



**HAL**  
open science

## Discovery and Classification of Glycan-Binding Proteins

Maureen Taylor, Kurt Drickamer, Anne Imberty, Yvette van Kooyk, Ronald Schnaar, Marilyn Etzler, Ajit Varki

► **To cite this version:**

Maureen Taylor, Kurt Drickamer, Anne Imberty, Yvette van Kooyk, Ronald Schnaar, et al.. Discovery and Classification of Glycan-Binding Proteins. Essentials of Glycobiology, 4th edition, 2022, 10.1101/glycobiology.4e.28 . hal-03663250

**HAL Id: hal-03663250**

**<https://hal.science/hal-03663250v1>**

Submitted on 10 May 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

NCBI Bookshelf. A service of the National Library of Medicine, National Institutes of Health.

Varki A, Cummings RD, Esko JD, et al., editors. Essentials of Glycobiology [Internet]. 4th edition. Cold Spring Harbor (NY): Cold Spring Harbor Laboratory Press; 2022. doi: 10.1101/glycobiology.4e.28

## Discovery and Classification of Glycan-Binding Proteins

### Chapter 28

#### Discovery and Classification of Glycan-Binding Proteins

Maureen E. Taylor, Kurt Drickamer, Anne Imberty, Yvette van Kooyk, Ronald L. Schnaar, Marilynn E. Etzler, and Ajit Varki

10.1101/glycobiology.4e.28

This chapter provides an overview of naturally occurring glycan-binding proteins (GBPs), the history of their discovery, some of their biological functions, ways in which GBPs are identified, and challenges in defining their biologically relevant ligands. The chapters that follow describe the analysis of glycan–protein interactions ([Chapter 29](#)), the physical principles involved ([Chapter 30](#)), and the structures and biological functions of important subclasses of GBPs ([Chapters 31–38](#)).

## TWO DISTINCT CLASSES OF GBPs

GBPs (which exclude glycan-specific antibodies) are found in all living organisms, and fall into two overarching groups—lectins and sulfated glycosaminoglycan (GAG)-binding proteins (compared in Online Appendix 28A). Lectins are further classified into evolutionarily-related families identified by carbohydrate-recognition domains (CRDs) based on primary and/or three-dimensional structural similarities ([Figure 28.1](#)). CRDs can exist as stand-alone proteins or as domains within larger multidomain proteins. They typically recognize terminal groups on glycans, which fit into shallow but well-defined binding pockets ([Chapters 29](#) and [30](#)). In contrast, proteins that bind to sulfated GAGs (heparan, chondroitin, dermatan, and keratan sulfates; [Chapter 17](#)) do so via clusters of positively charged amino acids that bind specific arrangements of carboxylic acid and sulfate groups along GAG chains. Most of these proteins are evolutionarily unrelated. GBPs that bind to the nonsulfated GAG hyaluronic acid (hyaladherins) share an evolutionarily conserved fold that facilitates recognition of short segments of the invariant hyaluronan repeating disaccharide ([Chapter 16](#)), so they are best classified as lectins rather than grouped with sulfated GAG-binding proteins. The rest of this chapter emphasizes lectins, different families of which are detailed in [Chapters 31–37](#). Sulfated GAG-binding proteins are also discussed, and are detailed further in [Chapter 38](#).

