of eliminating its expression in mice so far seem to be restricted to the B cell, with decreased signaling and proliferative responses and impaired antibody production (see Chapter 32). Few other obvious abnormalities have yet been found in organ structure and physiology, morphology, or behavior. If the specific intrinsic functions of the ST6Gal-I glycan product are in fact restricted to B

cells, why does the organism express it in so many other locations? Even more puzzling, why up-regulate its expression so markedly in the liver and endothelium during a so-called "acute phase" inflammatory response? Could it be that scattered expression of this structure in other locations represents a "smoke-screen" effect, restricting intraorganismal spread of an invading pathogen? Could it be that heavily glycosylated nonnucleated cells like erythrocytes act as a "sink" to divert viral pathogens that need nucleated cells for replication? The answers to these questions must take into account the evolutionary selection pressures (both intrinsic and extrinsic recognition phenomena such as host–pathogen interactions and innate immune contributions) on glycosyltransferase products. Many of these effects may also not be apparent in inbred genetically modified mice living in hygienic vivaria but may rather require population studies of animals in a natural, pathogen-rich environment. It is also possible that other gene products are masking the phenotypes in these model systems, by compensating for the genetic loss. Furthermore, it is likely we have not looked hard enough at such genetically modified mice nor applied the right environmental pressures to elicit phenotypes.

EVOLUTIONARY FORCES DRIVING GLYCAN DIVERSIFICATION IN NATURE

There is too little information available today to allow a comprehensive exposition of the evolution of even the major classes of glycans. On the basis of the available data, it is reasonable to suggest that glycan diversification in complex multicellular organisms has been driven by evolutionary selection pressures of both intrinsic and extrinsic origin relative to the organism under study (see Chapter 6). It is reasonable to postulate that glycans are particularly susceptible to the "Red Queen" effect, in which host glycans must keep on changing in order to stay ahead of the pathogens, which have extremely rapid evolutionary rates because of short generation times, high mutation rates, and much horizontal gene transfer. Given the rapid evolution of extrinsic pathogens and their frequent use of glycans as targets for host recognition, it seems likely that a significant portion of the overall diversity in vertebrate cell-surface glycan structure reflects such pathogen-mediated selection processes. Meanwhile, even one critical intrinsic role of a glycan would disallow its elimination as a mechanism to evade pathogens. Thus, the glycan expression patterns of a given organism may represent a compromise between evading pathogens and preserving intrinsic functions.

More gene disruption studies in intact animals would be helpful to differentiate between these intrinsic and extrinsic glycan functions. More systematic comparative glycobiology could also contribute, by making predictions about intrinsic glycan function; that is, the consistent (conserved) expression of the same structure in the same cell type across several taxa would imply a critical intrinsic role. Such work might also help define the rate of glycan diversification during evolution, better define the relative roles of the intrinsic and extrinsic selective forces, and eventually lead to a better understanding of the functional significance of glycan diversification during evolution. The possibility that glycan diversification might even drive the process of speciation (via reproductive isolation) also needs to be considered.

FURTHER READING

Warren L. 1963. The distribution of sialic acids in nature. *Comp. Biochem. Physiol.* **10:** 153–171.

Kishimoto Y. 1986. Phylogenetic development of myelin glycosphingolipids. *Chem. Phys. Lipids* **42:** 117–128.

Galili U. 1993. Evolution and pathophysiology of the human natural anti-α-galactosyl IgG (anti-Gal) antibody. *Springer Semin. Immunopathol.* **15:** 155–171.

Kappel T., Hilbig R., and Rahmann H. 1993. Variability in brain ganglioside content and composition of endothermic mammals, heterothermic hibernators and ectothermic fishes. *Neurochem. Int.* **22:** 555–566.

Martinko J.M., Vincek V., Klein D., and Klein J. 1993. Primate ABO glycosyltransferases: Evidence for trans-species evolution. *Immunogenetics* **37:** 274–278.

Dairaku K. and Spiro R.G. 1997. Phylogenetic survey of endomannosidase indicates late evolutionary appearance of this N-linked oligosaccharide processing enzyme. *Glycobiology* **7:** 579–586.

Drickamer K. and Taylor M.E. 1998. Evolving views of protein glycosylation. *Trends Biochem. Sci.* **23:** 321–324.

Gagneux P. and Varki A. 1999. Evolutionary considerations in relating oligosaccharide diversity to biological function. *Glycobiology* **9:** 747–755.

Freeze H.H. 2001. The pathology of N-glycosylation—Stay the middle, avoid the risks. *Glycobiology* **11:** 37G–38G.

Angata T. and Varki A. 2002. Chemical diversity in the sialic acids and related α-keto acids: An evolutionary perspective. *Chem. Rev.* **102:** 439–470.

Varki A. 2006. Nothing in glycobiology makes sense, except in the light of evolution. *Cell* **126:** 841–845.

Bishop J.R. and Gagneux P. 2007. Evolution of carbohydrate antigens—Microbial forces shaping host glycomes? *Glycobiology* **17:** 23R–34R.

CHAPTER 48

Glycomics

Carolyn R. Bertozzi and Ram Sasisekharan

THE TERM "GLYCOME" DESCRIBES THE COMPLETE REPERTOIRE of glycans and glycoconjugates that cells produce under specified conditions of time, space, and environment. "Glycomics," therefore, refers to studies that profile the glycome and is the topic of this chapter.

