Glycobiology vol. 11 no. 5 pp. 71R–79R, 2001

MINI REVIEW

Lectin-like proteins in model organisms: implications for evolution of carbohydrate-binding activity

Roger B. Dodd[1] and Kurt Drickamer[2]

Glycobiology Institute, Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, UK

Accepted on February 17, 2001

Classes of intracellular lectins that recognize core-type structures and mediate intracellular glycoprotein trafficking are present in vertebrates, model invertebrates such as *Caenorhabditis elegans* and *Drosophila melanogaster*, plants, and yeasts. Lectins that recognize more complex structures at the cell surface, such as C-type lectins and galectins, are also found in invertebrate organisms as well as vertebrates, but the functions of these proteins have evolved differently in different animal lineages.

Key words: carbohydrate-recognition domain/genomics/evolution/lectins

Introduction

Complex oligosaccharide structures displayed at cell surfaces, incorporated into the extracellular matrix, and attached to secreted glycoproteins can serve structural roles, mediate movement of glycoconjugates to the cell surface, or act as markers that mediate cell–cell and cell-matrix recognition events. The nonstructural roles of sugars generally require the participation of sugar-binding lectins (Drickamer and Taylor, 1998). Lectins are often complex, multidomain proteins, but sugar-binding activity can usually be ascribed to a single protein module within the lectin polypeptide. Such a module is designated a carbohydrate-recognition domain (CRD). CRDs in vertebrate lectins fall into a number of structurally distinct families of protein modules. Some of the best characterized of these CRD groups are summarized in Table I. The list is by no means complete, as sugar-binding activity has also been described for other protein modules that have folds not represented in this list. For example, several structurally distinct types of proteins have been shown to bind glycosaminoglycans. Although it has been proposed that these proteins might share a common local binding motif (Cardin and Weintraub, 1989), such a motif would have to be presented in the context of many different protein structures. For reasons of space, this survey of potential CRDs in model organisms discussed is restricted to the structural categories shown in Table I.

The lectins that contain CRDs listed in the Table I fall broadly in two categories. Lectins that contain CRDs in the first three structural groups are located mostly intracellularly, in luminal compartments. They function in the trafficking, sorting, and targetting of glycoproteins in the secretory and other pathways. CRDs in the remaining structural groups are found in lectins that function largely outside the cell and are either secreted or localized to the plasma membrane.

Comparisons of CRDs in each structural group allow the generation of amino acid sequence profiles or motifs that can be used to screen sequence databases for related protein modules. The conserved residues that define these profiles generally reflect the requirements for specific amino acid residues, mostly found in the protein interior, that determine a basic protein fold. Each profile thus serves to identify protein modules that are similar in overall structure to CRDs in a particular group. In many cases, however, such modules serve functions other than sugar binding. Sugar-binding to a bona fide CRD generally occurs in a shallow indentation on the protein surface. For each type of CRD, sugar-binding activity is thus determined by a second set of residues that function within the context of the structural fold associated with that type of CRD. Screening of genomic sequences for potential lectins is thus a two-step process. First, protein modules that have CRD-like folds are identified by profile analysis. The sequences of these domains are then examined for residues that form sugar-binding sites in known lectins to allow informed speculation about which of them may actually mediate carbohydrate binding.

The genomic sequences of the single-celled yeasts *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*, the nematode *Caenorhabditis elegans* (Consortium, 1998) and the fruit fly *Drosophila melanogaster* (Adams *et al.*, 2000) can provide insight into lectin functions that can be understood by analysis of these model organisms. Using an approach previously employed for the C-type CRDs of *C. elegans* (Drickamer and Dodd, 1999), sequence databases have been screened with various CRD profiles to identify protein containing potential CRDs. Sequence alignments have then been used to compare these modules to vertebrate CRDs to identify which ones are likely to bind sugars. The alignments that form the basis for the conclusions summarized here will be found at <http://ctld.glycob.ox.ac.uk>. Further examination of the complete sequences of the proteins,

[1]Present address: Wellcome Trust Centre for the Study of Molecular Mechanisms in Disease, Cambridge University, Wellcome Trust/MRC Building, Hills Road, Cambridge CB2 2XY, UK
[2]To whom correspondence should be addressed

Table I. Summary of lectin categories

Lectin group	Structure of CRD	Motifs	Length	Typical ligands	Examples of functions
Calnexin	Unknown			Glc_1Man_9 oligosaccharides	Protein sorting in the endoplasmic reticulum
L-type	β-sandwich	PS00307	230+	Various	Protein sorting in the endoplasmic reticulum
		PS00308			
		PF00138			
		PF00139			
		IPR000985			
		IPR001220			
P-type	Unique β-rich structure	PF00878	130+	Man 6-phosphate	Protein sorting post Golgi
		PF02157			
		IPR000479			
		IPR000296			
C-type	Unique mixed α/β structure	PS50041	115+	Various	Cell adhesion (selectins)
		PS00615			Glycoprotein clearance
		PF00059			Innate immunity (collectins)
		IPR001304			
Galectins	β-sandwich	PS00309	125+	β-Galactosides	Glycan crosslinking in the extracellular matrix
		PF00337			
		IPR001079			
I-type	Immunoglobulin superfamily	PF00047	120+	Sialic acid	Cell adhesion (siglecs)
		IPR003006			
R-type	β-trefoil	PF00652	125+	Various	Enzyme targeting
		IPR000772			Glycoprotein hormone turnover

Length is given in amino acid residues. The functions listed are primarily those that have been identified in vertebrates. PS motifs are from the ProSite database (<http://www.expasy.ch/prosite/>); PF motifs are from the Pfam protein families database (<http://www.sanger.ac.uk/Software/Pfam/>); IPR motifs are from the InterPro database (<http://www.ebi.ac.uk/interpro/>).

using hydropathy plots and a comprehensive set of protein module profiles, provides insight into their overall domain organization. The presence of other types of protein modules, such as membrane anchors, often suggests possible functions of the CRD-containing proteins. Comparison of the different structural classes of lectins in vertebrates, invertebrates, yeasts, bacteria, and plants also provides a basis for understanding the coevolution of glycan structures and glycan recognition processes.