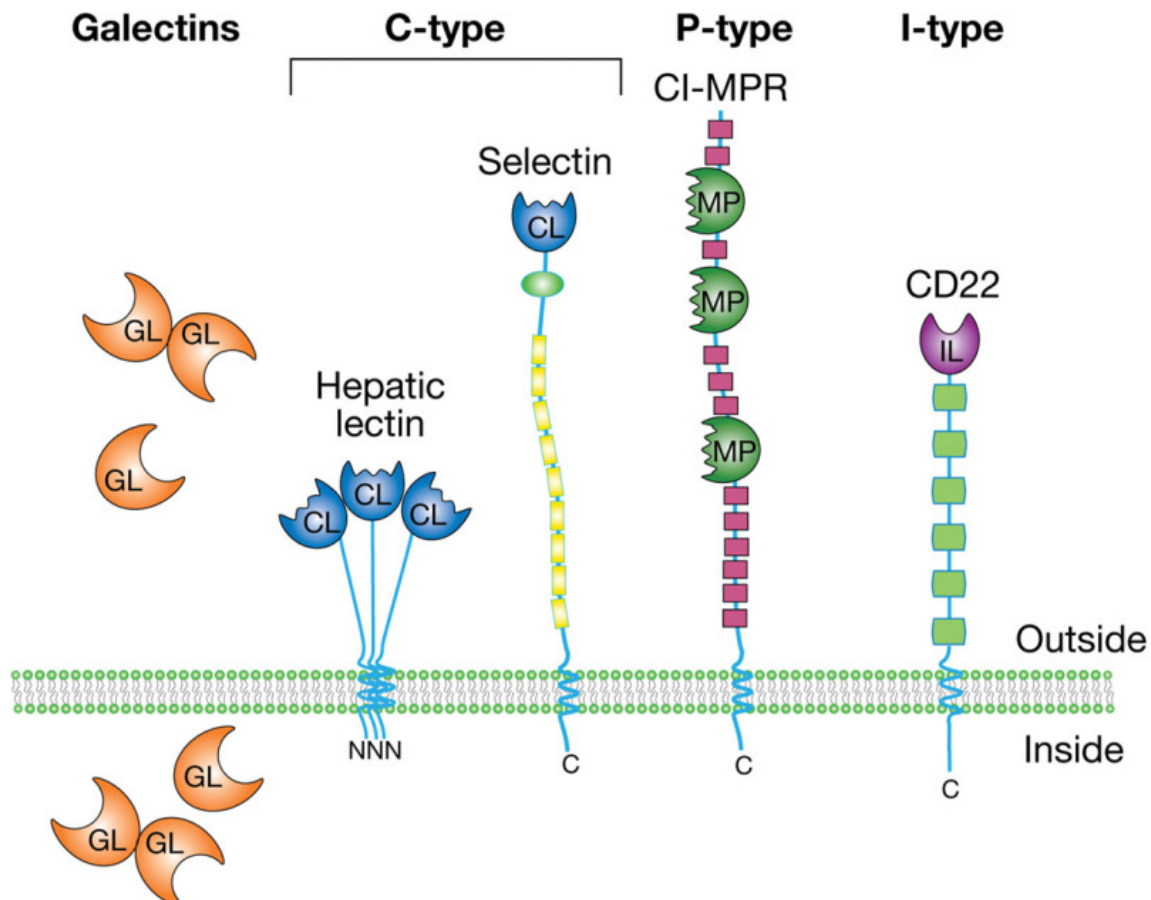


FIGURE 28.1. Representative structures from four common animal lectin families. The emphasis is on the extracellular domain structure and topology. The following are the defined carbohydrate-binding domains (CRDs) shown: (CL) C-type lectin; (GL) galectin; (MP) P-type lectin; (IL) I-type lectin. Other domains may include EGF-like domains; immunoglobulin C2-set domains; transmembrane domains; and complement regulatory repeats, depending on the family and specific family member. The number of domains accompanying the CRD varies among family members.

## DISCOVERY AND HISTORY OF LECTINS

Lectins were discovered in plants in 1888 when extracts of castor bean seeds were found to agglutinate animal red blood cells. Subsequently, seeds of many plants were found to contain such “agglutinins,” later renamed lectins (Latin for “select”) when they were found to distinguish human ABO blood groups (Chapter 14), important for blood transfusions. Lectins are particularly common in the seeds of leguminous plants and these “L-type” lectins, including concanavalin A and phytohemagglutinin, have been extensively studied. Although their specific glycan-binding activities make such plant lectins extremely useful scientific tools, their biological functions in plants remain mostly unknown.

The first animal lectin discovered was the asialoglycoprotein receptor (ASGPR) identified by Anatol Morell and Gilbert Ashwell in the late 1960s during investigations of the turnover of a serum glycoprotein, ceruloplasmin. Like most glycoproteins circulating in blood, ceruloplasmin has complex N-glycans with sialic acid termini. To prepare radiolabeled ceruloplasmin, the terminal sialic acids were removed, leaving an exposed galactose. Surprisingly, asialoceruloplasmin had a circulation half-life (in rabbits) of minutes whereas intact ceruloplasmin remained in the blood for hours. Glycoproteins with exposed Gal residues were rapidly cleared into liver cells via an endocytic cell-surface receptor that specifically bound to terminal  $\beta$ -linked Gal or GalNAc. ASGPR was purified by affinity chromatography using a column of immobilized asialoglycoprotein.

Other glycan-specific receptors involved in glycoprotein clearance and targeting were subsequently discovered, including mannose 6-phosphate receptors for targeting lysosomal enzymes to the lysosomes and mannose receptors that clear glycoproteins with terminal mannose or GlcNAc residues from the blood. Small soluble lectins specific for  $\beta$ -linked galactose (now called “galectins”; [Chapter 36](#)) were isolated by affinity chromatography of extracts from many biological sources ranging from the slime mold *Dictyostelium discoideum* to mammalian tissues. By the 1980s, the concept of vertebrate lectins that recognize specific glycans was well established. Although the first animal lectins identified were specific for endogenous glycans, many lectins specific for exogenous glycans of microorganisms were later found. Lectins recognizing exogenous glycans include soluble proteins that circulate in the blood of many species as well as membrane-bound receptors on cells of the immune system.

Lectins and sulfated GAG-binding proteins are also widespread in microorganisms, although they tend to be called by other names such as hemagglutinins and adhesins. The influenza virus hemagglutinin, which binds to sialic acid on host cells ([Chapter 15](#)) was the first GBP isolated from a microorganism. The viral hemagglutinins, like many plant lectins, can agglutinate red blood cells. Many bacterial lectins have been described. They fall into two general classes: adhesins on bacterial surfaces that recognize glycans on host cell surface glycolipids, glycoproteins, or GAGs to facilitate bacterial adhesion and colonization, and secreted bacterial toxins that bind to host cell surface glycolipids or glycoproteins ([Chapter 37](#)).

## DISCOVERY OF SULFATED GAG-BINDING PROTEINS

A large group of GBPs that defy classification based on sequence or structure recognize sulfated GAGs ([Chapter 38](#)). The best-studied example is the interaction of heparin with antithrombin. Heparin was discovered in 1916 by Jay McLean, as a medical student, but it was not until 1939 that heparin was shown to be an anticoagulant in the presence of “heparin cofactor,” which was then identified as antithrombin in the 1950s. Many other sulfated GAG-binding proteins were later discovered by affinity chromatography using columns of immobilized heparin. Growth factors and cytokines bearing clusters of positively charged amino acids along their protein surface interact with sulfated GAGs in a looser fashion—that is, they do not always show the high specificity seen with antithrombin. However, in some cases, specific GAG sequences mediate the formation of higher-order complexes, acting as a template for oligomerization or positioning of proteins such as fibroblast growth factor (FGF) and its cell-surface receptor.

## MAJOR BIOLOGICAL FUNCTIONS OF GBPs

GBPs function in communication between cells in multicellular organisms and in interactions between microbes and hosts and can also be involved in binding growth factors or cytokines. These interactions can take various forms, resulting in movement of molecules, cells, and information.