HISTORICAL PERSPECTIVE OF "OMICS" SCIENCE: GENOMICS, TRANSCRIPTOMICS, AND PROTEOMICS

The field of genomics arose from the availability of complete genome sequence data as well as computational methods for their analysis. One of the surprising findings from analysis of the human genome was the presence of fewer protein-encoding genes (a mere 25,000) than had been predicted earlier. Furthermore, the protein-encoding (i.e., "expressed") genes comprise a small fraction, less than 2%, of the human genome. These genes are transcribed into mRNAs that are often referred to collectively as the "transcriptome." The ability to analyze transcripts in a high-throughput parallel format using a DNA microarray, or "gene chip," has enabled researchers to probe global differences in gene expression, for instance, between healthy and diseased cells, between neurons and muscle cells, and

between drug-sensitive and drug-resistant cancer cells. Such "transcriptomic" comparisons have revealed networks of genes whose expression is linked to disease.

Although many scientific discoveries have emerged from genomic and transcriptomic approaches, this information still does not provide a complete picture of the physiology of a cell or organism. The proteins expressed by the cell, collectively termed the "proteome," perform many of the cell's functions. Most eukaryotic proteins are posttranslationally modified (e.g., by phosphorylation, oxidation, ubiquitination, lipidation, or glycosylation). These modifications, combined with alternative splicing in eukaryotes, render the proteome considerably more complex than the transcriptome. Although it is not known how many discrete proteins a particular human cell expresses, estimates between 50,000 and 120,000 have been suggested. Direct characterization of the proteome is required to understand both its complexity and its global functions. The global systems-level analysis of all proteins expressed by cells, tissues, or organisms is referred to as "proteomics."

Unlike the genome, which is fixed for most cells, the proteome is dynamic. The repertoire of proteins expressed by a cell is highly dependent on its tissue type, microenvironment, and stage within its life cycle. As cells receive cues in the form of growth factors, hormones, metabolites, or other agents, various genes are turned on or off. Thus, proteomes vary during cell differentiation, activation, trafficking, and during malignant transformation. Also, many proteins are secreted from cells and circulate in the blood or lymphatic fluid or are excreted in the saliva, mucus, tear fluid, or urine. These bodily fluids also have distinct proteomes.

WHAT IS "GLYCOMICS"?

Glycomic analyses seek to understand how a collection of glycans relates to a particular biological event. As described throughout this book, glycans participate in almost every biological process, from intracellular signaling to organ development to tumor growth. Understanding how the totality of glycans governs these processes is a central goal of glycobiology.

The glycomes of life-forms include all of the glycan and glycoconjugate types that have been described in this book. For example, vertebrates possess protein-associated N- and O-glycans, glycosaminoglycans, and GPI anchors, as well as lipid-associated glycans and free glycans such as hyaluronan (see Chapters 8–18). Other organisms possess their own distinct glycomes, with those of plants (see Chapter 22) and prokaryotes (see Chapter 20) differing greatly in composition from the vertebrate and invertebrate glycomes (see Chapter 25). And as with the proteome, each cell type has its own distinct glycome that is governed by local cues and the cell's internal state. The size of any particular glycome has not yet been established, but we know that glycomes can far exceed proteomes and transcriptomes with respect to complexity. For example, some estimates have placed the vertebrate glycome at more than one million discrete structures. Furthermore, it appears that the glycome is considerably more dynamic than the proteome or transcriptome.

The notion that glycans should be studied as a totality, as well as simply one at a time, is not a radical concept among glycobiologists. Indeed, researchers in the field have long known that glycans form patterns on cells that change during development (see Chapter 38) and cancer progression (see Chapter 44). Also, many glycan-binding proteins are oligomerized on cells and interact with multivalent arrays of glycans on opposing cells (see Chapter 27). In some cases, multiple discrete glycan epitopes work in concert to engage two cells or deliver a signal from one cell to the other. Thus, before "glycomics" was coined, scientists had already concluded that many aspects of glycobiology can be understood only with a sys-

tems-level analysis. Conversely, no systems-level analysis of a biological process is complete without interrogating the glycome in addition to the genome, transcriptome, and proteome.

RELATIONSHIP OF THE GLYCOME TO THE GENOME AND PROTEOME

Clues regarding the composition and complexity of the glycome can be found in the cell's genome, transcriptome, and proteome. As discussed in Chapter 7, genome "mining" using known sequences can identify many genes involved in glycan biosynthesis and processing. By such an analysis, more than 250 glycosyltransferases have been found encoded in the human genome as well as many nucleotide sugar biosynthetic enzymes and Golgi transporters (see Chapter 5). Some of the corresponding enzymes have been studied biochemically and their glycosyl donor and acceptor specificities have been defined, whereas others have been assigned predicted functions based on sequence relationships. Furthermore, expression patterns of many glycosyltransferases have been determined in human and mouse tissues using northern blots, quantitative PCR (polymerase chain reaction), and transcriptomic analyses.