Lectins in the endoplasmic reticulum: calnexin and calreticulin

Calnexin and calreticulin form part of the quality control system for glycoproteins in the endoplasmic reticulum (Trombetta and Helenius, 1998; Parodi, 2000). They bind to terminal glucose residues on N-linked oligosaccharides and retain misfolded glycoproteins in the endoplasmic reticulum. Calnexin is a transmembrane protein and calreticulin is a soluble protein retained in the lumen by a C-terminal retention signal. The luminal N-terminal portion of calnexin is very similar to calreticulin, although one of the repeated segments of calnexin is absent from calreticulin.

Homologues of calnexin and calreticulin have been identified in unicellular as well as multicellular eukaryotes. Based on overall sequence similarity and the absence of a membrane anchor from calreticulin, the homologues identified in yeasts resemble calnexin most closely (Parlati *et al.*, 1995). Homologues of both proteins are found in *C. elegans* as well as *Drosophila*. Pairwise comparisons show close similarity between calnexins from distant species: *S. pombe* and human calnexin sequences are 43% identical (Figure 1A). Recent independent duplications in the lineages leading to mammals and *Drosophila* have generated multiple forms of the membrane-anchored protein in these species. Calnexin and calreticulin do not appear to be part of a larger family of proteins.

A key component of the glycosylation-dependent quality control system is a glucosyl transferase that reglucosylates incorrectly folded glycoproteins. *Drosophila* express a functional glucosyltransferase, suggesting that calnexin orthologues function in a protein quality control pathway in invertebrates as well as vertebrates (Parker *et al.*, 1995). The cycle is also functional in *S. pombe*, although the glucosyltransferase is absent from *S. cerevisiae* (Fernandez *et al.*, 1998; Jakob *et al.*, 1998). Thus, the deglucosylation, recognition, and reglucosylation machinery was probably present in the early unicellular

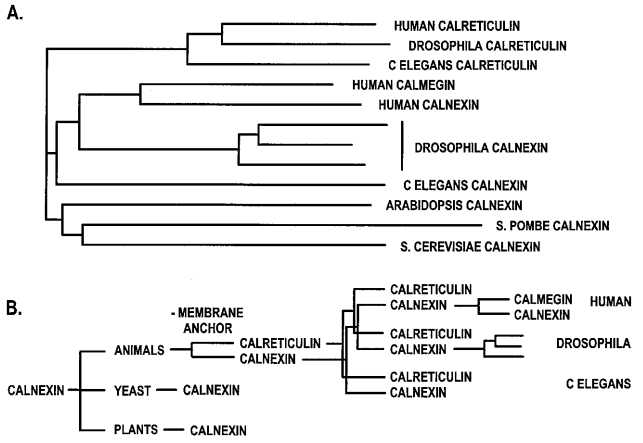

Fig. 1. Evolution of calnexin and calreticulin. (**A**) Dendrogram comparing sequences of the luminal domains of calnexin, calreticulin, and calmegin. In the absence of a suitable sequence profile, calnexin homologues in the protein sequence databases were identified by searching directly with the human protein. (**B**) Scheme summarizing evolution of calnexin and calreticulin.

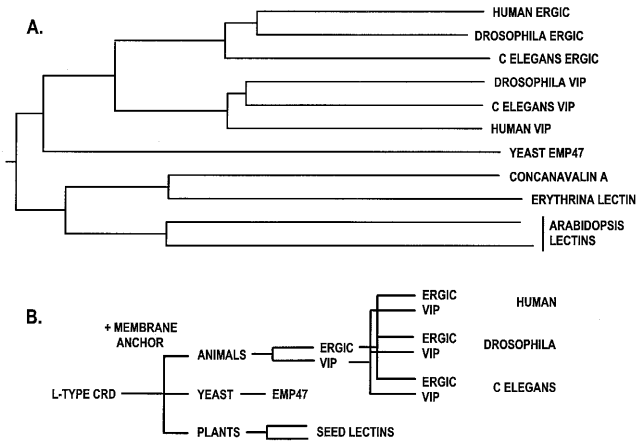

Fig. 2. Evolution of L-type lectins. (**A**) Dendrogram comparing sequences of L-type CRDs. Animal L-type lectins are not identified using the available profiles for legume lectins, so these lectins were identified by searching with the human ERGIC-53 and VIP-36 sequences and then again using relatively conserved fragments from the aligned sequences. (**B**) Scheme describing the evolution of L-type lectins.

eukaryotes that were common progenitors to plants, animals, and yeast, although it has become nonfunctional in *S. cerevisiae*. Calnexin and its relatives form an ancient and nearly universal sugar recognition system that has been highly conserved in eukaryotic cells.

L-type lectins in plants and animals

A second family of lectins involved in protein sorting in luminal compartments of animal cells is composed of two members. ERGIC-53 is localized to the endoplasmic reticulum–Golgi intermediate compartment (Itin *et al.*, 1996) and VIP-36 is found in Golgi and post-Golgi portions of the secretory pathway (Fiedler and Simons, 1994). Both ERGIC-53 and VIP-36 are type I transmembrane proteins. The luminal portions of these proteins correspond to the single folded domain of the soluble lectins found in abundance in the seeds of leguminous plants (Sharon and Lis, 1990). For this reason, they are designated L-type CRDs.

Plant and animal L-type lectins have divergent sequences and different molecular properties: the plant lectins are secreted, soluble proteins and are found at high level in specialized tissues, and the animal L-type lectins are membrane-bound luminal proteins and are found at low levels in many different cell types. These differences reflect the fact that plant and animal L-type lectins are likely to serve different functions. Nevertheless, it seems likely that the L-type CRDs have retained similar mechanisms of sugar binding. Certain key residues in four loop regions that contribute to the binding sites in the plant proteins (Sharma and Surolia, 1997; Rini, 1995) are conserved in all of the animal and plant L-type CRDs.

ERGIC-53 and VIP-36 orthologues are both found in *C. elegans* and *Drosophila* (Figure 2). These proteins contain the key sugar-binding residues and seem likely to serve sorting functions like ERGIC-53 and VIP-36. In contrast, the EMP47 gene product in *S. cerevisiae* is a relatively divergent homologue of the L-type plant and animal lectins. Virtually all of the key residues that form the sugar-binding sites in the legume lectins are absent from the yeast protein, suggesting that it probably lacks sugar-binding activity and serves a different function. The presence of an homologous protein in yeast indicates that the L-type CRD fold is ancient, but the sequence comparisons suggest that a sugar-binding L-type CRD first appeared in the common precursor of plants and animals.