### Trafficking, Targeting, and Clearance of Proteins

Directing movement of glycoproteins within and between cells is a common function for lectins in many organisms. In eukaryotic cells, including yeast as well as higher eukaryotes, several groups of lectins are important in glycoprotein biosynthesis and intracellular movement ([Chapter 39](#)). In the endoplasmic reticulum (ER), two lectins, calnexin and calreticulin bind monoglucosylated high-mannose glycans present on newly synthesized glycoproteins, forming part of a quality control system for protein folding. This binding keeps proteins in the ER until they are correctly folded. Other groups of lectins in the ER, including M-type lectins and proteins containing mannose 6-phosphate receptor homology domains take part in the process of ER-associated glycoprotein degradation (ERAD), binding partially processed high-mannose glycans on terminally misfolded glycoproteins, causing them to be retrotranslocated into the cytoplasm for deglycosylation, followed by degradation in the proteasome. One of the best characterized functions of GBPs is in delivery of newly synthesized lysosomal enzymes from the *trans*-Golgi to lysosomes. P-type lectins ([Chapter 33](#)) recognize mannose 6-phosphate residues that have been added to N-glycans on lysosomal enzymes in the Golgi apparatus, targeting them to endosomes for fusion with lysosomes.

Once released from cells, glycoproteins can also be taken up for degradation in lysosomes. As noted above, the ASGPR on mammalian hepatocytes controls turnover of many serum glycoproteins by recognition of terminal Gal or GalNAc residues. Similarly, the mannose receptor on macrophages and sinusoidal cells of the liver binds and clears glycoproteins with oligomannose N-glycans that are released from cells during inflammation and tissue damage.

Not all lectin-mediated targeting leads to degradation. Glycan-binding subunits of secreted bacterial and plant toxins typically target them to glycolipids on cell surfaces and facilitate entry of the toxins into cells ([Chapter 37](#)). Many enzymes contain glycan-binding domains that bring another domain with enzyme activity into close proximity with its substrates. One notable group includes bacterial cellulases in which cellulose-binding modules position the enzymatic domain for optimal degradation of cellulose fibers. Using a similar principle, GalNAc-binding domains in polypeptide-N-acetylgalactosaminyltransferases that initiate O-linked glycosylation in animals position these enzymes to add further GalNAc residues to regions of polypeptides that already bear O-glycans ([Chapter 10](#)).

### Cell Adhesion

Distinctive glycans on the surfaces of different eukaryotic and prokaryotic cells make them targets for GBPs. Binding of glycans on the surface of one cell by GBPs on another cell can induce recognition and adhesion, whereas crosslinking glycans on different cells by multivalent soluble lectins provides an alternative mechanism. Such interactions are exploited in specialized

situations exemplified by transient contacts between moving cells. The selectins, three receptors that function in interactions between white blood cells, platelets, and endothelia, provide the best characterized example of lectin–glycan interactions in cell–cell adhesion ([Chapter 34](#)). For example, L-selectin on lymphocytes binds glycans on specialized endothelial cells of lymph nodes to induce lymphocyte homing, wherein circulating lymphocytes leave the blood stream and enter the lymph node. Other mammalian GBPs that mediate binding of cells to each other or that recognize ligands on the same cell surface include Siglecs ([Chapter 35](#)) and galectins ([Chapter 36](#)). Lectins in multicellular organisms also mediate interactions between cells and the extracellular matrix and support the organization of matrix components. For example, proteins containing link modules that bind specifically to hyaluronan in cartilage (and other tissues) are essential for structuring the extracellular matrix ([Chapter 38](#)), and other extracellular proteins bind to sulfated GAGs to organize cell–cell and cell–matrix interactions ([Chapter 38](#)).

Many bacteria also use lectins to adhere to glycans on host cells, often keeping them from getting washed away. These adhesins are usually present at the ends of long structures called pili or fimbriae that project from the surface of the bacteria ([Chapter 37](#)). Adhesion can be part of the infection process. For example, a mannose-specific adhesin on pathogenic strains of *Escherichia coli* that cause urinary infections binds to epithelial cells of the urinary tract. Other glycan–protein interactions between host cells and bacteria provide a mechanism for coexistence. Several bacterial species that are part of the normal gut flora, including nonpathogenic *E. coli*, use adhesins to bind to glycolipids present on cells lining the large intestine.

## Immunity and Infection

Many lectins are involved in immune responses, in invertebrates as well as in “lower” vertebrates and mammals. Differences between glycans on host and microbial cell surfaces are commonly the basis for innate immune responses. Phagocytosis is a common outcome of the binding of macrophage lectins to non-host glycans on bacteria, parasites and fungi, but many of these macrophage lectins, like DC-SIGN also recognize host-glycans on viruses for phagocytosis ([Chapters 42 & 43](#)). Other lectins circulating in the blood, such as serum mannose-binding protein and ficolins, bind to pathogen cell surfaces and activate the complement cascade, leading to complement-mediated killing.