In principle, one might use all of this information to construct "virtual glycomes." However, this exercise is of limited value because the combinatorial action of glycosyltransferases in many competing biosynthetic pathways renders the complete glycome very difficult to predict with any accuracy. As an example, the reduced expression of a single glycosyltransferase can perturb the biosynthesis of dozens of glycan structures, some negatively and some positively. The direct glycan products of the glycosyltransferase will be reduced in expression, whereas glycans made by other enzymes that compete for common intermediates might increase in levels. Furthermore, unlike the genome and, to our knowledge, the proteome, the glycome can be sensitive to exogenous nutrient levels. Thus, variations in dietary monosaccharides, such as glucose, galactose, glucosamine, fucose, mannose (see Chapter 18), and *N*-glycolylneuraminic acid (see Chapter 14), can change the composition of the glycome. Because of these complexities, transcriptomic and proteomic data can at best guide hypotheses regarding the presence or absence of specific classes of structures. In contrast, the absence of particular genes (e.g., the sialic acid biosynthesis machinery in *Caenorhabditis elegans*; see Chapter 14) has been useful in assessing the relative compositions of various glycomes.

The numerous factors that influence the glycome (the genome, the proteome, and environmental nutrients, as well as the secretory machinery, pH, and many other determinants) create a system that is highly diverse and dynamic. Thus, the glycome can change dramatically in response to a subtle change in the cellular system. This feature makes glycomics research both exciting and also daunting. Since neither the proteome nor the transcriptome can accurately predict such a moving target, the glycome must be analyzed directly. Techniques that have been employed to characterize the glycome are summarized below.

TOOLS FOR CHARACTERIZING THE GLYCOME

The glycome can be described at many hierarchical levels of complexity (Figure 48.1). First, the glycome can be deconstructed into an inventory of glycan structures separated from their protein or lipid scaffolds and independent of their location in the cell, organ, or organism. This first hierarchical level is essentially a catalog of structures. It is an important starting point for any comprehensive glycome analysis. But how the parts in the catalog assemble to form the intact system is also important for understanding function. Thus, a second hierarchical level of analysis involves defining which glycans are associated with individual proteins or lipids. Analysis of the complete repertoire of a cell's glycoproteins, including

