PROTEINS: Structure, Function, and Bioinformatics 55:44–65 (2004)

Identification of Common Structural Features of Binding Sites in Galactose-Specific Proteins

M.S. Sujatha and Petety V. Balaji*
School of Biosciences and Bioengineering, Indian Institute of Technology Bombay, Powai, Mumbai 400 076, India

ABSTRACT **Galactose-binding proteins characterize an important subgroup of sugar-binding proteins that are involved in a variety of biological processes. Structural studies have shown that the Gal-specific proteins encompass a diverse range of primary and tertiary structures. The binding sites for galactose also seem to vary in different protein-galactose complexes. No common binding site features that are shared by the Gal-specific proteins to achieve ligand specificity are so far known. With the assumption that common recognition principles will exist for common substrate recognition, the present study was undertaken to identify and characterize any unique galactose-binding site signature by analyzing the three-dimensional (3D) structures of 18 protein-galactose complexes. These proteins belong to 7 nonhomologous families; thus, there is no sequence or structural similarity across the families. Within each family, the binding site residues and their relative distances were well conserved, but there were no similarities across families. A novel, yet simple, approach was adopted to characterize the binding site residues by representing their relative spatial dispositions in polar coordinates. A combination of the deduced geometrical features with the structural characteristics, such as solvent accessibility and secondary structure type, furnished a potential galactose-binding site signature. The signature was evaluated by incorporation into the program *COTRAN* to search for potential galactose-binding sites in proteins that share the same *fold* as the known galactose-binding proteins. *COTRAN* is able to detect galactose-binding sites with a very high specificity and sensitivity. The deduced galactose-binding site signature is strongly validated and can be used to search for galactose-binding sites in proteins. PROSITE-type signature sequences have also been inferred for galectin and C-type animal lectin-like *fold* families of Gal-binding proteins. Proteins 2004;55:44–65.**

Key words: lectins; substrate recognition; protein-carbohydrate interactions; distance matrix; fuzzy recognition; recognition motif; binding site signature; aromatic stacking interaction; functional genomics

INTRODUCTION

Characterization of biochemical function that involves the study of ligand-binding property, mechanism of catalysis or antigenic site prediction is mainly undertaken on the basis of the knowledge of three-dimensional (3D) structure of the protein. However, similarity in overall fold does not necessarily imply similarity in biochemical function because local 3D structure, rather than the overall fold, is important for recognition and binding to a ligand. Thus, proteins sharing the same fold may perform different functions; conversely, proteins that share similar active site features perform the same function despite having different folds.[1,2] Examples of protein pairs that share functional similarity in the absence of any sequence or structural similarity include chymotrypsin/subtilisin[3,4] and α- and β-carbonic anhydrases.[5] Thus, proteins that bind a common substrate or share a common catalytic mechanism can be expected to at least have a similar spatial disposition of the functional groups that interact with the ligand.

Galactose-binding proteins form an important subgroup of sugar-binding proteins, and they mediate several key biological processes.[6–8] Structural studies have shown that the Gal-specific proteins differ not only in their primary but also in their tertiary structures (Table I). The binding sites of galactose also seem to vary in different protein-galactose complexes: all the hydroxyl groups of galactose are stabilized by hydrogen bonds in *Erythrina corallodendron* lectin.[9] In S-lac lectin, Gal:O2 does not form a hydrogen bond with any of the protein atoms.[10] In some proteins such as S-lectin, water molecules mediate galactose-protein interactions,[11] whereas in a few others, a divalent calcium ion mediates the sugar binding (tunicate C-type lectin[12]). Despite this apparent dissimilarity in the binding site architecture, all these proteins are specific to galactose. With the assumption that common recognition principles exist for common substrate recognition, the present work was initiated to identify the common features of the galactose-binding sites using 3D structures of a nonredundant set of 18 protein-galactose

Grant sponsor: Council of Scientific and Industrial Research, India; Grant number: 37(1110/02/EMR-II).

*Correspondence to: P.V. Balaji, School of Biosciences and Bioengineering, Indian Institute of Technology Bombay, Powai, Mumbai 400076, India. E-mail: balaji@iitb.ac.in

Received 26 March 2003; Accepted 30 July 2003

TABLE I. Proteins With Similar Sequence and Structure but Different Ligand Specificities[†]

Protein	GI number	Percent similarity	E-value	RMS deviation	Ligand
Reference protein: *Erythrina corallodendron* lectin (GI: 3212463; Gal-specific)					
Peanut lectin	3891464	57	$6e^{-38}$	1.6	Gal
Winged bean acidic lectin	15826315	70	$5e^{-4}$	1.5	Gal
Concanavalin A	576421	57	$3e^{-1}$	1.3	Man/Glc
Pea lectin	6729957	59	$3e^{-50}$	1.1	Glc/Man
Agglutinin II precursor	Q39529	58	$2e^{-44}$	—	Glc/Man
Reference protein: Calcium binding protein (GI: 1431805; Gal-specific)					
Mannose-binding-like lectin	NP_571645	51	$2e^{-16}$	—	Man
Monoglyceride lipase	NP_071788	51	$2e^{-13}$	—	Gal/GalNAc
Reference protein: Tunicate C-type lectin (GI: 5822426; Gal-specific)					
Galactose-specific lectin	P21963	45	0.12	—	Gal
Mannose-binding-like lectin	NP_571645	41	0.031	—	Man
Reference protein: Jacalin (GI: 2392381; Gal-specific)					
Artocarpin	20150328	66	$5e^{-34}$	0.9	Man

[†]The percent similarity and the e-value were obtained by performing a pair-wise BLAST analysis of each protein with its corresponding reference protein. All the protein sequences were retrieved from the GenBank using the given GI numbers. For protein pairs with known 3D structures, the root-mean square deviation obtained from the DALI server (http://www2.ebi.ac.uk/dali/) indicates the structural similarity.

complexes belonging to 7 nonhomologous protein families (Table II). It is shown that the common features do indeed exist for galactose recognition and that the common features so identified do characterize the galactose-binding site signature fairly uniquely.