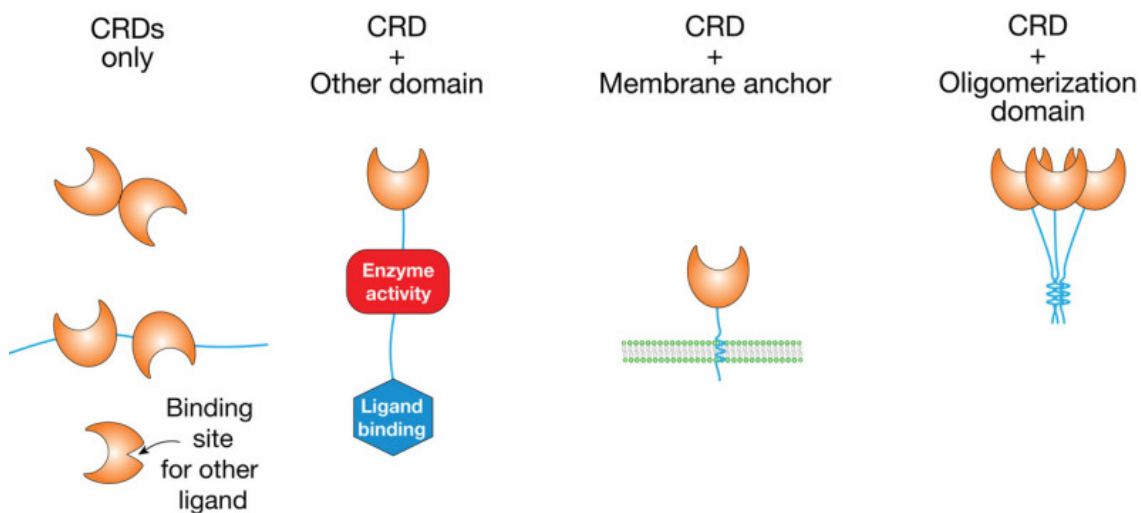
Binding of glycans to lectins on immune cells can also trigger intracellular signaling that activates or suppresses cellular responses. Receptors that recognize self-glycans such as sialic acid, as well as several that are specific for glycans characteristic of microorganisms can initiate such signaling. For example, binding of  $\alpha$ 2-6 linked Sia to CD22, a member of the Siglec family of vertebrate lectins found on B-lymphocytes, initiates signaling that inhibits activation to prevent self-reactivity ([Chapter 35](#)). In contrast, binding of trehalose dimycolate, a glycolipid found in the cell wall of *Mycobacterium tuberculosis* by the macrophage C-type lectin Mincle, induces a signaling pathway that causes the macrophage to secrete proinflammatory cytokines.

Finally, viruses often use their own GBPs to attach to host cells during infection ([Chapter 37](#)). Proteins on virus surfaces, including those on influenza virus, reovirus, Sendai virus, and polyomavirus, bind to sialic acids. In addition to bringing the virus into contact with their cell targets, these hemagglutinins typically induce membrane fusion, facilitating virus entry and delivery of nucleic acids into the cytosol. Glycan-binding receptors on viruses are often highly specific for a particular linkage; human influenza virus binds to sialic acids linked  $\alpha$ 2-6 to Gal,

whereas bird influenza virus binds to  $\alpha$ 2-3-linked sialic acid. Other viruses, such as herpes simplex virus, have adhesins that bind to heparan sulfate proteoglycans on cell surfaces.

## ORGANIZATION OF LECTINS

An important concept in identifying, defining, and classifying lectins is that glycan-binding activity is embodied in discrete protein modules or domains, referred to as CRDs. CRDs are typically independently folding segments of proteins; often one can separate the glycan-binding activity from other activities of the protein by expressing its CRD in isolation. In some cases, the CRDs constitute the entire GBP (Figure 28.2).



*FIGURE 28.2. Arrangements of carbohydrate-recognition domains (CRDs) in lectins. Proteins containing just CRDs or CRDs associated with other types of functional domains, with membrane anchors or with oligomerization domains, are depicted schematically. A single lectin can contain all of these additional domains.*

When a lectin is comprised simply of its CRD, its functions often are dependent on multivalency, which endows lectins with the ability to cross-link glycan-containing structures. This arrangement explains the ability of many plant lectins to agglutinate cells and to cluster glycoproteins on cell surfaces, which can induce mitogenesis. Other GBPs that function this way include the galectins, which can bridge glycans on one cell surface or between cells. Sometimes other activities are encoded within the structure of the same domain that binds glycans; some cytokines comprised of a single folded domain may have distinct sites for binding glycans and other target receptors. More commonly, other activities of lectins reside in separate modules in multidomain proteins (Figure 28.2). Such arrangements are widespread. The domains associated with CRDs perform many different functions, including binding other types of ligands, performing enzymatic reactions, anchoring proteins to membranes, and directing oligomerization. GBPs often contain multiple modules, combining several functions in one protein.

Membrane anchors in lectins take multiple forms, but they often span the membrane, linking extracellular CRDs with cytoplasmic domains. This arrangement facilitates the flow of information between glycan-binding sites on the extracellular surface and the cytoplasm. Simple sequence motifs in the cytoplasmic domains of transmembrane lectins often control trafficking of receptors and their bound glycan ligands. Common functions of such intracellular movements are internalization of cell-surface receptors, directing bound ligands to endosomes and lysosomes, and movement through intracellular compartments such as the ER and Golgi apparatus to the cell surface. Flow of information in the opposite direction can lead to stimulation of signaling complexes on the cytoplasmic side of the membrane in response to binding of glycans at the cell surface.

Clustering of glycan-binding sites (multivalency) is often critical to both recognition and biological functions and is achieved in different ways: by formation of simple oligomers of CRDs, as a result of the presence of multiple CRDs in a single receptor polypeptide, and through association of CRD-containing polypeptides via independent oligomerization domains. Some oligomers are stable, whereas others, such as those formed by some galectins, are in equilibrium with monomers. These arrangements facilitate multivalent binding to increase avidity and direct the geometrical arrangement of binding sites. Multiple CRDs may face in the same direction for surface recognition or in opposite directions to facilitate crosslinking. Multivalent CRDs may have fixed spacing or flexible spacing to accommodate different target glycans. In some cases, oligomerization domains also form structural features, serving as stalks that project CRDs from the cell surface. Oligomerization domains can also embody other functions, such as the protease-binding sites in the collagen-like domains of mannose-binding protein.

## CLASSIFICATION OF LECTINS BASED ON STRUCTURAL SIMILARITIES

It is convenient to classify lectins based on the structures of the CRDs that they contain ([Figure 28.3](#)). CRDs are found in a large number of different structural categories, indicating that many different protein folds can accommodate glycan binding ([Chapter 30](#)). Based on this observation, glycan recognition must have evolved independently many times and the diversity of CRD structures must have arisen to address a diversity of functions.