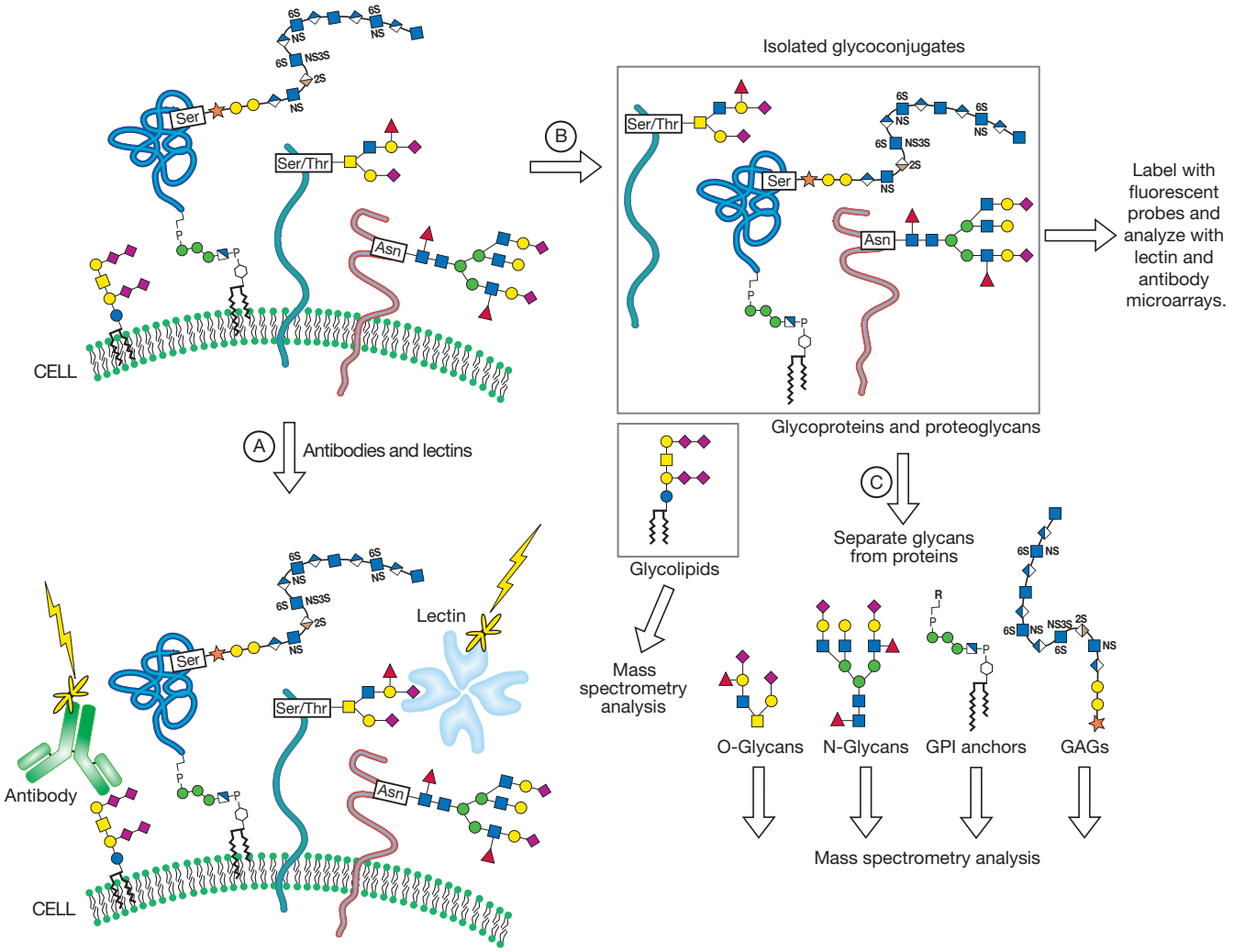

FIGURE 48.1. Multiple approaches for profiling a cell's glycome at various hierarchical levels of complexity. (Step A) Cells can be directly probed for glycan expression using labeled lectins and glycan-specific antibodies. This top–down experiment provides a global view of the distribution of certain glycan epitopes on cells and tissues but does not afford detailed structural information. (Step B) Glycoproteins and glycolipids can be isolated from cell lysates and then analyzed using lectin and antibody microarrays and mass spectrometry methods. Glycolipids can be sequenced directly, whereas glycoproteins are often further deconstructed into separate glycans and proteins before structural analysis. (Step C) Isolated glycans are separated based on type (i.e., N-glycans, O-glycans, glycosaminoglycans, etc.) and their sequences determined by mass spectrometry. Alternatively, intact glycoproteins can be digested with trypsin and the glycopeptides characterized by mass spectrometry. This approach retains the glycan–peptide linkage and allows assignment of sites of protein glycosylation. Collectively, steps B and C comprise a bottom–up glycomic analysis.

their glycan structures and sites of attachment, lies at the intersection of glycomics and proteomics and is often referred to with the term "glycoproteomics." A third level of complexity involves determining which glycans or glycoconjugates are expressed on specific cells or tissues. This level of glycomic profiling is essential if the goal is to reveal new functions in cell–cell communication or to correlate particular glycomes with disease tissue. A final level that has yet to be investigated involves visualizing how glycoconjugates are actually organized relative to each other within the cell, at the cell surface, and in the extracellular matrix.

As described below, numerous techniques have been developed for interrogating the glycome at these various hierarchical levels. No single technique can define all aspects of the gly-

come. Thus, several approaches are typically employed in parallel, allowing one to assemble a picture of the glycome both from the "bottom up" (i.e., from the individual glycan repertoire) and from the "top down" (i.e., from a global tissue expression analysis). A significant challenge in analyzing the glycome derives from its enormous structural diversity. Different approaches and techniques are required to characterize the structures of glycoproteins versus glycolipids, N-glycans versus O-glycans, and sulfated glycosaminoglycans versus neutral glycans (see Chapter 47). By contrast, a single technique, the DNA microarray, can be used to interrogate all RNA transcripts at once. Thus, at present, the techniques for glycomic analysis remain relatively low-throughput and specific for a particular glycan type, although considerable effort is being directed toward methods that encompass all glycan classes.

Mass Spectrometry

High-resolution mass spectrometry (see Chapter 47) is the primary technique for characterizing the structures of individual glycans when only small quantities are available, as is the case in most glycomic studies. In a typical experiment, a glycoprotein- or glycolipid-enriched sample is prepared from cell lysates and analyzed by multiple rounds of mass spectrometry. In the case of glycoproteins, the N-glycans can be selectively released enzymatically or chemically, separated by HPLC (high-pressure liquid chromatography) methods, and actually sequenced. Separately, the O-glycans are released chemically and sequenced as well. Glycolipids can often be directly sequenced without separation of the lipid component. Glycosaminoglycans are more problematic because of their large size, but small fragments can be sequenced by mass spectrometry in conjunction with enzymatic digestion (see Chapter 16). An advantage of mass spectrometric glycan profiling is that multiple glycans of any given subtype can be profiled at once, increasing the throughput of the glycomic analysis. Still, there is no method at present by which highly complex samples possessing many glycan subtypes can be analyzed in one mass spectrometry experiment. Furthermore, many of the techniques routinely used tend to partially or completely destroy the sample or miss potentially important modifications such as sulfation and O-acetylation.

Mass spectrometry can also be employed to define sites of attachment of glycans to the underlying protein scaffold (i.e., for glycoproteomic analysis). Typically, the glycoprotein-associated glycans are first trimmed to remove peripheral epitopes such as sialic acid and fucose residues. This procedure simplifies the diversity of the pool and therefore sacrifices some of the information in the glycome. Then, the glycoprotein is subjected to tryptic digest and the peptides are analyzed by mass spectrometry using a technique that leaves the glycans attached to the peptide during the analysis (termed electron transfer dissociation [ETD] mass spectrometry). In parallel, the protein can be stripped of its glycans before the tryptic digest and the masses of those naked peptides can be compared to those of the glycopeptides. The differential allows prediction of the attached glycan structure. Furthermore, if an endoglycosidase such as PNGase F is used to release N-glycans (see Chapter 47), the resulting change from asparagine to aspartic acid can mark the site of the original glycosylation. Similarly, β-elimination of O-glycans changes the amino acid at the site of elimination, and can also be followed by Michael addition of nucleophiles such as dithiothreitol to mark the original O-linked sites.

A common problem encountered in proteomic studies is a lack of sensitivity for low-abundance species. The range of protein levels in cells and bodily fluids is thought to span more than eight orders of magnitude. Mass spectrometry detection often suffers from saturation by the most abundant species, leaving those with lower abundance impossible to detect. For glycoproteomic analyses, therefore, it is often beneficial to enrich the sample in glycoproteins and to discard those proteins that do not bear glycans. Lectins have been artfully employed for this purpose. As discussed in Chapter 45, a variety of glycan-binding

proteins (e.g., the plant lectins) are commercially available. These proteins can be immobilized on agarose beads and used for affinity purification of glycoproteins from cell lysates or body fluids. Once enriched, the glycoproteins can be analyzed in the absence of abundant unglycosylated protein contaminants.