Mannose 6-phosphate receptors

A third family of lectins involved in intracellular trafficking consists of two types of receptors that recognize mannose 6-phosphate residues on oligosaccharides of hydrolases that must be directed from the Golgi apparatus to their lysosomal destination. The cation-dependent mammalian mannose 6-phosphate receptor contains a single P-type CRD, and the cation-independent receptor contains 15 homologous domains, 2 of which have mannose 6-phosphate-binding activity (Figure 3) (Dahms *et al.*, 1989). Because the binding sites for the sugar ligand and the divalent metal ion cofactor are formed mostly from backbone amide and carbonyl groups, only five amino acid side chains in the sugar-binding subsite are common to the domains that interact with mannose 6-phosphate (Roberts *et al.*, 1998). Overall, the sugar-binding P-type CRDs are no more closely related to each other than they are to the non-binding domains.

Drosophila, *C. elegans*, and yeast proteins that contain P-type CRD-like sequences are diagrammed in Figure 3. However, these proteins lack almost all of the residues found in the sugar-binding sites of mammalian P-type CRDs and are thus unlikely to be mannose 6-phosphate receptors. This conclusion is consistent with the absence of such a receptor in cells from invertebrates such as *Dictyostelium discoideum*, even though modified forms of mannose 6-phosphate are attached to glycoproteins in this species (Mehta *et al.*, 1996). Thus, compared to

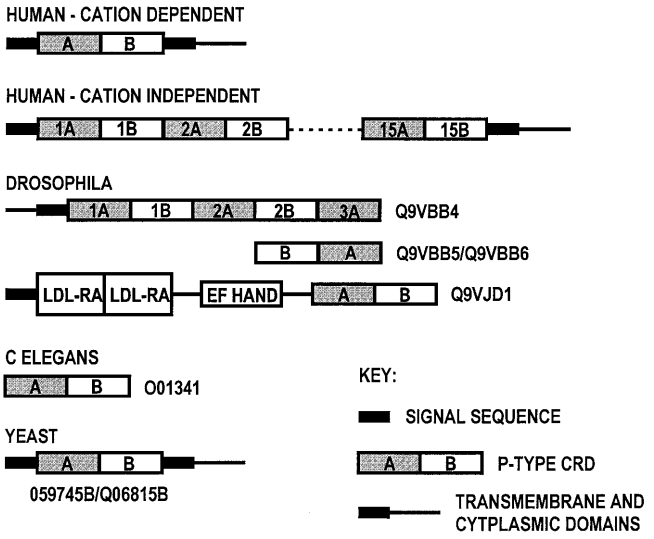

Fig. 3. Domain organization of mannose 6-phosphate receptors. The N-terminal and C-terminal β sheets of the P-type CRD are denoted A and B. In addition to screening with the motifs indicated in Table I, further searches were conducted using various relatively conserved sequences within the P-type CRDs. One of the *Drosophila* proteins appears to have undergone a cyclic permutation. LDL-RA, low density lipoprotein receptor class A domain.

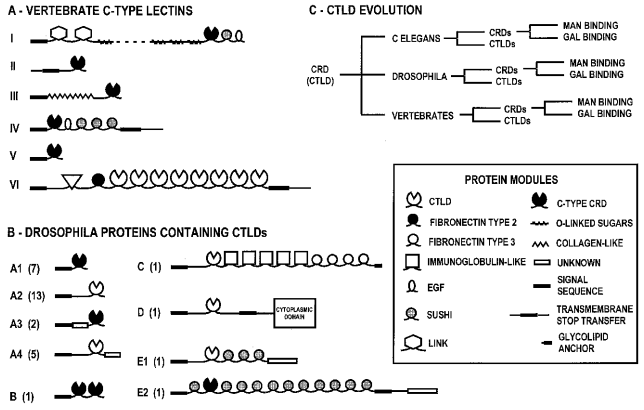

Fig. 4. Evolution of C-type lectins. (**A**) Domain organization of vertebrate proteins that contain C-type CRDs. Proteins are grouped by domain organization (Drickamer, 1993). (**B**) Domain organization of *Drosophila* proteins that contain CTLDs. Following identification of protein domains using the Pfam library of protein motifs, structural categories were defined by the arrangement of protein domains around the CTLDs. Thus, groups A1 to A4 contain signal sequences and single CTLDs with no spacer regions (group A1), short spacers consisting of repetitive sequence (group A2), or with flanking regions that might form small globular domains that do not correspond to known protein folds (groups A3 and A4). Group B contains a pair of CTLDs and the remaining groups of characteristic patterns of other types of protein modules. The numbers of different *Drosophila* gene products that fall into each structural category are indicated in parentheses. For structural groups in which one or more members contain residues associated with sugar-binding activity, the CTLDs are shaded. A subset of these domains were previously detected when a portion of the genome was screened (Theopold *et al.*, 1999). Profiles employed to screen for CTLDs are listed in Table I. (C) Summary of stages in the evolution of CTLDs.

the calnexin-related proteins and the L-type lectins, the mannose 6-phosphate receptors appear to be a more recent and possibly more specialized carbohydrate recognition system of animal cells.

C-type lectin-like proteins

C-type lectins are the most diverse family of animal lectins. These lectins are generally multidomain proteins, in which C-type CRDs provide Ca^{2+}-dependent sugar-recognition activity and a variety of other modules then initiate a broad range of biological processes, such as adhesion, endocytosis, and pathogen neutralization (Drickamer and Taylor, 1993; Weis *et al.*, 1998). The domain organizations of some of the vertebrate C-type lectins are summarized in Figure 4A. The sugar-binding sites in vertebrate C-type CRDs are formed in part by a bound Ca^{2+}, which must be present for sugar binding to occur. The C-type CRDs form a subgroup of a larger family of protein domains that share a common protein fold and are designated C-type lectin-like domains (CTLDs) (Weis *et al.*, 1998). Currently about a hundred human proteins that contain CTLDs have been described, and roughly half of these have been proposed to function as C-type CRDs. Many CTLDs bind to protein ligands rather than to sugars, and only some of these binding interactions are Ca^{2+}-dependent.

The domain organization of the 32 *Drosophila* proteins that contain CTLDs (Figure 4B) is strikingly different to the organization of both the known mammalian C-type lectins and the *C. elegans* proteins that contain CTLDs (Drickamer and Dodd, 1999). The only common domain architecture consists simply of an isolated CTLD. In multidomain proteins, *Drosophila* CTLDs are associated mostly with immunoglobulin, sushi, and fibronectin type 3 modules, whereas CTLDs in *C. elegans* are combined with CUB, epidermal growth factor, and von Willebrand factor modules. The lack of similarity in overall architecture of these proteins suggests that most of them serve distinct functions in the vertebrates and in each of the invertebrates.