GBPs appear across all kingdoms of life, but the types of lectins in each kingdom vary considerably. Several families appear in both prokaryotes and eukaryotes, but their distributions suggest different evolutionary histories. The malectin domain, although conserved in structure and widely distributed in prokaryotes, plants, and animals, is found in proteins with distinct domain organization and different functions in the three groups. Animal malectin is a membrane-anchored CRD of the ER that binds N-linked glycans during glycoprotein biosynthesis. In plants, the malectin CRD is expressed at the cell surface and is linked to a cytoplasmic kinase domain. Bacterial malectins consist of CRDs associated with glycohydrolase domains. Similarly, R-type CRDs ([Chapter 31](#)) in plants form the cell-surface-binding component of toxins such as ricin and are linked to glycohydrolase genes in bacteria, but in animals they appear in two distinct contexts. In polypeptide-N-acetylgalactosaminyltransferases that initiate O-GalNAc glycans ([Chapter 10](#)) and in the mannose receptor family. Although these CRDs have been adapted to serve different



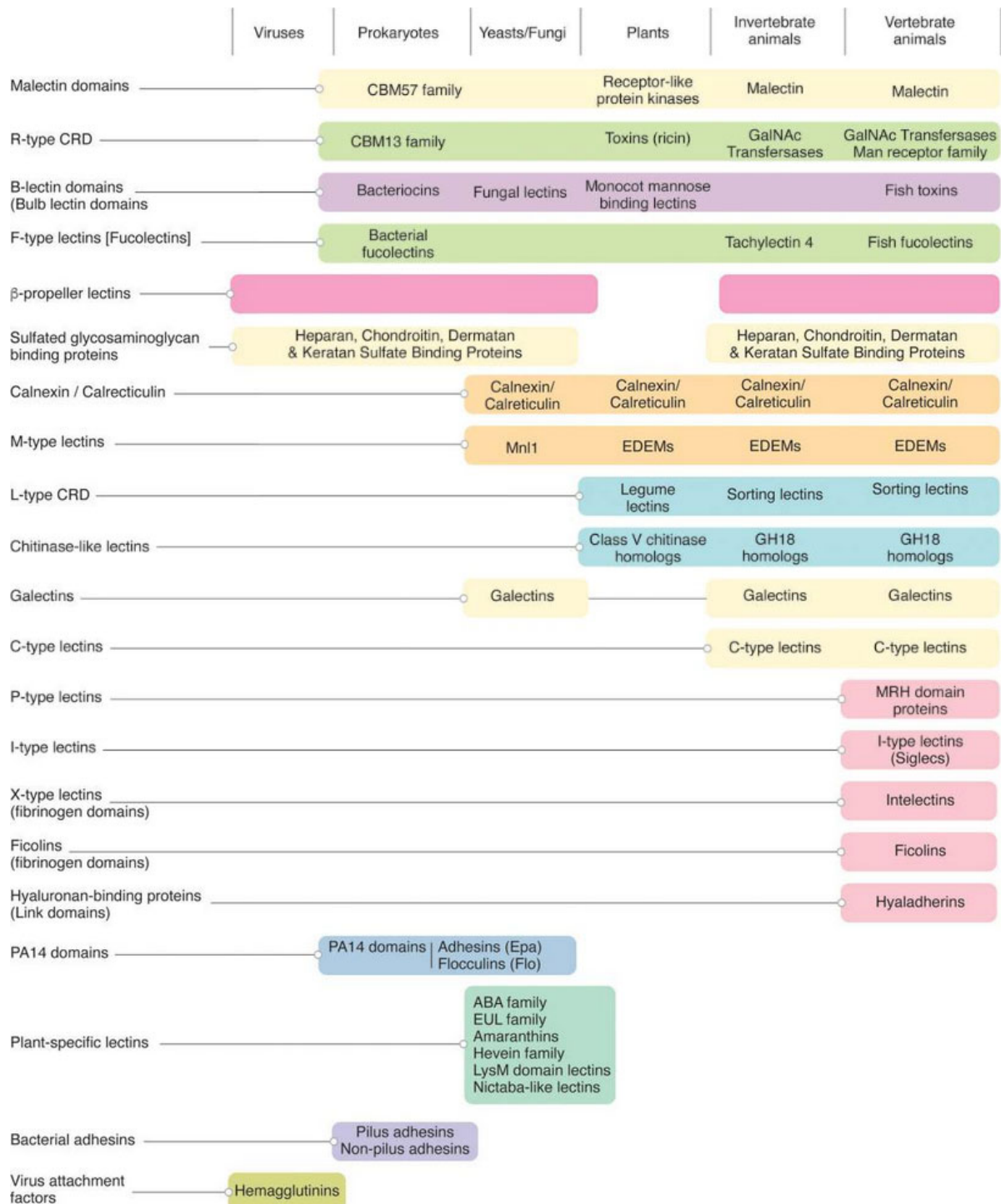


FIGURE 28.3. Several major structural families of glycan-binding proteins (GBPs) and their biological distributions. CBM, carbohydrate-binding module; EUL, *Euonymus europaeus* lectin; ABA, *Agaricus bisporus* agglutinin; EDEM, ER degradation-enhancing  $\alpha$ -mannosidase-like proteins; GH, glycohydrolase; MRH, mannose receptor homology.

functions in different kingdoms, a glycan-binding function appears to have evolved early and been preserved in subsequent lineages.

In contrast to CRDs with broad evolutionary distribution, two other groups of lectins have sporadic distributions. B-lectin domains, which are broadly distributed in bacteria in association with hydrolase domains, are found as isolated or tandem CRDs in monocot plants but not in other plants, in bony fishes but not in other animals, and in some fungi. The recently identified F-type lectins appear in a few bacterial species and several lower vertebrates, but have not been found in mammals. In these cases, the presence of related domains in evolutionarily distant species may reflect lateral gene transfer rather than the presence of a precursor lectin in the distant common ancestor that they share. A different pattern of evolution is observed for PA14 domains, the only other type of CRD found in both bacteria and eukaryotes. Although the PA14 fold is relatively widespread, suggesting that it originated early and was retained across species, only a subset have been shown to have glycan-binding activity: CRDs associated with bacterial glycohydrolases and in adhesins and flocculation factors on the surface of yeast. Similarly,  $\beta$ -propeller lectins have evolved in multiple contexts due to their high avidity for glycans present on cells surfaces, and are present in bacteria, fungi, and animals.