RESULTS

Primary and Tertiary Structure Similarities Are Absent Across the Galactose-Binding Protein Families

Significant sequence similarity exists only among legume lectins (38–63%), galectins (26–50%), and between the two ricin B-like proteins (48%). There is no detectable sequence similarity between any of the other protein pairs constituting the nonredundant data set. These 18 Gal-specific proteins belong to seven nonhomologous protein families. Of these, only legume lectins and galectins share the same fold (Table III). According to the SCOP database, the C-type animal lectins belong to the *alpha + beta class,* whereas the rest belong to the *all beta class;* within the latter class, the proteins belong to five distinct fold types (Table III).

Even the Type of Amino Acid Residues or the Functional Groups That Constitute the Galactose-Binding Site Are Not Conserved

The galactose-binding sites were examined to identify common features that may be shared by these proteins. Examination of the amino acid residues that constitute the binding site (i.e., those that are within 4.0 Å from galactose) revealed the presence of a potential hydrogen bond donor/acceptor around the hydroxyl groups of galactose and an aromatic residue stacking against the *b* face of galactose (Fig. 1). *These are the only two common features* shared by these 18 proteins (Table IV). However, the number and nature of residues constituting the binding site are not same. In some cases, the residue is acidic, in

TABLE II. Data Set Used for Galactose-Binding Site Analysis

Protein	Symbol, chain identifier, and residue number of galactose	PDB ID: chain identifier, resolution[a]
Family: Legume lectins		
Erythrina corallodendron lectin	GAL: 402	1AX1, 1.95
Soybean agglutinin	GAL: S: 2	1G9F:A, 2.5
Winged bean lectin	AMG: A: 400	1WBL:A, 2.5
Peanut lectin	GAL: E: 400	1BZW:A, 2.7
Winged bean acidic lectin	AMG: 400	1F9K:A, 3.0
Family: Galectins		
Human galectin3	GAL: 500	1A3K, 2.1
Congerin I	GAL: B: 137	1CIL:A, 1.5
Toad ovary galectin	GAL: A: 2	1GAN:A, 2.23
S-lectin	GAL: 402	1SLT:A, 1.9
S-lac lectin	GAL: A: 998	1HLC:A, 2.9
Human galectin 7	GAL: 998	2GAL:A, 1.95
Family: C-type animal lectins		
Gal-specific mutant MBP-A[b]	MGA: 1: 1	1AFA: 1, 2.5
Tunicate C-type lectin	GAL: A: 1	1TLG:A, 2.2
Family: Ricin B-like		
Ricin, B chain (lectin)	GAL: 264	2AAI:B, 2.51
Ebulin, B chain (lectin)	GAL: 280	1HWM:B, 2.8
Family: Mannose-binding lectins		
Moraceae plant lectin	AMG: A: 200	1JAC:A, 2.43
Family: Bacterial AB5 toxins, B-subunit		
Heat labile enterotoxin	GAL: 1104	1DJR:D, 1.3
Family: Galactose-binding domain		
Neuraminidase, C-terminal sugar-binding domain	GAL: 2	1EUU, 2.5

[a]From the Protein Data Bank.[48] There are no chain identifiers in 1AX1. 1A3K, and 1EUU.

[b]MBP mannose-binding protein-A, lectin domain, has been genetically engineered to confer galactose specificity.

TABLE III. SCOP Classification of Proteins in the Data Set[†]

Class	Fold	Superfamily	Family	PDB code of protein in the data set
All beta proteins	Concanavalin A-like lectins/glucanases	Concanavalin A-like lectins/glucanases	Legume lectins	1AX1, 1G9F, 1WBL, 1BZW, 1F9K
			Galectin (animal S-lectin)	1A3K, 1C1L, 1GAN, 1SLT, 1HLC, 2GAL
	OB-fold	Bacterial enterotoxins	Bacterial AB5 toxins, B-subunit	1DJR
	Galactose-binding domain-like	Galactose-binding domain-like	Galactose-binding domain	1EUU
	Beta-Trefoil	Ricin B-like lectins	Ricin B-like	1HWM, 2AAI
	Beta-Prism I	Mannose-binding lectins	Mannose-binding lectins	1JAC
Alpha and beta proteins (a+b)	C-type lectin-like	C-type lectin-like	C-type lectin domain	1AFA, 1TLG

[†]From the SCOP database.[49]