In general, the *Drosophila* CTLD sequences are more closely related to each other than they are to CTLDs from *C. elegans* or vertebrates, a finding that is consistent with the idea that there has been a largely independent radiation of the CTLDs in different animal lineages. The only exceptions to this rule are the two *Drosophila* proteins in group E, one of which is the product of the *furrowed* gene (Leshko-Lindsay and Corces, 1997), which are distantly related to the product of the C54G4.4 locus in *C. elegans*. In addition to the weak similarity in the CTLD sequences, these three proteins share a similar domain organization, suggesting that they may have related functions.

A comprehensive analysis of the *C. elegans* genome identified 19 of the 183 CTLDs that contain most of the five residues needed to form the primary Ca^{2+}-binding site in vertebrate C-type CRDs (Drickamer and Dodd, 1999). Of these, seven have sequences consistent with formation of mannose- or N-acetylglucosamine-binding sites similar to those in vertebrate C-type CRDs. A similar analysis of the 32 *Drosophila* CTLDs reveals that only 6 show conservation of potential Ca^{2+}-liganding residues at positions that correspond to the five residues that form the primary Ca^{2+}- and sugar-binding site in

mammalian C-type CRDs (Weis *et al.*, 1992). In two cases, the pattern of Ca^{2+}-liganding residues is identical to that seen in mannose-binding C-type CRDs, and in two cases it corresponds to the arrangement seen in galactose-binding C-type CRDs. A protein product of one of the genes predicted to encode a galactose-binding CTLD has in fact been characterized as a galactose-binding lectin (Haq *et al.*, 1996), providing evidence for the utility of the comparative approach to identification of potential binding activity.

The presence of the Ca^{2+}- and sugar-binding site residues in conserved positions in a subset of CTLDs in mammals, *C. elegans*, and *Drosophila* might suggest that at least one CTLD present in their common progenitor contained these residues. However, the potential galactose- or mannose-binding C-type CRDs in these different species are less similar in overall sequence than are the various CTLDs within each species. Thus, it appears that the sugar-binding activity originated independently in each lineage (Figure 4C). Although this sequence of events might seem unlikely, it should be noted that the number of residues required to generate a sugar-binding site in the CTLD framework is quite small.

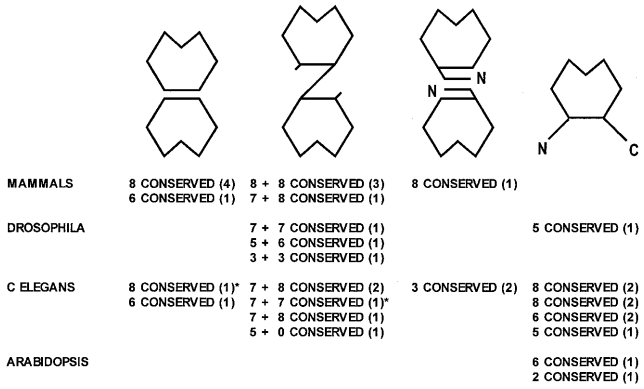

Fig. 5. Summary of galectin domain organization and conserved residues. One class of mammalian galectins contain N-terminal, proline-rich tails that contribute to dimerization. Although some of the invertebrate and plant galectin-like proteins contain longer N- or C-terminal extensions, these are not related in sequence. The number of residues associated with galactoside binding that are conserved in each class of galectin are shown below the diagrams. Values in parentheses indicate the number of proteins of each type identified in each species. Asterisks denote proteins that have been shown to have β-galactose binding activity (Adams, 2000). Galectin-like proteins were identified using profiles listed in Table I.

Galectins

Although mammalian galectins lack conventional signal sequences, they reach the cell surface by a novel mechanism and bind to glycoconjugates in the plasma membrane and in the extracellular matrix (Barondes *et al.*, 1994). The galectins consist of globular galectin-type CRDs with relatively minor accessory domains (Cooper and Barondes, 1999). A galectin-type CRD comprises a β sandwich similar in overall topology to the L-type CRDs. However, the lack of sequence similarity and the different way in which sugar-binding sites are constructed in these two families of domains suggest that this topological similarity results from convergent evolution. Most galectins contain multiple sugar-binding sites, due to the presence of two galectin-type CRDs in a single polypeptide or as a result of dimerization (Figure 5). A common function of the galectins may be to crosslink N-acetyllactosamine-containing structures found at cell surfaces and in the extracellular matrix. Studies of knockout mice suggest that multiple galectins provide distinct but overlapping functions (Colnot *et al.*, 1998).

Galectin-like proteins are ubiquitous in multicellular organisms, including *C. elegans* and *Drosophila*, but have not been identified in yeast (Kasai and Kirabayashi, 1996; Cooper and Barondes, 1999). Comparison of all the galectin sequences reveals conservation primarily of inwardly facing hydrophobic residues in β strands in the β-sandwich of the galectin fold (Lobsanov *et al.*, 1993; Liao *et al.*, 1994; Leonidas *et al.*, 1998). The eight residues that form the galactoside-binding site in most mammalian galectins are conserved in some invertebrate homologues, but in many cases some or all of these residues are not present. A vertebrate galectin containing only six of the canonical galactose-binding residues interacts with mannose rather than galactose (Swaminathan *et al.*, 1999), so it is possible that some of the invertebrate galectin homologues bind ligands other than galactosides.

The overall sequences of the mammalian galectin-type CRDs are marginally more similar to each other than they are to sequences of any of the invertebrate galectins. In other words, no invertebrate galectin is particularly similar to any one mammalian protein. Some of the invertebrate proteins also contain N- or C-terminal extensions that are different from those found in any known mammalian galectins. These results suggest independent radiation of galectins in the vertebrate and invertebrate lineages. Thus, some of the invertebrate galectins may perform functions that are distinct from the functions of mammalian galectins.

I-type lectins

The siglec family of cell surface adhesion receptors are sialic acid–binding proteins that contain I-type CRDs derived from the immunoglobulin fold (Crocker *et al.*, 1998). The evolution of sugar-binding activity in these I-type CRDs is best considered in the context of the evolution of immunoglobulin superfamily modules (Teichmann and Chothia, 2000). It has been noted that evolution of the siglecs parallels the appearance of sialic acid–containing ligands at cell surfaces (Angata and Varki, 2000).