The intracellular sorting lectins mentioned earlier, such as calnexin, calreticulin, and M-type lectins, are the most broadly distributed lectins that evolved from a common eukaryotic ancestor. Their distribution and the conservation of their functions probably reflect an ancient and conserved role in intracellular trafficking of glycoproteins in eukaryotes. Two other groups of CRDs appear to be found in metazoans but not simpler eukaryotes. The L-type CRDs have diverged in function between animals, in which they function in intracellular glycoprotein sorting and trafficking, and plants, in which they serve a protective function ([Chapter 32](#)). Chitinase-like glycan-binding domains across a range of species retain the ability to bind polymers of GlcNAc, but their biological functions are not well understood, so it is unclear if they have shared roles in plants and animals.

In addition to the widely distributed families, certain CRD families are evolutionarily restricted. In addition to animal-specific and vertebrate-specific lectin groups, there are also groups such as the I-type lectins found only in mammals ([Chapter 35](#)). The pattern of evolution of animal-specific lectins varies. Galectins seem to be similar in organization in vertebrates and invertebrates and it may be possible to identify orthologs in quite diverse species ([Chapter 36](#)). In contrast, C-type CRDs have undergone independent radiation in vertebrates and invertebrates, and identifying orthologs even between mouse and human proteins in some cases is difficult ([Chapter 34](#)). Of the 12 different protein folds found in plant lectins, six appear to be unique to plants. It is also noteworthy that viruses seem to have developed their own approaches to binding glycans rather than borrowing from hosts ([Chapter 37](#)).

In addition to families of proteins that share evolutionarily related CRDs, there are individual proteins that bind glycans through domains that are not related to CRDs in other proteins. Examples include proteins with dedicated glycan-binding domains, such as some laminin G domains that recognize glycans on  $\alpha$ -dystroglycan ([Chapter 45](#)), pentraxins that bind modified and phosphorylated glycans, and macrophage  $\alpha_M\beta_2$  integrin that binds fungal glucans and exposed GlcNAc residues on glycoproteins. Other proteins bind to glycans through domains that also have other ligands: annexin V binds bisecting GlcNAc residues as well as phospholipids, and several cytokines have been reported to bind glycans as well as nonglycan receptors. Sulfated-GAG-binding proteins have also largely evolved by convergent evolution.

## IDENTIFYING GBPs BY BIOLOGICAL AND BIOCHEMICAL FUNCTION AND STRUCTURAL SIMILARITY

There are multiple ways in which glycan recognition can be implicated in specific biological processes. One common approach is to show the ability of simple monosaccharides or small glycans to compete with a process. Information can often also be gained by modifying glycans on cells and glycoproteins with enzymes that add or remove glycans, by genetic manipulation, and by chemical inhibitors of glycan metabolism. These strategies have provided information about the glycans involved—for example, those needed for virus or toxin binding or those required for endocytosis of glycoproteins. Based on this information, it is then possible to look for GBPs that target these particular glycans, which can then be linked to the biological process.

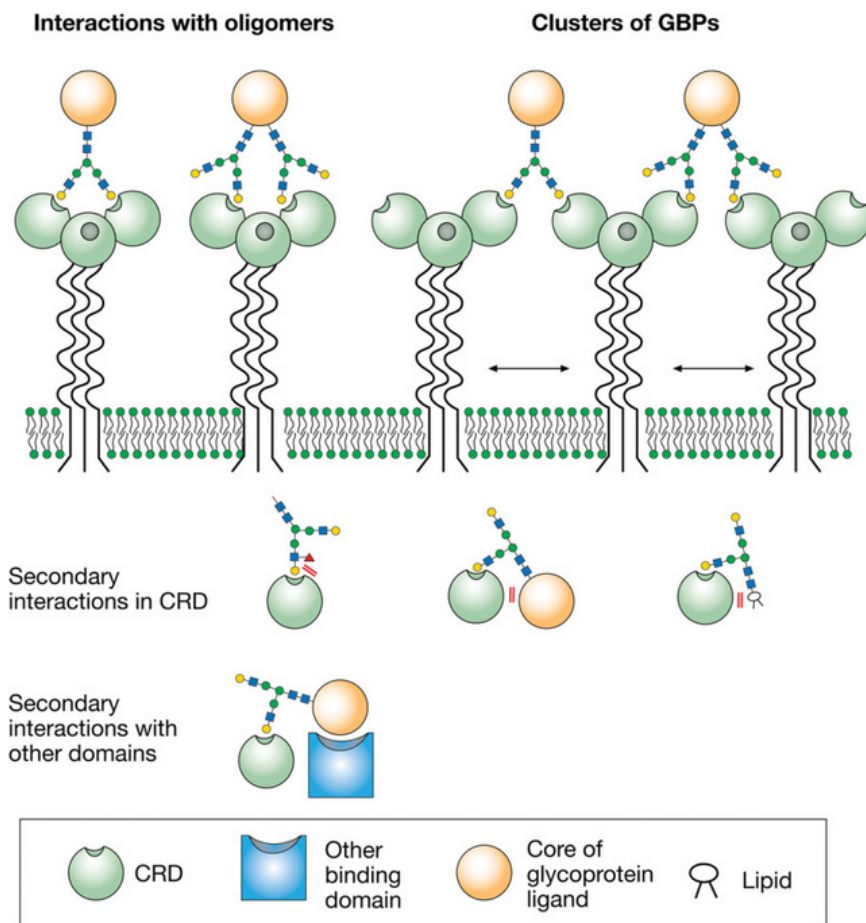
The ability to bind specific glycans, assessed in various biochemical assays, has often been the basis for direct identification of novel GBPs without reference to a particular biological function. For example, galectins ([Chapter 36](#)) share a binding preference for  $\beta$ -galactosides and F-type lectins for fucosyl residues. In addition to forming a basis for binding and competition assays, the binding activity is commonly used as a means of isolating these proteins by using affinity chromatography with appropriate immobilized glycan ligands. A wide variety of methods for coupling monosaccharides and complex glycans to create affinity resins have been developed. As mentioned above, many sulfated-GAG-binding proteins have been discovered by affinity chromatography using immobilized GAG chains. A limitation of these approaches is that binding activity does not directly indicate a biological function and the roles of many well-characterized GBPs have not been fully determined.