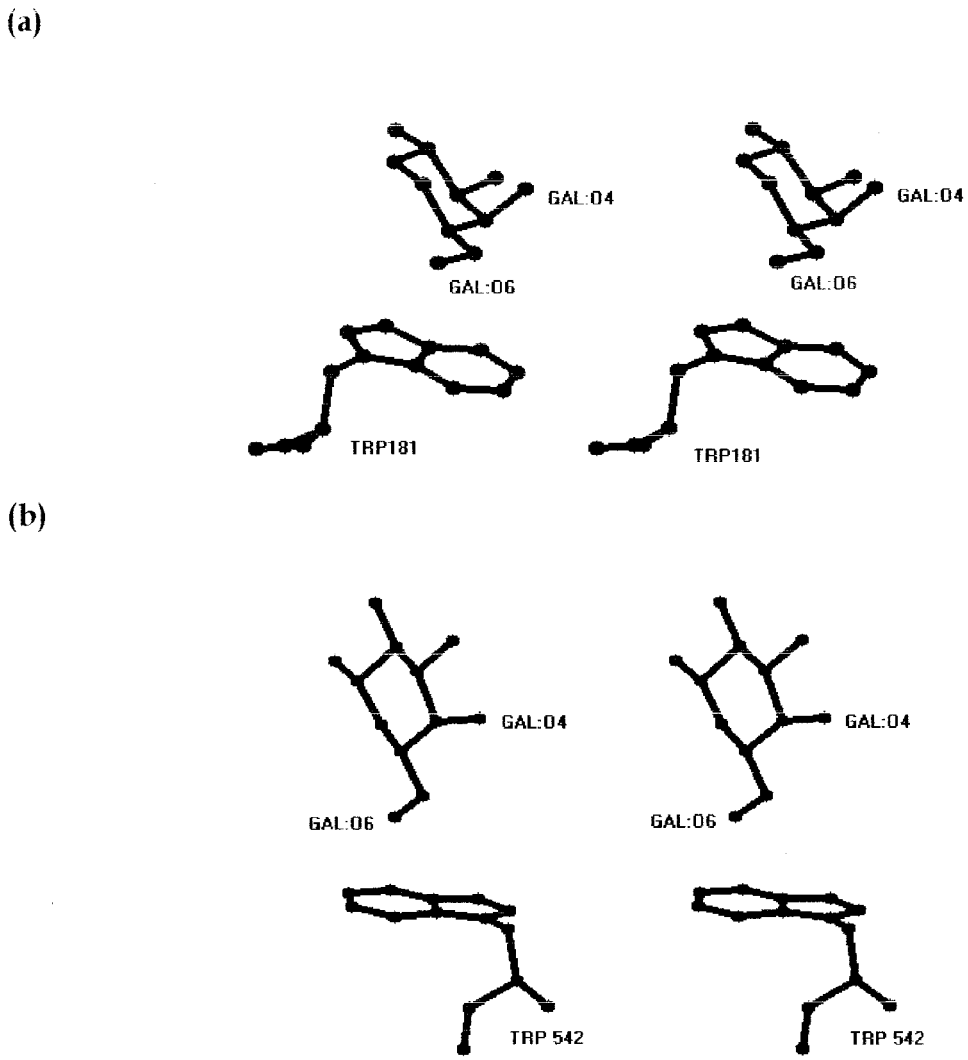


Fig. 1. The stacking interaction of the aromatic residue against the **b** face of galactose is a feature shared by all the Gal-specific proteins characterized to date. However, as can be seen from this stereo diagram, there can be differences in the mode of stacking: the Gal:H3 and Gal:H4 atoms are above Trp in human galectin 3 (1A3K; **a, top**); in contrast, Gal:H61 and Gal:H62 are above Trp in neuraminidase (1EUU; **b, bottom**). By convention, the axial hydroxyl group at C4 points toward the *a* face of galactose, whereas the H1, H3, and H5 atoms point toward the *b* face.

TABLE IV. Hydrogen-Bonding Atoms† and Stacking Residues in Galactose-Binding Sites

	Interacting atom of galactose					Stacking
PDB code	-OH at C-2	-OH at C-3	-OH at C-4	Ring O	-OH at C-6	residue
Family: Legume lectins						
1AX1	Nδ2 N133	Oδ1 D89 N G107	Oδ2 D89	N A218	Nε2 Q219	F131
1G9F	Nδ2 N130	Oδ1 D88 N G106	Oδ2 D88	N L214	Oδ1 D215	F128
1WBL	Nδ2 N128	Oδ2 D87 N G105	Oδ1 D87	N D212	Nε2 H84	F126
1BZW	Nδ2 N127	Oδ2 D83 N G104	Oδ1 D83	Oγ S211	Oδ2 D80	Y125
1F9K	Nδ2 N129	Oδ1 D88 N G106	Oδ2 D88	N Y215	Oε1 Q216	F127
Family: Galectins						
1A3K	HOH 1105	HOH 1021	Nε2 H158	Nη2 R162	Nδ2 N174 Oε2 E184	W181
1C1L	HOH 218	Nη2 R29 Nη1 R29	Nε2 H44	Nη2 R48	Nδ2 N61 Oε2 E73	W70
1GAN	Nε2 H53	HOH 14	Nε2 H45	Nη2 R49	Nδ2 N62 Oε2 E72	W69
1SLT	Nε2 H52	HOH 25	Nε2 H44	Nη2 R48	Nδ2 N61 Oε2 E71	W68
1HLC		Nη2 R120	Nε2 H45	Nη1 R49	Oε1 E68 Nδ2 N58	W65
2GAL	HOH 32	HOH 15	Nε2 H49	Nη2 R53	Nδ2 N62 Oε2 E72	W69
Family: C-type animal lectins						
1AFA	Nδ2 N210	Oε1 E198	Oδ1 D187	HOH 252	HOH 256	W189
1TLG		Oε1 E86	Oδ1 D107	HOH 34	HOH 42	W100
Family: Ricin B-like						
2AAI	Nζ K40	Oδ2 D22 Nδ2 N46	Oδ1 D22	N D25	Oδ1 D25	W37
1HWM		Oδ2 D24	Oδ2 D24	N N27	Nη2 R115	W39
Others						
1JAC		N G1	Oδ2 D125	N Y122	Oδ1 D125	Y78
1DJR	Nδ2 N90	Oδ1 N90 Nζ K91	Oε1 E51		Nε2 Q61	W88
1EUU		Nη2 R572	Nη1 R572			W542

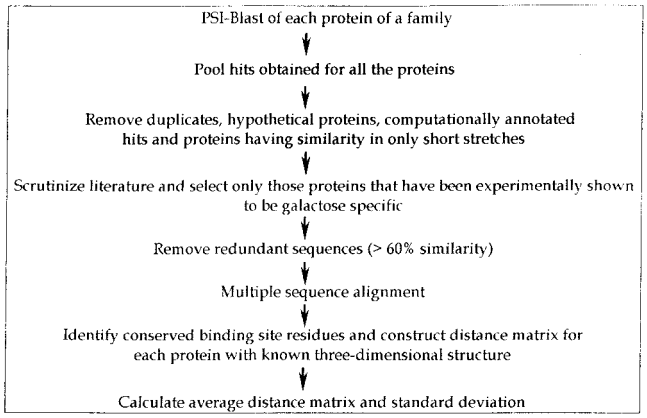
†The amino acid residue names are denoted by their single-letter code. Besides the potential hydrogen-bonding atoms and stacking residue listed here, other atoms/residues are also found within 4 Å from galactose. Residues that do not form optimal hydrogen-bonding interactions as deduced by visual inspection have been omitted.

some it is basic, and in a few others, it is neutral; around any given hydroxyl group of galactose, the interacting functional groups are also not the same; besides, the functional group(s) hydrogen bonding with a specific hydroxyl group of galactose can be a hydrogen bond donor or acceptor or both in some proteins.