Ricin-like domains

The ricin-like or R-type CRDs are the only sugar-binding protein modules from animal lectins that have also been found in bacteria. The galactose-binding B chain of the plant toxin ricin is formed of two homologous domains, each consisting of three lobes arranged as a β-trefoil around a threefold axis (Rutenber and Robertus, 1991; Murzin *et al.*, 1992). Many bacterial hydrolases resemble the ricin precursor, in which N-terminal hydrolytic domains are attached to C-terminal R-type CRDs (Figure 6) (Fujimoto *et al.*, 2000). In mammals, UDP-N-acetylgalactosamine: polypeptide N-acetylgalactosaminyltransferases that initiate synthesis of O-linked oligosaccharides in the *cis* Golgi

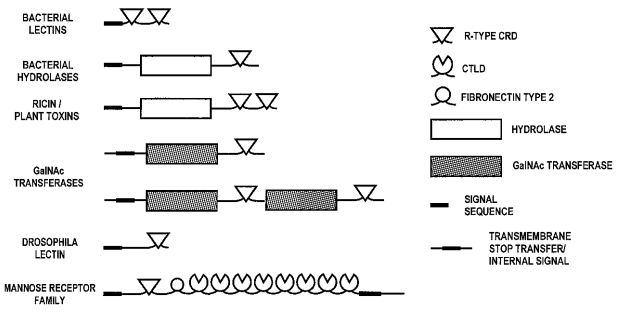

Fig. 6. Organization of domains in proteins containing R-type CRDs. The designation of domains is indicated by the key at the right. Potential R-type CRDs were identified using the profiles listed in Table I.

have a related organization (Clausen and Bennett, 1996). The C-terminal R-type CRDs in these proteins probably direct them to substrates (Hassan *et al.*, 2000; Fujimoto *et al.*, 2000). In contrast, proteins in the macrophage mannose receptor family contain R-type CRDs that are combined with a different set of protein modules (Taylor, 1997). In the mannose receptor, this domain interacts with sulfated N-acetylgalactosamine residues on glycoprotein hormones, leading to their clearance from the circulation (Fiete *et al.*, 1998).

The three lobes of the R-type CRD show evidence of early duplication (Rutenber *et al.*, 1987) and in some cases two or three of the lobes interact with glycans through a conserved set of residues located on the outer edge of these lobes. A different set of residues in the third lobe of the R-type CRD in the mannose receptor interacts with 4-sulfo-N-acetylgalactosamine (Liu *et al.*, 2000). Examination of the genomic sequences of *C. elegans* and *Drosophila* reveals families of N-acetylgalactosaminyltransferases that have activities like their mammalian counterparts and contain R-type CRDs (Hagen and Nehrke, 1998). Thus, this particular domain organization was probably established and duplicated early in the animal lineage. Although the common presence of R-type CRDs in bacteria and animal cells suggests an early origin for these domains, the possibility of lateral gene transfer into bacteria cannot be ruled out.

Discussion

Three general patterns emerge from the comparative analysis of lectins summarized in Figures 7 and 8. First, sugar-binding activities evolved from core recognition toward recognition of terminal elaborations (Drickamer and Taylor, 1998). Second, biological functions associated with sugar binding evolved from intracellular to extracellular. Third, both within and between species, the diversity of the lectins, the sugars that