The observation that many lectins fall into structural families provides an alternative way to identify novel GBPs through analysis of protein sequences. Sequence motifs characteristic of CRDs are routinely used to screen sequences from whole genome sequencing. These motifs can also be used to screen specific cDNA and gene sequences of interest because of their association with biological functions. Detection of an appropriate motif suggests the presence of a CRD, and structural knowledge of known glycan-binding sites can suggest whether a novel protein is likely to retain glycan-binding activity. In some cases, it can even suggest potential ligands. Such predictions often motivate testing for glycan-binding activity, either by specifically examining binding to predicted ligands or by screening more generally using glycan arrays.

Although structure-based predictions do not directly yield information about biological function, the organization of CRDs and their association with other domains often provide information about potential functions. This type of top-down analysis is limited to discovery of GBPs that contain domains resembling known CRDs. As glycan array screening becomes more widely accessible, more broad-based screening can be envisioned.

## GLYCAN LIGANDS FOR LECTINS

Monosaccharides or small oligosaccharides in isolation tend to be low-affinity ligands for GBPs, often with dissociation constants in the millimolar range. These intrinsic affinities are enhanced in several ways (Figure 28.4). At the level of individual glycans, affinity can be enhanced by linking the glycan to other types of structures. Conjugation of glycans to proteins and lipids can lead to enhanced CRD binding. For example, some GBPs such as the macrophage receptor Mincle bind to glycolipids with much higher affinity than they bind to free oligosaccharides. In this case, enhanced affinity can result from the presence of an extended or accessory binding site in a CRD adjacent to the glycan-binding site, which is able to accommodate the hydrophobic tail of the lipid. Other GBPs bind selectively to a particular glycan conjugated to a specific polypeptide motif. Optimal binding of P-selectin to the ligand PSGL-1 requires an O-linked glycan bearing a sialyl Lewis x structure on a peptide with adjacent acidic residues and sulfated tyrosines (Chapter 34). In yet other cases, glycan recognition is combined with other binding domains on a protein. The mannose receptor contains C-type CRDs that bind high-mannose oligosaccharides and a fibronectin type II repeat that binds to triple helical polypeptides. Together, these two modalities facilitate binding to fragments of collagen released at sites of inflammation.



*FIGURE 28.4. Mechanisms of enhanced binding of natural ligands to lectins. Within individual carbohydrate-binding domains (CRDs), secondary interactions beyond the primary binding site can be with glycan, protein, or lipid portions of glycoconjugate ligands. Multivalent interactions can reflect interaction of single branched oligosaccharides or multiple oligosaccharides attached to a glycoprotein with multiple CRDs brought together within receptor oligomers or in glycan-binding protein (GBP) clusters on the cell surface.*

A major determinant of binding to natural lectin ligands is often the interaction of multivalent glycans with clustered CRDs, resulting in high-avidity binding. Ligand clustering can result from multiple binding epitopes in a single oligosaccharide or polysaccharide, multiple glycans attached to a single protein scaffold or the presence of adjacent glycoproteins or glycolipids in a cell membrane. Similarly, CRD clustering can reflect multiple CRDs in a single polypeptide, formation of polypeptide oligomers that each contains a single CRD, or from clustering of CRD-containing proteins in the cell membrane. Each of these levels of CRD organization has the potential to place geometrical constraints on the optimal arrangement of ligands, depending on the degree to which CRDs are held in a fixed arrangement or are flexibly linked. Clustering of glycans attached to a single polypeptide, particularly in heavily O-glycosylated proteins such as mucins, can also affect their ability to take on different conformations. Because lectins typically interact with a single conformation, there is an entropic penalty associated with binding any one of these which may be reduced when the glycan has fewer potential conformations. In vitro biochemical assays, including glycan arrays, reflect only some of these types of clustering of CRDs and ligands, so they must be interpreted with some caution. In some cases, binding of a CRD to isolated glycans may be essentially undetectable even though binding of the intact CRD-containing protein to its endogenous glycoconjugate may be highly selective and quite strong. Care must also be exercised in use of the term ligand, to distinguish the glycan part of a ligand from the entire natural glycoconjugate or even cell surface.