The stacking of galactose against the aromatic amino acid is a well-documented feature of Gal-specific lectins.[13–17] For example, introduction of a tryptophan in the mannose-binding protein A resulted in an increased affinity for galactose.[18] A naturally occurring isolectin of the protein RIPt has Trp37→Leu and Asn46→Lys mutations and the binding site for galactose in this isolectin has been suspected to be inactive.[19] It has also been shown that substitution of the aromatic residue by a nonpolar residue results in decreased affinity for galactose and improved affinity for GalNAc.[20]

The Distances Between the Binding Site Residues Are the Same Within a Family But Are Different Across Families

The relative distances between the binding site residues (i.e., those that are within 4.0 Å from galactose) were computed to check for conservation among all 18 proteins of the nonredundant data set. To enable comparison of the relative distances of the binding site residues across the families, family-specific distance matrices were computed first and were then used for comparison. To ensure that the distance matrix is a true representative of all the proteins of the family, only those binding site residues that are conserved in all the proteins were considered for distance matrix computation. The conservation of the binding site residues was inferred on the basis of a multiple-sequence alignment. A nonredundant data set (i.e., only proteins that have no more than 60% sequence



Scheme 1. Steps involved in construction of family-specific distance matrix.

similarity among themselves) was considered for the multiple-sequence alignment to avoid bias, and such a data set was obtained following the procedure outlined in Scheme 1.

Galectin family

About 350–400 hits were obtained from PSI-Blast for each protein of the galectin family. Screening of the hits as specified in Scheme 1 resulted in 50 hits that were specific to galactose and one that was specific to mannose (charcot-Leyden-crystal or CLC protein). Removal of redundant sequences from among the 50 Gal-specific proteins resulted in a nonredundant data set of 17 proteins. Multiple-sequence alignment of these proteins showed that seven of the binding site residues are conserved in all the proteins [Fig. 2(a)]. Residues such as Arg29 of congerin I, which form hydrogen bond with galactose, are not conserved in

(a)

```
                                  * * *           * *         * *
1A3K        140 PNANRIALDFQR----GNDVAFHFNPRFNENNR-RVIVCNTKLDNN--WGREERQSVFPF 192
1C1L         75 NSPQRFSVNVGE---SMNSLSLHLDHRFNYGADQNTIVMNSTLKGDNGWETEQRSTNFTL 131
1GAN         76 PDCKGFAVNLGE---DASNFLLHFNARFDLHGDVNKIVCNSKEADA--WGSEQREGVFPF 130
1SLT         79 ADAKSFLLNLGK---DDNNLCLHFNPRFNAHGDVNTIVCNSKDAGA--WGAEQRESAFPF 130
1HLC         76 DGTDGFVINLGQ---GTDKLNLHFNPRFSES----TIVCNSLDGSN--WGQEQREDHLCF 126
2GAL         78 PNASRFHVNLLCGEEQGSDAALHFNPRLDTSE----VVFNSKEQGS--WGREERGPGVPF 131
1QMJ         75 SDAKGFSVNVGK---DSSTLMLHFNPRFDCHGDVNTVVCNSKEDGT--WGEEDRKADFPF 130
AAL37895     52 TDENRFQVDFQMGISDTNNVAFHFNPRFEGSG---YVVCNTMQLGN--WGPEEKKMQMPF 106
BAB83247     76 KDAKSFAINLGK---DSSNYVIHFNPRFDHHGDTNKIICNSKEENH--WGKEQRENAFPF 130
A42812       78 KGTNRFAVDLGT---DSRNLICHCNPRFEYSVDKNTIVLNSKQNDV--WDIEKKETAFPF 132
P07583       77 PNAKSFVMNLGK---DSTHLGLHFNPRFDAHGDVNLIVCNSKKMEE--WGTEQRETVFPF 131
1304250A     77 PNAKSFVMNLGK---DSTHLGLHFNPRFDAHGDVNLIVCNSKKMEE--WGTEQRETVFPF 131
Q9YIC2       61 SNATRFSINVGE---STDSIAMHMDHRFSYGADQNVLVLNS-LVHNVGWQQEERSKKFPF 117
A28302       61 IDSTNFAINVGN---SAEDLALHINPRFDAHGDQQAVVVNSFQGGN--WGTEQREGGFPF 117
AAF61069     60 PDATNFALNIGP---TDQDIVMHINPRFNAHGDENAVVCNSYIGRQ--WCEELREGGFPF 116
CAB71314     40 TTGDRFEINLRT----HSETAMHFNPRFDEDT----VVFNSTSGGD--WQNEER-VQMPF 88
CAA93822     35 HDQFNINLQTGPNTNPRDDTALHISIRPRDG----VIIRNSIQFRN--WGIEERFGGCPV 88
```

(b)

Residue	H158	N160	R162	V172	N174	W181	E184
H158		5.5	9.5	6.6	5.2	7.1	9.2
N160	0.1		4.7	4.5	8.8	10.7	9.7
R162	0.1	0.1		5.5	11.4	12.6	9.9
V172	0.2	0.1	0.2		6.8	10.1	5.9
N174	0.1	0.1	0.1	0.2		5.6	5.6
W181	0.1	0.2	0.2	0.2	0.2		8.2
E184	0.1	0.1	0.2	0.1	0.1	0.1	