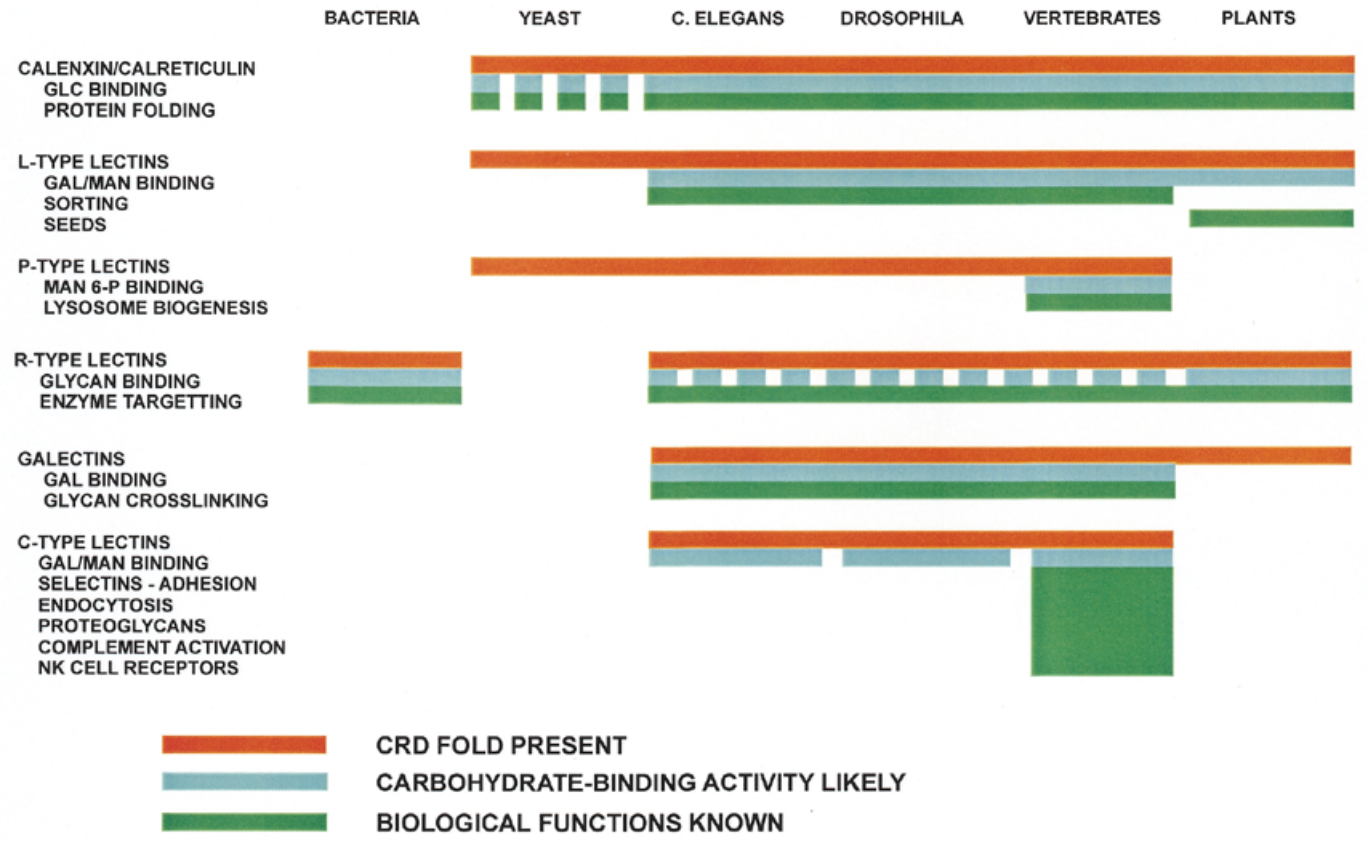

Fig. 7. Different evolutionary patterns observed for different structural classes of lectins. Orange bars indicate the presence of domains with structures defined by the different types of CRDs in animal lectins. Blue bars denote the existence of members of the different structural categories with demonstrated or predicted sugar-binding activity. Green bars indicate shared biological activities in different species.

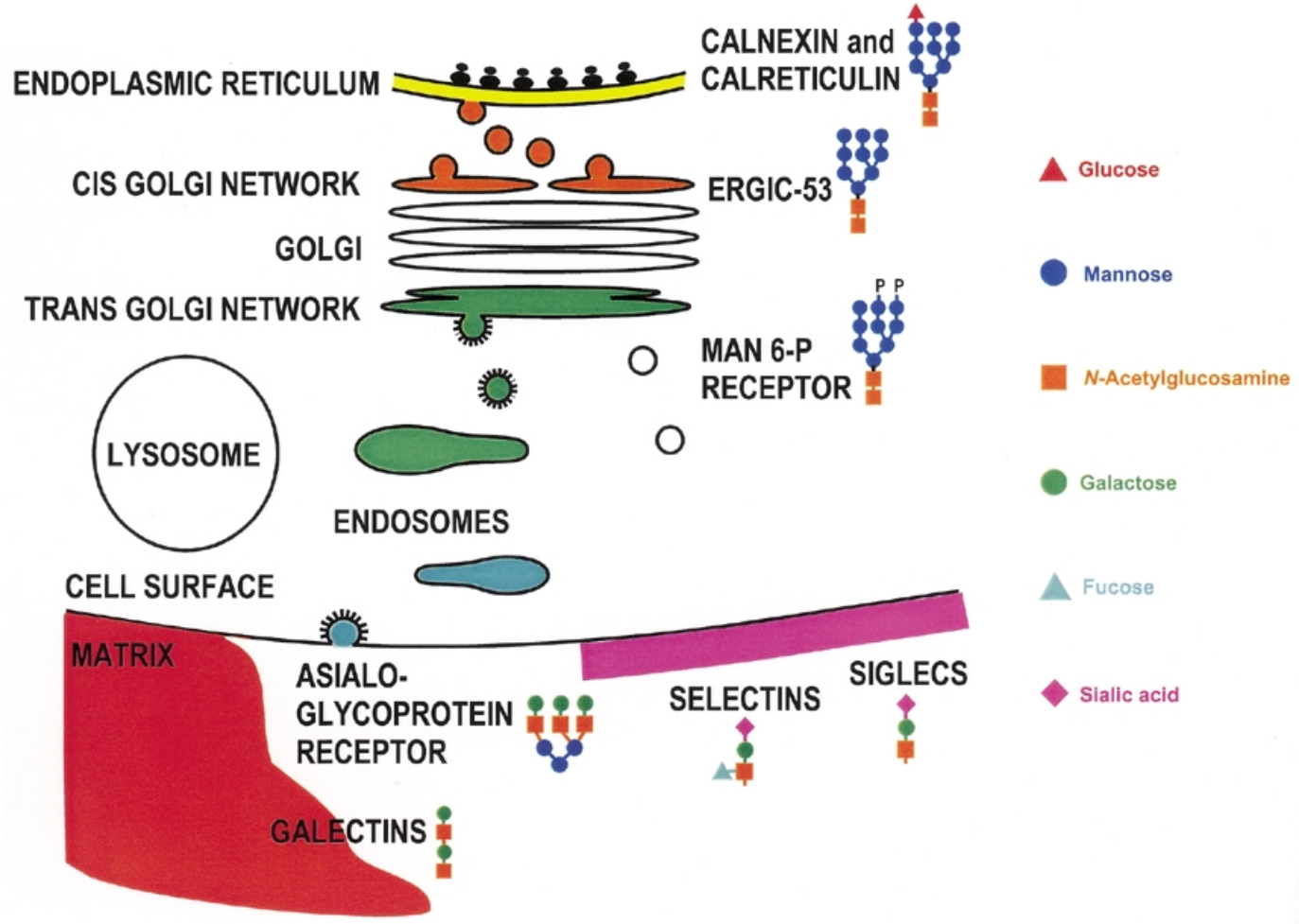

Fig. 8. Comparison of intracellular localization and sugar-binding activities of different lectin classes.

they recognize, and the biological functions associated with this recognition is greatest for the extracellular lectins that have evolved most recently.

Intracellular sorting functions associated with the luminal lectins must have first appeared at an early stage in the evolution of eukaryotes, suggesting that these functions are important in the basic physiology of cells in all of these organisms. In contrast, the C-type lectins have evolved largely independently in vertebrate and invertebrate lineages. The evolution of this class of lectins is paralleled by the appearance of distinct sets of terminal elaborations on complex N- and O-linked oligosaccharide chains on glycoproteins and on glycolipids at the cell surface and outside the cell. Better understanding of glycoconjugates in *Drosophila* (Seppo and Tiemeyer, 2000) and *C. elegans* (DeBose-Boyd *et al.*, 1998; Hagen and Nehrke, 1998; Chen *et al.*, 1999) will provide insight into the functions of C-type lectins in these organisms. Galectins have also undergone independent radiation in the different animal lineages, but at least some members of the family probably perform analogous sugar-binding functions in each of the lineages. The lactosamine units to which they bind probably appeared more recently than the core structures recognized by the luminal lectins and before the terminal elaborations that are bound by many of the C-type lectins (Drickamer and Taylor, 1998).