## TERMINOLOGY FOR SPECIFIC GBP LIGANDS

Based on the above considerations, GBPs may bind optimally to a glycan only when it is conjugated to a particular protein or lipid. In such instances, the GBP ligand is neither the glycan by itself nor the carrier by itself. Examples include P-selectin, which binds to sialyl Lewis x adjacent to sulfated tyrosines on the PSGL-1 protein of leukocytes (see above) and E-selectin, which binds to the same glycan on a variant form of the protein CD44 on hematopoietic stem cells (designated HCELL). Although the concept that GBPs bind glycans in the context of their carriers is well established, there is no standard terminology to designate a particular glycoform as the ligand for a GBP. Stating that a protein (e.g., PSGL-1 or CD44) is “the ligand” for a GBP (P-Selectin and E-Selectin, respectively) is inaccurate, in that the protein without the specific glycan is not a ligand. However, giving a ligand a completely different name (e.g., HCELL) does not identify the polypeptide carrier. An option is to use a superscript “L” (ligand) to designate these molecules as PSGL-1<sup>PSL</sup> and CD44<sup>ESL</sup>, respectively. This terminology also has the advantage of distinguishing different glycoforms of the same polypeptide as ligands for different GBPs. For example, subsets of the glycoprotein CD24 recognized by P-selectin and Siglec-10 can be designated as CD24<sup>PSL</sup> and CD24<sup>S10L</sup>, respectively. Regardless of the nomenclature used, direct

proof of a functional interaction in vivo is needed to definitively assign a particular glycoprotein or glycolipid as a physiological GBP ligand.

## ACKNOWLEDGMENTS

The authors acknowledge contributions to previous versions of this chapter by Richard Cummings and Jeffrey Esko and appreciate helpful comments and suggestions from Lingquan Deng, Yuki Ohkawa, Paeton L. Wantuch, and Ryan Weiss.

## FURTHER READING

- Stillmark, H. 1888. Inaugural dissertation. Uber Ricin, Ein giftiges Ferment aus den Samen von *Ricinus communis* L. und einigen anderen Euphorbiaceen. University of Dorpat, Dorpat (now Tartu), Estonia.
- Goldstein, IJ, Hughes, RC, Monsigny, M, Osawa, T, Sharon, N. 1980. What should be called a lectin. *Nature* 285: 66.
- Ashwell, G, and Harford, J. 1982. Carbohydrate-specific receptors of the liver. *Annu Rev Biochem* 51: 531–554.
- Drickamer, K. 1988. Two distinct classes of carbohydrate-recognition domains in animal lectins. *J Biol Chem* 263: 9557–9560.
- Powell, LD, and Varki, A. 1995. I-type lectins. *J Biol Chem* 270: 14243–14246.
- Lee, RT, and Lee, YC. 2000. Affinity enhancement by multivalent lectin–carbohydrate interaction. *Glycoconj J* 17: 543–551.
- Casu, B, and Lindahl, U. 2001. Structure and biological interactions of heparin and heparan sulfate. *Adv Carbohydr Chem Biochem* 57: 159–206.
- Esko, JD, and Selleck, SB. 2002. Order out of chaos: Assembly of ligand binding sites in heparan sulfate. *Annu Rev Biochem* 71: 435–471.
- Drickamer, K, and Taylor, ME. 2003. Identification of lectins from genomic sequence data. *Methods Enzymol* 362: 560–567.
- Lee, JK, Baum, LG, Moremen, K, Pierce, M. 2004. The X-lectins: A new family with homology to the *Xenopus laevis* oocyte lectin XL-35. *Glycoconj J* 21: 443–450.
- Sharon, N, and Lis, H. 2004. History of lectins: from hemagglutinins to biological recognition molecules. *Glycobiology* 14: 53R–62R.
- Blundell, CD, Almond, A, Mahoney, DJ, DeAngelis, PL, Campbell, ID, Day, AJ. 2005. Towards a structure for a TSG-6-hyaluronan complex by modeling and NMR spectroscopy: Insights into other members of the link module superfamily. *J Biol Chem* 280: 18189–18201.
- Varki, A, and Angata, T. 2006. Siglecs—The major subfamily of I-type lectins. *Glycobiology* 16: 1R–27R.
- Van Damme, EJM, Lannoo, N, Peumans, WJ. 2008. Plant lectins. *Adv Bot Res* 48: 107–209.
- Dam, TK, Gerken, TA, Brewer, CF. 2009. Thermodynamics of multivalent carbohydrate-lectin cross-linking interactions: importance of entropy in the bind and jump mechanism. *Biochemistry* 48: 3822–3827.
- Taylor, ME, and Drickamer, K. 2009. Structural insights into what glycan arrays tell us about how glycan-binding proteins interact with their ligands. *Glycobiology* 19: 1155–1162.
- Adrangi, S, and Faramarzi, MA. 2013. From bacteria to human: A journey into the world of chitinases. *Biotechnol Adv* 31: 1786–1795.

- Gilbert, HJ, Knox, JP, Boraston, AB. 2013. Advances in understanding the molecular basis of plant cell wall polysaccharide recognition by carbohydrate-binding modules. *Curr Opin Struct Biol* 23: 669–677.
- Cohen, M, and Varki, A. 2014. Modulation of glycan recognition by clustered saccharide patches. *Int Rev Cell Mol Biol* 308: 75–125.
- Nagae, M, and Yamaguchi, Y. 2014. Three-dimensional structural aspects of protein–polysaccharide interactions. *Int J Mol Sci* 15: 3768–3783.
- Taylor, ME, and Drickamer, K. 2014. Convergent and divergent mechanisms of sugar recognition across kingdoms. *Curr Opin Struct Biol* 28C: 14–22.
- Bishnoi, R, Khatri, I, Subramanian, S, Ramya, TN. 2015. Prevalence of the F-type lectin domain. *Glycobiology* 25: 888–901.
- Drickamer, K, and Taylor, ME. 2015. Recent insights into structures and functions of C-type lectins in the immune system. *Curr Opin Struct Biol* 34: 26–34.
- Pees, B, Yang, W, Zarate-Potes, A, Schulenburg, H, Dierking, K. 2017. High innate immune specificity through diversified C-type lectin-like domain proteins in invertebrates. *J Innate Immun* 8: 129–142.
- Bonnardel, F, Mariethoz, J, Salentin, S, Robin, X, Schroeder, M, Perez, S, Lisacek, F, Imberty, A. 2019. UniLectin3D, a database of carbohydrate binding proteins with curated information on 3D structures and interacting ligands. *Nucleic Acid Res* 47: D1236–D1244