Lectin and Antibody Arrays

A major benefit of mass spectrometry is the detailed information it provides regarding the structure of a glycan. A drawback, however, is its relatively low throughput and the need for different experimental protocols for each glycan subtype. Lectin and antibody arrays can be employed to interrogate the glycome with much higher throughput, although in considerably less structural detail. As described in Chapter 45, nature has provided a considerable collection of lectins and many have been biochemically characterized. These lectins possess a range of specificities; some recognize a particular monosaccharide in virtually any context, whereas others are very specific for higher-order glycan epitopes or single residues within a defined context. For those epitopes that lack a naturally occurring lectin, one can generate monoclonal antibodies (see Chapter 45). This can be accomplished by immunization of a rodent with a synthetic or isolated version of the glycan attached to an antigenic carrier protein such as keyhole limpet hemocyanin. The monoclonal antibodies generated in this fashion can serve as "artificial lectins." Single-chain-fragment antibodies generated in bacteria have also proven useful for analysis of glycosaminoglycans. However, recent studies have noted that some antibodies once thought to be highly specific for certain glycans can cross-react with others.

The lectin array employs the same architecture as the DNA microarray or the glycan microarray described in Chapter 27. Lectins (or glycan-specific antibodies) are spatially arrayed on a glass chip by covalent attachment (Figure 48.2). Glycoproteins from the cell lysate or fluid sample of interest are nonspecifically labeled with a fluorescent dye. The sample is then incubated with the array and the fluorescence associated with each pixel is quantified. The pattern of bright spots reflects the glycome of the particular sample. For comparative purposes, two samples can be analyzed in parallel, one labeled with a green dye and one with a red dye. Combining the two samples onto one array allows direct analysis of changes in the glycome. In principle, samples that are even more complex than glycoproteins can be probed using lectin arrays, and indeed intact bacterial cells have been probed using this technique.

The lectin array provides global information about the types of glycan epitopes that are present in the sample but does not give any detailed structural information, nor does the experiment provide information regarding which proteins the glycans are attached to. However, the high-throughput platform allows for rapid comparison of many glycomes in search of global changes that might motivate further mass spectrometry studies.

Cell and Tissue Analysis Using Lectins and Antibodies

Glycobiologists routinely use lectins and glycan-specific antibodies as histological probes of glycan expression. This approach still holds an important place in any comprehensive glycomic analysis, as the cellular and tissue distribution of glycans is an important element of the glycome. In the modern era, tissue-expression patterns observed using lectins and antibodies can be correlated with lectin array data and mass spectrometry profiling data as well as genomic and proteomic data to create a more complete picture of the glycome. In the future, such in situ labeling followed by laser-capture microdissection of specific regions from tissue sections could potentially allow all the techniques to come together.