Fig. 2. **a:** Multiple-sequence alignment of 17 galactose-specific proteins that were identified by PSI-Blast analysis (and subsequent screening as described in Scheme 1) as homologues to proteins belonging to the galectin family (see Table II). These proteins do not share >60% sequence similarity among them and hence, constitute a nonredundant data set. Conserved binding site residues are in bold and marked by an asterisk (*) above the alignment. The PDB code, if available, or the GenBank accession number is given: 1A3K, 1C1L, 1GAN, 1SLT, 1HLC, and 2GAL, as in Table II. 1QMJ, chicken galectin; AAL37895, *Ovis aries* galectin-14; BAB83247, *Xenopus laevis* galectin-Ia; A42812, 16K lactose-binding lectin from African clawed frog; P07583, beta-galactoside-binding lectin (14-kD lectin; C-14), 1304250A, beta galactoside-binding lectin from chicken embryo; Q9YIC2, congerin II; A28302, beta-galactoside-binding lectin from electric eel; AAF61069, *Paralichthys olivaceus* galectin; CAB71314, *Haemonchus contortus* galectin; CAA93822, *Anopheles gambiae* lectin. **b:** Distance matrix of binding site residues, averaged for the first seven proteins in the multiple-sequence alignment for which 3D structure data are available. Residue numbering corresponds to human galectin 3, 1A3K. The upper triangle of the matrix shows the average distance, and the lower triangle shows the corresponding standard deviation.

all the proteins of this family (1C1L; Table IV). The conserved residues either are directly involved in galactose binding (e.g., His158) or are probably involved in maintaining the binding site architecture (e.g., Asn160).

The multiple-sequence alignment was used to infer a PROSITE-type[21] signature sequence -H-x-[NDS]-x-R-x(6,10)-[VI]-x-N-x(6,8)-W-x(2)-E- for this family of galactose-binding proteins. Scanning the PROSITE database, which includes the sequences deposited in Swiss-Prot, TrEMBL, TrEMBL-new, and PDB for this signature sequences resulted in 155 hits. All the hits are galectins except for meiotic endonuclease 12-kDa subunit (Swiss-Prot ID: Q00358), urate transporter/channel protein (Q9XSM8, Q9XSM9), porcine adenovirus 4 putative fiber protein (Q83467), VHSV-induced protein-9 (Q8QGB1), Lgals8 protein (BAB23560), myosinase- III (BAC16240). Thus, this signature sequence can be considered to be representing the galectin family of proteins.

The 3D structure data are available for 7 proteins belonging to this 17 protein nonredundant data set. Of these seven proteins, the structure of chicken galectin (1QMJ) has been determined in the absence of bound galactose, whereas the other six structures are with bound galactose. The relative distance between the seven conserved residues as deduced from the multiple-sequence alignment was calculated in each protein, and these were used to calculate the average relative distances [Fig. 2(b)]. The relative distances of the binding site residues are similar in structures determined with (e.g., human galectin 3; 1A3K) and without (chicken galectin; 1QMJ) bound galactose, suggesting that galactose binding is not accompanied by conformational changes of the binding site residues. The standard deviations in the relative distances of the binding site residues in the seven proteins are small [<0.2; Fig. 2(b)], suggesting that the relative spatial disposition of the binding site residues is well conserved in these proteins and hence, can be considered to be characteristic for this family.

A similar multiple-sequence alignment of about 20 galectins, which included even proteins that have >60% sequence similarity, was also performed by Dodd and Drickamer.[22] From such an alignment, eight residues were identified as constituting the binding site, of which seven are the same as those identified in the present study. The remaining residue (Arg186 of human galectin 3) is >4 Å away from the ligand. Three of the conserved residues are mutated in the homologous mannose-specific CLC-protein (Asn160→Gln, Arg162→Cys, and Glu184→Gln; human galectin 3 numbering). Such a partial conservation has been shown to lead to significant changes in the topology and chemical nature of the carbohydrate recognition domain from X-ray crystallographic studies, and such changes have been implicated in the altered ligand specificity.[23]

Ricin B-like family

About 500 hits were obtained from PSI-Blast analysis for each of the two proteins, ebulin and ricin B chains (Table II). The hits included proteins that were specific to

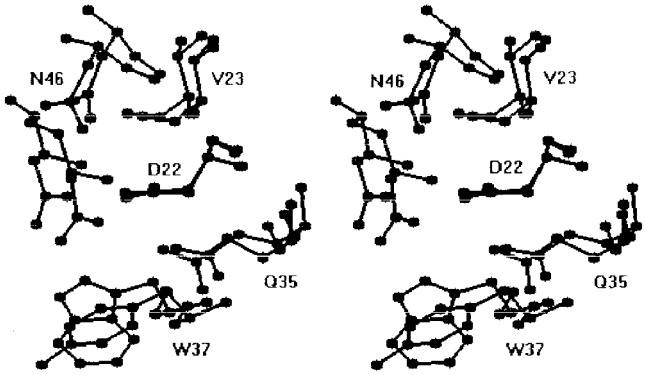


Fig. 3. Stereo diagram showing the binding site residues of ricin B-chain (2AAI) superposed over those of xylanase (1XYF). The superposition has been done with respect to Asp22 (in 2AAI) and Asp325 (in 1XYF). All the residues superpose well except for Asp25 (of 2AAI) and Asn328 (of 1XYF). Xylanases have been shown to bind to both galactose and xylose/xylooligosaccharides in the same binding pocket albeit in different modes. The structure of ricin B chain has been determined with bound galactose, whereas that of xylanase was determined without bound galactose.

Gal, UDP-GalNAc, and xylan. Some of the Gal-specific proteins also bind GalNAc. The xylan-binding domain (XBD) of xylanases have been shown to specifically bind to galactose-containing sugars thought to have evolved from the ancient ricin superfamily to bind additional sugar targets.[24a–c] The sugar-binding residues seen in ricin/lactose complex have also been found to be spatially conserved in XBD by X-ray crystallographic studies (Fig. 3).[25] Hence, xylan-specific proteins were also included for the multiple-sequence alignment.