Relatively simple invertebrate organisms may serve as useful models for some (but not all) of the functions of sugar-binding proteins in mammals. The early intracellular sorting events involving calnexin and L-type lectins as well as the role of R-type CRDs in glycosyltransferases are likely to be quite similar, whereas later sorting events involving the mannose 6-phosphate receptors will probably be different. At the cell surface, the role of some of the galectins may be similar in all animals, so that genetic and developmental analysis of the model invertebrates is likely to illuminate studies of the vertebrate proteins as well. In contrast, the greater diversity of invertebrates and vertebrates proteins containing CTLDs suggests that these proteins probably participate in more specialized functions of glycans that are unique to different groups of animals.

Acknowledgments

This work was funded by grant 041845 from the Wellcome Trust. We thank Maureen Taylor for critical reading of the manuscript. Use of SRS software and FASTA searching provided on the European Bioinformatics Institute Web site

(<http://ebi.ac.uk>), the profile scanning software of the Swiss Institute for Experimental Cancer Research (<http://www.isrec.isb-sib.ch/software/PFSCAN_form.html>) and the sequence analysis software on the ExPASy Molecular Biology Server provided by the Swiss Institute of Bioinformatics (<http://www.expasy.ch/cgi-bin/protscale.pl>) is acknowledged.

Abbreviations

CRD, carbohydrate-recognition domain; CTLD, C-type lectin-like domain.

References

Adams, M.D., Celniker, S.E., Holt, R.A., Evans, C.A., Gocayne, J.D., Amanatides, P.G., Scherer, S.E., Li, P.W., Hoskins, R.A., Galle, R.F., and others (2000) The genome sequence of *Drosophila melanogaster*. *Science*, **287**, 2185–2195.

Angata, T., and Varki, A. (2000) Cloning, characterization, and phylogenetic analysis of siglec-9, a new member of the CD33-related group of siglecs: evidence for co-evolution with sialic acid synthesis pathways. *J. Biol. Chem.*, **275**, 22127–22135.

Barondes, S.H., Cooper, D.N.W., Gitt, M.A., and Leffler, H. (1994) Galectins: structure and function of a large family of animal lectins. *J. Biol. Chem.*, **269**, 20807–20810.

Cardin, A.D., and Weintraub, H.J. (1989) Molecular modeling of protein-glycosaminoglycan interactions. *Arteriosclerosis*, **9**, 21–32.

Chen, S., Zhou, S., Sakar, M., Spence, A.M., and Schachter, H. (1999) Expression of three *Caenorhabditis elegans* N-acetylglucosaminyltransferase I genes during development. *J. Biol. Chem.*, **274**, 288–297.

Clausen, H., and Bennett, E.P. (1996) A family of UDP-GalNAc: polypeptide N-acetylgalactosaminyl-transferases control the initiation of mucin type O-linked glycosylation. *Glycobiology*, **6**, 635–646.

Colnot, C., Fowlis, D., Ripoche, M.A., Bouchaert, I., and Poirier, F. (1998) Embryonic implantation in galectin 1/galectin 3 double mutant mice. *Dev. Dyn.*, **211**, 306–313.

Consortium, T.C.e.S. (1998) Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science*, **282**, 2012–2018.

Cooper, D.N.W., and Barondes, S.H. (1999) God must love galectins; He made so many of them. *Glycobiology*, **9**, 979–984.

Crocker, P., Clark, E.A., Filbin, M., Gordon, S., Jones, Y., Kehrl, J.H., Kelm, S., Le Douarin, N., Powell, L., Roder, J., and others (1998) Siglecs: a family of sialic-acid binding lectins. *Glycobiology*, **8**, v.

Dahms, N.M., Lobel, P., and Kornfeld, S. (1989) Mannose 6-phosphate receptors and lysosomal enzyme targeting. *J. Biol. Chem.*, **264**, 12115–12118.

DeBose-Boyd, R.A., Nyame, A.K., and Cummings, R.D. (1998) Molecular cloning and characterization of an α1, 3 fucosyltransferase, CEFT-1, from *Caenorhabditis elegans*. *Glycobiology*, **9**, 905–917.

Drickamer, K. (1993) Evolution of Ca^{2+}-dependent animal lectins. *Prog. Nucleic Acid Res. Mol. Biol.*, **45**, 207–232.

Drickamer, K., and Dodd, R.B. (1999) C-Type lectin-like domains in *Caenorhabditis elegans*: predictions from the complete genome sequence. *Glycobiology*, **9**, 1357–1369.

Drickamer, K., and Taylor, M.E. (1993) Biology of animal lectins. *Annu. Rev. Cell. Biol.*, **9**, 237–264.

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

Fernandez, F., D'Alessio, C., Fanchiotti, S., and Parodi, A.J. (1998) A misfolded protein conformation is not a sufficient condition for *in vivo* glucosylation by the UDP-Glc:glycoprotein glucosyltransferase. *EMBO J.*, **17**, 5877–5886.

Fiedler, K., and Simons, K. (1994) A putative novel class of animal lectins in the secretory pathway homologous to leguminous lectins. *Cell*, **77**, 625–626.

Fiete, D.J., Beranek, M.C., and Baenziger, J.U. (1998) A cysteine-rich domain of the "mannose" receptor mediates GalNAc-4-SO_4 binding. *Proc. Natl. Acad. Sci. USA*, **95**, 2089–2093.

Fujimoto, Z., Kuno, A., Kaneko, S., Yoshida, S., Kobayashi, H., Kusakabe, I., and Mizuno, H. (2000) Crystal structure of *Streptomyces olivaceoviridis* E-86 β-xylanase containing xylan-binding domain. *J. Mol. Biol.*, **300**, 575–585.

Hagen, F.K., and Nehrke, K. (1998) cDNA cloning and expression of a family of UDP-N-acetylgalatcosamine:polypeptide N-acetylgalactosaminyltransferase sequence homologs from *Caenorhabditis elegans*. *J. Biol. Chem.*, **273**, 8268–8277.

Haq, S., Kubo, T., Kurata, S., Kobayashi, A., and Natori, S. (1996) Purification, characterization, and cDNA cloning of a galactose-specific C-type lectin from *Drosophila melanogaster*. *J. Biol. Chem.*, **271**, 20213–20218.

Hassan, H., Reis, C.A., Bennett, E.P., Mirgorodskaya, E., Roepstorff, P., Hollingsworth, M.A., Burchell, J., Taylor-Papadimitiou, J., and Clausen, H. (2000) The lectin domain of UDP-N-acetyl-D-galactosmine:polypeptide N-acetylgalactosaminyltransferase-T4 direct its glycopeptide specificities. *J. Biol. Chem.*, **275**, 38197–38205.