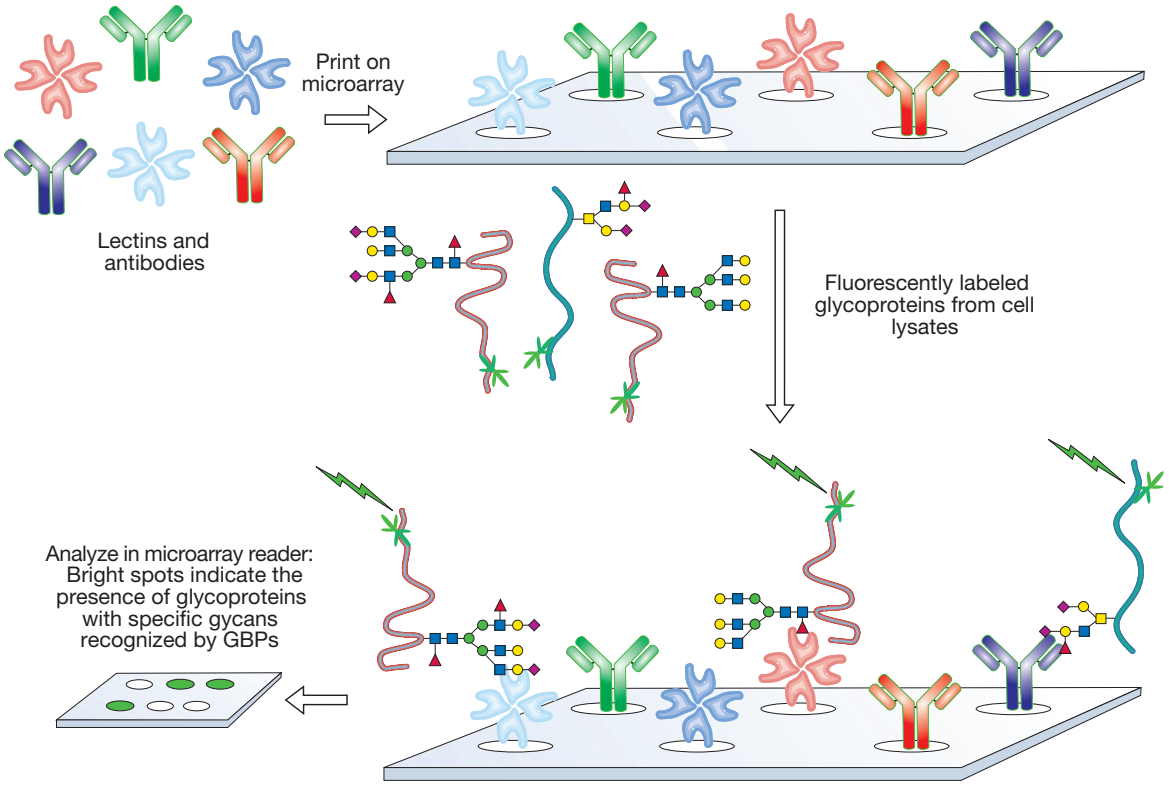

FIGURE 48.2. Analysis of cellular glycomes using lectin and antibody arrays. The arrays are generated by immobilization of lectins and glycan-specific antibodies on a chip. Glycoproteins from cell or tissue samples are labeled with a fluorescent dye and then incubated with the array. The fluorescent spots reflect the presence of glycoproteins bearing glycans recognized by the corresponding lectin or antibody. The technique provides minimal structural detail but permits rapid high-throughput analysis of many samples. Intact cells or virus particles can also be interrogated on lectin micrarrays.

Imaging the Glycome by Metabolic and Covalent Labeling

A recent addition to the arsenal of tools for glycome analysis is the use of metabolic labels that allow covalent tagging of glycans with imaging probes. As shown in Figure 48.3, cells or organisms can be treated with monosaccharide substrates bearing azido groups, and the downstream metabolic products are incorporated into cellular glycans. The azido groups can be covalently reacted with azide-specific imaging reagents. Once labeled, the glycans can be visualized on cells or tissues by fluorescence microscopy. This procedure has been used to image changes in the glycome during zebrafish embryonic development. The technique provides little structural information regarding the elements of the glycome that are labeled but has the advantage that global changes in the glycome can be monitored in vivo and in real time. By contrast, lectin and antibody reagents are largely restricted to ex vivo analysis of glycomes in cultured cells or tissues.

COMPARATIVE GLYCOMICS

Because the glycome is influenced by both genetic and environmental factors, the information contained therein might shed light on intraspecies and interspecies variations as well as on changes that have occurred over evolutionary history. From the immediate clin-

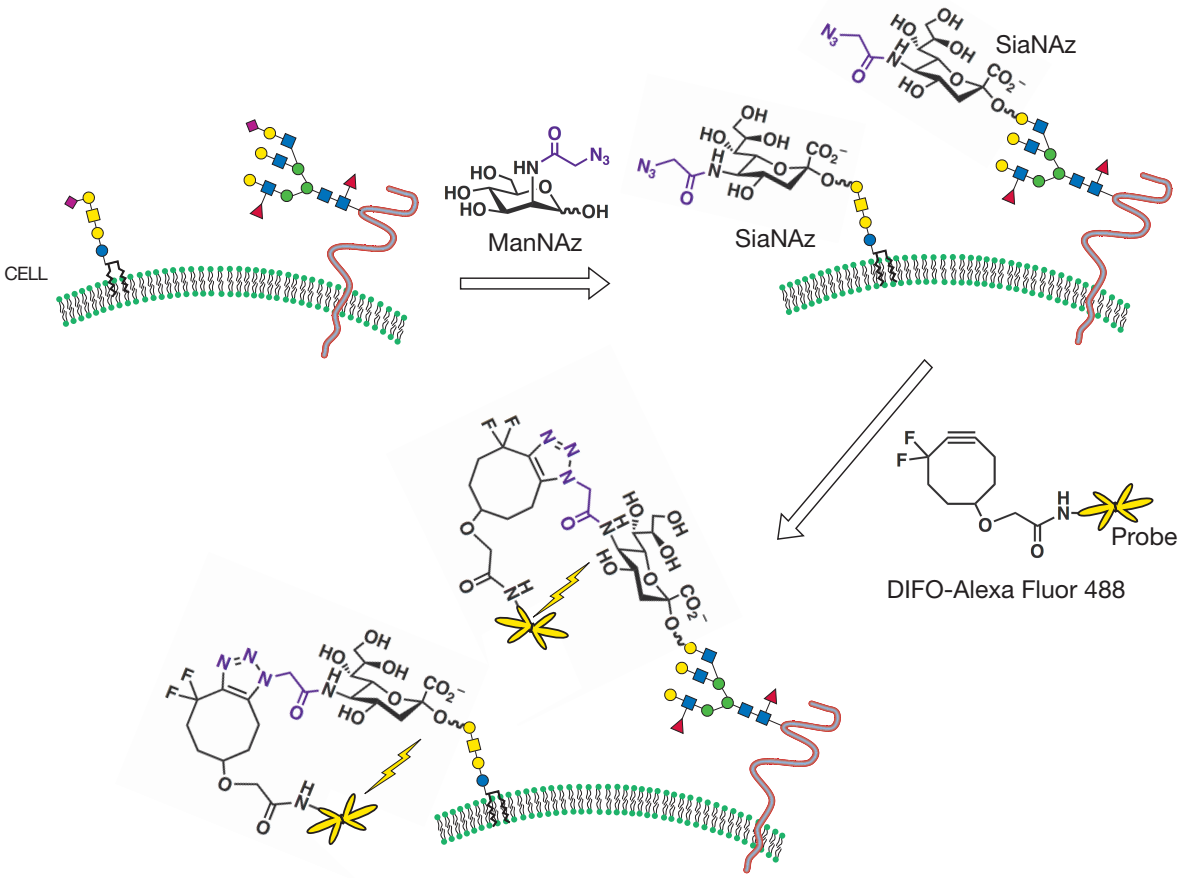

FIGURE 48.3. Metabolic and covalent labeling of glycans for in vivo imaging of the glycome. Cells metabolize azido sugars such as *N*-azidoacetylmannosamine (ManNAz), which is converted into the corresponding azido sialic acid (SiaNAz) and incorporated into cellular glycans. The azides are selectively reacted with a fluorescent probe conjugated to a cyclooctyne reagent termed "DIFO." The fluorescent dye-labeled glycans are then visualized. Using this two-step method, changes in the glycome can be probed within the physiologically authentic environment of a live organism.