Removal of redundant entries from among the Gal-specific proteins resulted in a nonredundant data set of eight proteins (seven lectins and one xylanase). A multiple-sequence alignment of this data set showed that the binding site residues Asp22 and Asn46 (ricin B numbering) are conserved in all members of the nonredundant data set [Fig. 4(a)]. The residue Trp37 is conserved in all the proteins except xylanase (1XYF) where it is replaced by a Tyr. The residues Val23 and Asp25 also have conservative replacements [Fig. 4(a)]. Lys40, which forms a hydrogen bond with O2 hydroxyl group of galactose is not conserved in this family of proteins.

Gln35 (ricin B numbering, 2AAI) is conserved in all the proteins except abrin-A (GI: P11140) and *Lumbricus terrestris* 29-kDa galactose-binding lectin (GI: BAA36395) wherein it is replaced by isoleucine. 3D structure data of ricin B chain (2AAI) and ebulin (1HWM) show that Gln35 forms hydrogen bonds with both Gal:O4 and Gal:O6. It is not clear which residues, if any, compensate for the loss of interaction due to the Glu35→Ile mutation in abrin-A and *L. terrestris* Gal-binding lectin. It is of interest that Gln35 and Trp37 are part of the conserved *Gly-X-X-X-Gln-X-Trp* sequence motif; this motif also is present in abrin-A and *L. terrestris* Gal-binding lectin[25a] but in a different segment of the polypeptide chain than that shown Figure 4(a).

The binding site residues conserved in the Gal-specific proteins [Fig. 4(a)] are not conserved in the UDP-GalNAc-specific proteins that were obtained from the PSI-Blast

(a)

```
                          **  *              *           *
2AAI(B chain)          22DVRD-GRFHNGNAIQLWPCKSN-TDANQLW46
1HWM(B chain)          24DVRN-GYDTDGTPIQLWPC-G--TQRNQQW49
1CE7(B chain)          19DVRD-DDFHDGNQIQLWPSKSN-NDPNQLW46
1XYF                  325DVPN-ASTTDGTQVQLYDCHSAT---NQQW350
P33183(B chain)        43DVRN-GYDTDGTPLQLWPC-G--TQRNQRW66
AAF37219              353DVRN-ESNNDGIPIQLWPC-G--AQRNQQW378
P11140                288DVYD-NGYHNGNRIIMWKCDKR-LEENQLW315
BAA36395              146DI-EGQNPAPGSKIITWDQKKGPTAVNQLW174
```

(b)

Residue	D22	V23	D25	W37	N46
D22		4.2	8.5	5.9	5.1
V23	0.0		7.1	9.4	4.6
D25	0.4	0.3		9.4	8.4
W37	0.2	0.1	0.4		9.2
N46	0.2	0.3	0.1	0.2	

Fig. 4. **a:** Multiple-sequence alignment of eight Gal-specific proteins identified by PSI-Blast analysis (and subsequent screening as described in Scheme 1) as homologues to proteins belonging to the ricin B-like family (see Table II). These proteins do not share >60% sequence similarity among them and hence, constitute a nonredundant data set. Glu148 in BAA36395 has been manually aligned with Asp25 in 2AAI. Conserved binding site residues are shown in bold and marked by an asterisk (*) above the alignment. The PDB code, if available, or the GenBank accession number is given: 2AAI and 1HWM, as in Table II. 1CE7, Mistletoe lectin I from *Viscum Album*; 1XYF, endo-1,4-beta-xylanase from *Streptomyces olivaceoviridis*; P33183, Nigrin b precursor; AAF37219, *Polygonatum multiflorum* ribosome inactivating protein RIPt; P11140, abrin-a precursor; BAA36395, *Lumbricus terrestris* 29-kDa galactose-binding lectin. **b:** Distance matrix of binding site residues, averaged for the first four proteins in the multiple-sequence alignment for which 3D structure data are available. Residue numbering corresponds to ricin B chain, 2AAI. The upper triangle of the matrix shows the average distance, and the lower triangle shows the corresponding standard deviation.

analysis. For example, UDP-GalNAc/polypeptide N-acetylgalactosaminyltransferase (GI number: 1582794) has Val23→Asn and Asp25→Ala (ricin B numbering) mutations; although Trp37 is replaced by a Phe, this is a conservative replacement.

The 3D structure data are available for four of the eight proteins in the nonredundant data set. Of these, the structures of mistletoe lectin I (1CE7) and endo-1,4-β-xylanase (1XYF) have been determined in the absence of bound galactose, whereas the other two (ricin B and ebulin) are with bound galactose. The relative distances of the binding site residues are similar in all four of the structures as indicated by the small standard deviations [<0.4; Fig, 4(b)]. This points to the conservation of the relative distances of the binding site residues in these proteins and hence, can be considered to be characteristic for this family. The similarity in the relative distances of the binding site residues with and without bound galactose implies that galactose binding is not accompanied by any significant conformational change in the binding pocket. This is similar to that observed in the galectin family of proteins (see above).

C-type animal lectin family

Lectins belonging to this family bind to galactose with the help of a Ca^{2+} ion which, besides interacting with the ligand, is also critical for maintaining the architecture of the binding site.[26] PSI-Blast analysis was performed for the two proteins that belong to this family (Table II). About 500 and 1400 hits were obtained for Gal-specific mutant MBP-A and tunicate C-type lectin, respectively. Literature scrutiny of these hits showed that some were specific to Gal, whereas others were specific to Man/GlcNAc. Some of the Gal-specific proteins were also able to bind GalNAc. Removal of redundancy from among the Gal-specific hits resulted in a nonredundant data set of eight proteins. This set included the rat MBP-A, which has been genetically engineered to confer Gal specificity by five substitution mutations (Glu185Gln, Asn187Asp, His189Trp, Gly190Tyr, and Ser191Gly) and an insertion of a Gly-rich loop (-His-Gly-Leu-Gly-Gly-) after residue 191.[27] Multiple-sequence alignment of the nonredundant data set proteins showed that the binding site residues Gln185, Asp187, Trp189, Glu198, Asn210, and Asp211 (Gal-specific mutant MBP-A numbering) are conserved in all the proteins of the nonredundant data set [Fig. 5(a)]. Ile212 is also conserved but has conservative replacements in few proteins.