Itin, C., Roche, A.C., Monsigny, M., and Hauri, H.P. (1996) ERGIC-53 is a functional mannose-selective and calcium-dependent human homologue of leguminous lectins. *Mol. Biol. Cell*, **7**, 483–493.

Jakob, C.A., Burda, P., Roth, J., and Aebi, M. (1998) Degradation of misfolded endoplasmic reticulum glycoproteins in *Saccharomyces cerevisiae* is determined by a specific oligosaccharide structure. *J. Cell Biol.*, **142**, 1223–1233.

Kasai, K., and Hirabayashi, J. (1996) Galectins: a family of animal lectins the decipher glycocodes. *J. Biochem. (Tokyo)*, **119**, 1–8.

Leonidas, D.D., Vatzaki, E.H., Vorum, H., Celis, J.E., Madsen, P., and Acharya, K.R. (1998) Structural basis for the recognition of carbohydrates by human galectin-7. *Biochemistry*, **37**, 13930–13940.

Leshko-Lindsay, L.A., and Corces, V.G. (1997) The role of selectins in *Drosophila* eye and bristle development. *Development*, **124**, 169–180.

Liao, D.-I., Kapadia, G., Ahmed, H., Vasta, G.R., and Herzberg, O. (1994) Structure of S-lectin, a developmentally regulated vertebrate α-galactoside-binding protein. *Proc. Natl. Acad. Sci. USA*, **91**, 1428–1432.

Liu, Y., Chirino, A.J., Misulovin, Z., Leteux, C., Feizi, T., Nussenzweig, M.C., and Bjorkman, P.J. (2000) Crystal structure of the cysteine-rich domain of mannose receptor complexed with a sulfated carbohydrate ligand. *J. Exp. Med.*, **191**, 1105–1115.

Lobsanov, Y.D., Gitt, M.A., Leffler, H., Barondes, S.H., and Rini, J.M. (1993) X-ray crystal structure of the human dimeric S-Lac lectin, L-14-II, in complex with lactose at 2.9-Å resolution. *J. Biol. Chem.*, **268**, 27034–27038.

Mehta, D.P., Ichikawa, M., Salimath, P.V., Etchison, J.R., Haak, R., Manzi, A., and Freeze, H.H. (1996) A lysosomal cysteine proteinase from *Dictyostelium discoideum* contains N-acetylglucosamine-1-phosphate bound to serine but not mannose-6-phosphate on N-linked oligosaccharides. *J. Biol. Chem.*, **271**, 10897–10903.

Murzin, A.G., Lesk, A.M., and Chothia, C. (1992) Beta-trefoil fold patterns of structure and sequence in the Kunitz inhibitors, interleukins-1β and 1α and fibroblast growth factors. *J. Mol. Biol.*, **223**, 531–543.

Parker, C.G., Fessler, L.I., Nelson, R.E., and Fessler, J.H. (1995) *Drosophila* UDP-glucose:glycoprotein glucosyltransferase: sequence and characterization of an enzyme that distinguishes between denatured and native proteins. *EMBO J.*, **14**, 1294–1303.

Parlati, F., Dominguez, M., Bergeron, J.J.M., and Thomas, D.Y. (1995) *Saccharomyces cerevisiae* CNE1 encodes an endoplasmic reticulum (ER) membrane protein with sequence similarity to calnexin and calreticulin and functions as a constituent of the ER quality control apparatus. *J. Biol. Chem.*, **270**, 244–253.

Parodi, A.J. (2000) Role of N-oligosaccharides endoplasmic reticulum processing reactions in glycoprotein folding and degradation. *Biochem. J.*, **348**, 1–13.

Rini, J.M. (1995) Lectin structure. *Annu. Rev. Biophys. Biomol. Struct.*, **24**, 551–577.

Roberts, D.L., Weix, D.J., Dahms, N.M., and Kim, J.-J.P. (1998) Molecular basis of lysosomal enzyme recognition: three-dimensional structure of the cation-dependent mannose 6-phosphate receptor. *Cell*, **93**, 639–648.

Rutenber, E., and Robertus, J.D. (1991) Structure of ricin B-chain at 2.5 Å resolution. *Proteins*, **10**, 260–269.

Rutenber, E., Ready, M., and Robertus, J.D. (1987) Structure and evolution of ricin B chain. *Nature*, **326**, 624–626.

Seppo, A., and Tiemeyer, M. (2000) Function and structure of *Drosophila* glycans. *Glycobiology*, **10**, 751–760.

Sharma, V., and Surolia, A. (1997) Analyses of carbohydrate recognition by legume lectins: size of the combining site loops and their primary specificity. *J. Mol. Biol.*, **267**, 433–445.

Sharon, N., and Lis, H. (1990) Legume lectins: a large family of homologous proteins. *FASEB J.*, **4**, 3198–3208.

Swaminathan, G.J., Leonidas, D.D., Savage, M.P., Ackerman, S.J., and Acharya, K.R. (1999) Selective recognition of mannose by the human

eosinophil Charcot-Leydon crystal protein (Galectin 10): a crystallographic study at 1.8 Å resolution. *Biochemistry*, **38**, 13837–13843.

Taylor, M.E. (1997) Evolution of a family of receptors containing multiple motifs resembling carbohydrate-recognition domains. *Glycobiology*, **7**, R5-R8.

Teichmann, S.A., and Chothia, C. (2000) Immunoglobulin superfamily protein in *Caenorhabditis elegans*. *J. Mol. Biol.*, **296**, 1367–1383.

Theopold, U., Rissler, M., Fabbri, M., Schmidt, O., and Natori, S. (1999) Insect glycobiology: a lectin multigene family in *Drosophila melanogaster*. *Biochem. Biophys. Res. Commun.*, **261**, 923–927.

Trombetta, E.S., and Helenius, A. (1998) Lectins as chaperones in glycoprotein folding. *Curr. Opin. Struct. Biol.*, **8**, 587–592.

Weis, W.I., Drickamer, K., and Hendrickson, W.A. (1992) Structure of a C-type mannose-binding protein complexed with an oligosaccharide. *Nature*, **360**, 127–134.

Weis, W.I., Taylor, M.E., and Drickamer, K. (1998) The C-type lectin superfamily in the immune system. *Immunol. Rev.*, **163**, 19–34.