ical perspective, the glycome might provide indicators of disease that can be used for diagnosis and for monitoring the efficacy of drugs. Comparative glycomics is therefore an exciting frontier in biology and medicine.

As discussed in detail in Chapter 44, numerous changes in the glycome have been associated with malignancy and metastasis, including altered N- and O-glycosylation, up-regulation of sialylated and fucosylated antigens, and altered heparan sulfates. In some cases, the tumor-associated glycans have been shown to be functionally relevant, whereas in other cases, the association of the glycan with cancer remains correlative. No matter its functional consequence, however, a change in the glycome that is highly correlated with malignancy (or any disease) can serve as a diagnostic biomarker. Given the dearth of such biomarkers for cancer screening, it is not surprising that considerable effort is being directed toward analysis of cancer-associated glycomes.

Already, mass spectrometry has been artfully employed in several glycomic studies of serum samples from healthy and cancer patients. In these studies, O-glycans were released from serum glycoproteins and analyzed by mass spectrometry. Although many structures were similar in the samples from healthy donors and cancer patients, a handful were markedly elevated in the latter. This observation suggests that changes in the serum gly-

come accompany cancer and that such changes might be used for diagnostic purposes. Notably, we do not need to understand *how* the disease of interest, in this case cancer, causes changes in the serum glycome in order to use that information for clinical benefit. Indeed, glycans that are altered in the disease sample might not be directly related to the disease at all. Rather, they may reflect downstream consequences of the disease on remote organs or they may reflect changes in the patient's immune system.

A fundamental question that remains unanswered is what is the extent of natural variation among individual human glycomes? Since the glycome can respond, in principle, to dietary and environmental changes, an equally interesting question is how glycomes around the world vary as a function of local dietary habits and/or medicine use, and further, how do glycome variations relate to acquired disease susceptibility (see Chapter 43)?

Studies of evolutionary biology have also much to gain from comparative glycomics. Evolution of the vertebrate immune system, for example, was accompanied by the acquisition of new glycan-binding proteins including the Siglec (see Chapter 32) and selectin (see Chapter 31) family members. The process by which the glycome evolved to accommodate these developments is of considerable interest. Likewise, the glycomes of microbes and their vertebrate hosts may have coevolved in some instances. The human blood group antigens, for example, are thought to reflect selective pressure induced by bacterial pathogens bearing similar glycan epitopes (see Chapter 39). This dramatic observation may be one of many examples of glycome coevolution across species. Comparative glycomic analysis of humans and their resident microbes, both pathogenic and symbiotic, may be highly revealing.

FUNCTIONAL GLYCOMICS USING GLYCAN MICROARRAYS

Taking inventory of the glycome provides the basis for hypotheses regarding biological function. Studying the functions of glycans is, of course, the central goal of the field of glycobiology. When this goal is pursued using high-throughput techniques, the term "functional glycomics" is often applied. As an example, the relative binding activity of glycans to glycan-binding proteins can be probed in high-throughput parallel fashion using glycan microarrays. These arrays seek to represent a fraction of the glycome and are often generated using glycans isolated from cells or tissue sources. The glycans are typically immobilized on a chip and then exposed to the protein of interest labeled with a fluorescent dye. Binding of the protein to the various glycans is detected by fluorescence imaging.

Arrays have been used for the analysis of a number of glycan-binding proteins, such as plant and microbial lectins, glycan-binding proteins involved in the innate and adaptive immune system, glycan-specific antibodies, viral glycan-binding proteins, and whole cells (see Chapter 27). Two examples of such applications are briefly described below.

DC-SIGN and DC-SIGNR

DC-SIGN and DC-SIGNR belong to the type II transmembrane receptor subfamily of C-type lectins (see Chapter 31). DC-SIGN is abundantly expressed on dendritic cells and plays a key role in adhesion of T cells as well as in the recognition of pathogens such as HIV. Indeed, binding of HIV to DC-SIGN on dendritic cells enhances T-cell infection. The related protein DC-SIGNR shares 77% sequence identity with DC-SIGN, but screening of these two similar proteins using glycan arrays revealed distinct ligand specificities. In addition to the high-mannose structures bound by both receptors, DC-SIGN recognized certain fucosylated ligands that were not bound by DC-SIGNR. This finding may shed light on the functional distinction between the two related proteins.

Influenza Virus Hemagglutinin

Influenza A virus subtypes are avian viruses that are named according to their surface antigens: hemagglutinin (HA) and neuraminidase (NA) (see Chapter 39). These viruses bind to sialylated glycans on host epithelial cells to initiate infection. The HA glycoprotein mediates host-cell recognition and is therefore an important determinant of species tropism. Human viral HA preferentially recognizes glycans terminated by NeuAcα2-6Gal, whereas avian HA preferentially recognizes glycans containing NeuAcα2-3Gal. Likewise, the upper airway epithelial cells in humans contain mainly NeuAcα2-6Gal, whereas in birds both the airways and intestine contain mainly NeuAcα2-3Gal linkages (see Chapters 13 and 14, and the cover figure).