A PROSITE-type signature sequence -[QE]-x-D-x-W-x(8)-E-x(11)-N-D-x(0,1)-[VIF]- has been deduced from the multiple-sequence alignment to represent the proteins belonging to this family, with the exception of tunicate C-type lectin (see below). Searching the PROSITE database, which includes the sequences from Swiss-Prot, TrEMBL, TrEMBL-new, and PDB, for this signature sequence re-

(a)

```
                          *  *  *        *              **  *
1AFA          180NWKKDQPDDWYGHGLGGGEDCVTIVDNGLWND-ISCQASH218
1DV8          234NWRPEQPDDWYGHGLGGGEDCAHFTDDGRWNDDVCQRPYR273
NP_031519     251NWAFTQPDNWQGHEQGGGEDCAEILSDGHWNDNFCQQVNR290
AAA41522      251NWAFTQPDNWQGHEEGGSEDCAEILSDGLWNDNFCQQVNR290
NP_001172     258NWAVTQPDNWHGHELGGSEDCVEVQPDGRWNDDFCLQVYR297
NP_071788     256HWAPKQPDNWYGHGLGGGEDCAHFTSDGRWNDDVCQRPYR295
AAD31028      254NWAPLQPDNWFGHGLGGGEDCAHITTGGPWNDDVCQRTFR293
NP_006335     238NWKPGQPDDWQGHGLGGGEDCAHFHPDGRWNDDVCQRPYH277

1TLG           82 WSPNEPSNPQ----SWQLCVQIWSKYNLLDD-V109
-------------------------------------------------

1HUP          187NWNEGEPN-----NAGSDEDCVLLLKNGQWNDVPCSTSHL221
1RDL          185NWNEGEPN-----NVGSGENCVVLLTNGKWNDVPCSDSFL219
P02707        164FWKEGEPN-----NRGFNEDCAHVWTSGQWNDVYCTYECY198
NP_066978     342YWNRGEPN-----NVGE-EDCAEFSGNG-WNDDKCNLAKF374
NP_571645     209NWGPNQPD-----NYKGAQDCGAIADSGLWDDVSCDSLYP243
AAF63470      204NWGPGQPD-----DYKGLQDCGVIEDTGLWDDGGCGDIRP238
O02659        206NWNDGEPN-----NASPGEHCVTLLSDGTWNDIACSASFL242
AAD45377      195NWNDGEPN-----NADSAEHCVEILKDGKWNDIFCSSQLS233
AAA82010      203NWNDGEPN-----NTGDGEDCVVILGNGKWNDVPCSDSFL237
BAA04983      294NWADGEPNN---SDEGQPENCVEIFPDGKWNDVPCSKQLL330
P42916        259NWAPGEPGNR-AKDEG-PENCLEIYSDGNWNDIECREERL296
P07439        111YWSSNNPNN---WENQDCGVVNYDTVTGQWDDDDCNKNKN147
```

(b)

Residue	Q185	D187	W189	E198	N210	D211	I212	Ca^{2+}
Q185		5.1	10.0	8.9	5.4	5.8	11.6	4.8
D187	0.2		6.2	6.0	7.7	5.4	10.1	4.0
W189	0.3	0.2		6.2	9.9	8.2	8.6	6.1
E198	0.0	0.0	0.3		7.3	4.1	5.1	4.8
N210	0.0	0.2	0.4	0.0		4.0	7.6	4.2
D211	0.1	0.0	0.3	0.1	0.1		6.4	3.1
I212	3.6	2.1	0.7	1.1	4.1	3.2		7.2
Ca^{2+}	0.2	0.1	0.2	0.1	0.2	0.0	2.9	

Fig. 5. **a:** Multiple-sequence alignment of eight Gal-specific (Block I; above the line), tunicate C-type lectin, and 12 Man/GlcNAc-specific (Block II; below the line) proteins that were identified by PSI-Blast analysis (and subsequent screening as described in Scheme 1) as homologues to proteins belonging to the C-type animal lectin family (see Table II). The eight Gal-specific proteins do not share >60% sequence similarity among them and hence, constitute a nonredundant data set. Pairwise alignment of 1AFA and 1TLG could be obtained only with an expected value of 100,000, and this too resulted in alignment in a short stretch; this was manually extended to include the neighboring residues and is shown here. The 12 Man/GlcNAc-specific proteins also constitute a similar nonredundant data set. Conserved binding site residues are shown in bold and marked by an asterisk (*) above the alignment. The PDB code, if available, or the GenBank accession number is given: Gal-specific proteins: 1AFA, as in Table II. 1DV8, carbohydrate recognition domain of human asialoglycoprotein receptor H1 subunit; NP_031519, mouse asialoglycoprotein receptor 2; AAA41522, rat asialoglycoprotein receptor; NP_001172, human asialoglycoprotein receptor 2 isoform a; NP_071788, rat Gal/GalNAc-specific lectin; AAD31028, mouse macrophage Gal/GalNAc-specific C-type lectin; NP_006335, human macrophage lectin 2 (calcium dependent). 1TLG, as in Table II. Man/Glc specific proteins: 1HUP, human mannose-binding protein lectin domain; 1RDL, rat mannose-binding protein, subtilisin digest fragment; P02707, chicken hepatic lectin; NP_066978, human CD209 antigen, dendritic cell-specific ICAM3-grabbing nonintegrin; NP_571645, *Danio rerio* mannose binding-like lectin; AAF63470, *Carassius auratus* mannose binding-like lectin precursor; O02659, bovine mannose-binding protein C precursor; AAD45377, *Sus scrofa* mannose-binding lectin; AAA82010, mouse mannose-binding protein C; BAA04983, bovine conglutinin precursor; P42916, bovine collectin-43; P07439, *Megabalanus rosa* lectin BRA-3 precursor. **b:** Distance matrix of binding site residues, averaged for the first two proteins in the multiple-sequence alignment for which 3D structure data are available. Residue numbering corresponds to Gal-specific mutant MBP-A (1AFA). The upper triangle of the matrix shows the average distance, and the lower triangle shows the corresponding differences; standard deviation was not calculated because only two entries are present.

sulted in 35 hits. All these hits are Gal/GalNAc-specific proteins; thus, this signature sequence may be considered as representative of the galactose-specific C-type lectin family.