In view of the rapid geographic spread of avian influenza A subtypes such as H5N1 and the increasing numbers of confirmed human cases, it is critical to survey potential influenza outbreaks and monitor human adaptation, which is a key step in the emergence of a pandemic virus. Glycan array technologies have proven to be powerful tools for this purpose. Arrays patterned with sialylated glycans of various linkages and topologies have been generated and screened with viral HA proteins as well as intact viruses. The microarray studies revealed striking glycan binding preferences that were governed not only by the sialic acid linkage and glycan modifications such as fucosylation or sulfation but also by underlying glycan structures. Once integrated into a portable format, glycan arrays may be employed in the field for influenza surveillance, which is a very practical application of functional glycomics techniques.

THE INFORMATICS CHALLENGES OF DIVERSE GLYCOMIC DATA

As mentioned above, a complete picture of the glycome can only be assembled using both top–down and bottom–up approaches. The structures of the individual glycans, their assembly on proteins and lipids, their distribution on cells and tissues, and their relation to each other on cells and within the extracellular matrix all warrant interrogation at the systems level. However, each corresponding experimental platform produces very different types of data. Mass spectral data, which can guide the assignment of a glycan's primary sequence, have a different form than lectin microarray data. The integration of disparate forms of data to generate a comprehensive picture of the glycome is a major frontier in informatics associated with glycomics research.

A comprehensive systems-level analysis would correlate data that define the glycome at various hierarchical levels with data derived from transcriptomic and proteomic experiments. One would like to know, for example, how the relative expression levels of genes that encode glycan biosynthetic enzymes compare to the glycome observed in a cell or tissue type. The effects of pharmacological agents or gene knockdowns on the glycome as compared to the transcriptome and proteome are also of interest. The functional consequences of perturbing a cell's glycome on its interactions with other cells might also be cataloged and correlated with additional systems-level data.

At present, we do not have a clear picture of how expression levels of glycan biosynthesis and processing genes relate, at the systems level, to the composition of the glycome. Efforts to correlate large data sets obtained from glycomic, transcriptomic, genomic, and proteomic studies have met with several challenges. Representation of glycan chemical structures is difficult because of their complexity and branching patterns. The use of single alphabet codes, as employed to describe nucleic acid and amino acid sequences, is not applicable to glycans. Rather than the conventional character-based codes used for

sequence information in transcriptomic and proteomic data sets, numerical or object-based codes are better suited to link the complex glycan structure information in various glycomic data sets to each other and, ultimately, to the transcriptomic and proteomic data sets. The field is in need of a comprehensive bioinformatics platform that stores, integrates, and processes data from glycomic and other "omic" studies and disseminates them in a meaningful fashion via the Internet to the scientific community.

In recent years, academic and commercial organizations have made a significant effort toward building new databases and bioinformatics platforms that fulfill this goal (GlycoSuiteDB, Sweet, KEGG GLYCAN). International organizations have formed to develop community resources. The Consortium for Functional Glycomics (CFG), EuroCarb, and the Japanese Glycomics Consortia are collaborating to develop technologies for advancing glycomics. These collaborative efforts have resulted in the development of novel experimental resources as well as online searchable databases.

With growing effort directed toward new technology and bioinformatics platforms, the future of comprehensive "omics" science is bright and potentially exciting. The reformatting of existing genomic and proteomic data sets for compatibility with emerging glycomic data sets is under way and stand-alone glycan structure databases are already in place. Once achieved, the integration of large data sets that link the glycome, genome, transcriptome, and proteome will generate a wealth of hypotheses for pursuit by future generations of scientists.

FURTHER READING

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

Campbell C.T. and Yarema K.J. 2005. Large-scale approaches for glycobiology. *Genome Biol.* **6:** 236.

Paulson J.C., Blixt O., and Collins B.E. 2006. Sweet spots in functional glycomics. *Nat. Chem. Biol.* **2:** 238–248.

Prescher J.A. and Bertozzi C.R. 2006. Chemical technologies for probing glycans. *Cell* **126:** 851–854.

Raman R., Venkataraman M., Ramakrishnan S., Lang W., Raguram S., and Sasisekharan R. 2006. Advancing glycomics: Implementation strategies at the consortium for functional glycomics. *Glycobiology* **16:** 82R–90R.

Sasisekharan R., Raman R., and Prabhakar V. 2006. Glycomics approach to structure–function relationships of glycosaminoglycans. *Annu. Rev. Biomed. Eng.* **8:** 181–231.

Pilobello K.T. and Mahal L.K. 2007. Deciphering the glycocode: The complexity and analytical challenge of glycomics. *Curr. Opin. Chem. Biol.* **11:** 300–305.

Timmer M.S., Stocker B.L., and Seeberger P.H. 2007. Probing glycomics. *Curr. Opin. Chem. Biol.* **11:** 59–65.

Turnbull J.E. and Field R.A. 2007. Emerging glycomics technologies. *Nat. Chem. Biol.* **3:** 74–77.

Mahal L.K. 2008. Glycomics: Towards bioinformatic approaches to understanding glycosylation. *Anticancer Agents Med. Chem.* **8:** 37–51.