The primary structure of the tunicate C-type lectin (1TLG) is reported to have 20–30% sequence similarity with the vertebrate C-type lectins.[28] However, its binding site sequence did not align correctly with the rest of the sequences during multiple-sequence alignment. A partial pairwise sequence alignment of tunicate C-type lectin (1TLG) with Gal-specific mutant MBP-A (1AFA) could be obtained with an expected value of 100,000; this was manually extended on either side to include other binding site residues [Fig. 5(a)]. However, despite the absence of overall sequence similarity, the binding site residues in tunicate C-type lectin, with the exception of Trp100, are well conserved both in sequence [Fig. 5(a)] and in 3D (Fig. 6). Galactose interacts with tryptophan in the same orientation in both Gal-specific mutant MBP-A and tunicate C-type lectin; however, the interaction of galactose with other binding site residues is different because of the change in the position of Trp100 (Fig. 6).[12]

The hits obtained from PSI-Blast analysis for the two proteins belonging to this family also included proteins that were specific to Man/GlcNAc. All the Man/GlcNAc-specific hits from PSI-Blast analysis were grouped, and a nonredundant data set of 12 proteins was obtained. Comparison of the multiple-sequence alignment of these Man/GlcNAc-specific proteins with that of Gal-specific proteins shows a five-residue deletion, which includes the aromatic residue that stacks against the *b* face of galactose [Fig. 5(b)]. This finding highlights the importance of the aromatic residue for galactose recognition.[29,30]

The 3D structure data are available for three of the nine proteins that form the nonredundant data set for this family. Of these, two have been crystallized with galactose (1AFA, 1TLG). The relative distances between the conserved binding site residues in Gal-specific mutant MBP-A (1AFA) and asialoglycoprotein receptor (1DV8) were found to be very similar, with the exception of Ile212 (1AFA) and Val267 (1DV8), which do not superpose on each other because of a gap between Asp211 and Ile212 in the sequence alignment [Fig. 5(a)]. In tunicate C-type lectin (1TLG), the relative distances of all the binding site residues except those of Trp100 are similar to those found in Gal-specific mutant MBP-A and the asialoglycoprotein receptor.

Legume lectin family

PSI-Blast analysis gave 400–600 hits for each protein belonging to this family, and the hits included both Gal- and Man/Glc/GlcNAc-specific proteins. Literature scrutiny was performed to select only those hits that are specific to Gal. Redundant entries were removed from among the Gal-specific proteins, which resulted in a nonredundant data set of 10 proteins, and a multiple-sequence alignment was performed for all the proteins.

Four loop regions, A, B, C, and D, constitute the binding site in this family of proteins; residues that are part of loops A, B, and C are well conserved [Fig. 7(a)].[31,32] It was also found that the length of loop D is small in Glc/Man-specific lectins compared with those specific for Gal/GalNAc [Fig. 7(a)].[32] Binding site residues that are part of loop D are not conserved either in sequence or in space[32,33] [Fig. 7(b)] and hence, aligning them even based on 3D structure is not unambiguous. Only one binding residue in loop A is conserved in all the proteins of this nonredundant data set (Asp89; *Ecor*L, 1AX1, numbering). One Gly residue in loop B is conserved, and the aromatic amino acid that stacks with the galactose and an Asn residue are conserved in loop C.

The 3D structure data are available for 6 of the 10 Gal-specific proteins in this family. Except for the structure of the bark lectin from *Robinia pseudoacacia* (1FNZ), the structures of five other proteins (Table II) have been determined with bound galactose. The average distance matrix calculated for the conserved binding site residues [Fig. 7(a)] in the six galactose-specific proteins of the legume lectin family shows that the relative distances of the binding site residues that are part of loops A, B, and C are similar as reflected by the small standard deviations [Fig. 7(c)].

Distance matrices for other proteins of the nonredundant data set (Table II)

PSI-Blast analysis of *Escherichia coli* heat-labile enterotoxin (1DJR) belonging to the toxin family gave about 50 hits. All the hits are toxins that have specificity toward galactose-containing oligosaccharides. Gal-specific toxins with known 3D structure data share 80–95% sequence similarity; the standard deviations in the average distances between the binding site residues are very small (data not shown). PSI-Blast analysis of jacalin (1JAC) gave 286 hits of which only jacalin was specific to galactose; all others were mannose-specific lectins. The distance matrix of binding site residues was calculated by considering only jacalin (data not shown). PSI-Blast analysis of the carbohydrate-binding domain (CBD) of neuraminidase (1EUU)[34] followed by literature scrutiny of the hits to select galactose-specific proteins resulted in galactose oxidases and CBDs of other neuraminidases. The alignment to galactose oxidase is partial and corresponds to the Gal-binding site in the noncatalytic N-terminal domain.[34a] In this alignment, only three of the five Gal-binding site residues (i.e., those that are within 4 Å from galactose) in neuraminidase (His539, Trp542, Arg572) are conserved in galactose oxidase (His40, Tyr43, Arg73); the other two (Ser575 and Glu578) align with glycine residues (Gly76 and Gly80). The 3D structure data for CBDs of other neuraminidases are not available. Hence, distance matrix of the binding site residues was computed (data not shown) by considering only the CBD of neuraminidase (1EUU) and the noncatalytic N-terminal domain galactose oxidase (1GOG). The relative His … Trp(Tyr) and Trp(Tyr) … Arg distances vary by >1 Å in the two proteins; this is indicative of the differences in the relative spatial distribution of the binding site residues in these two proteins.

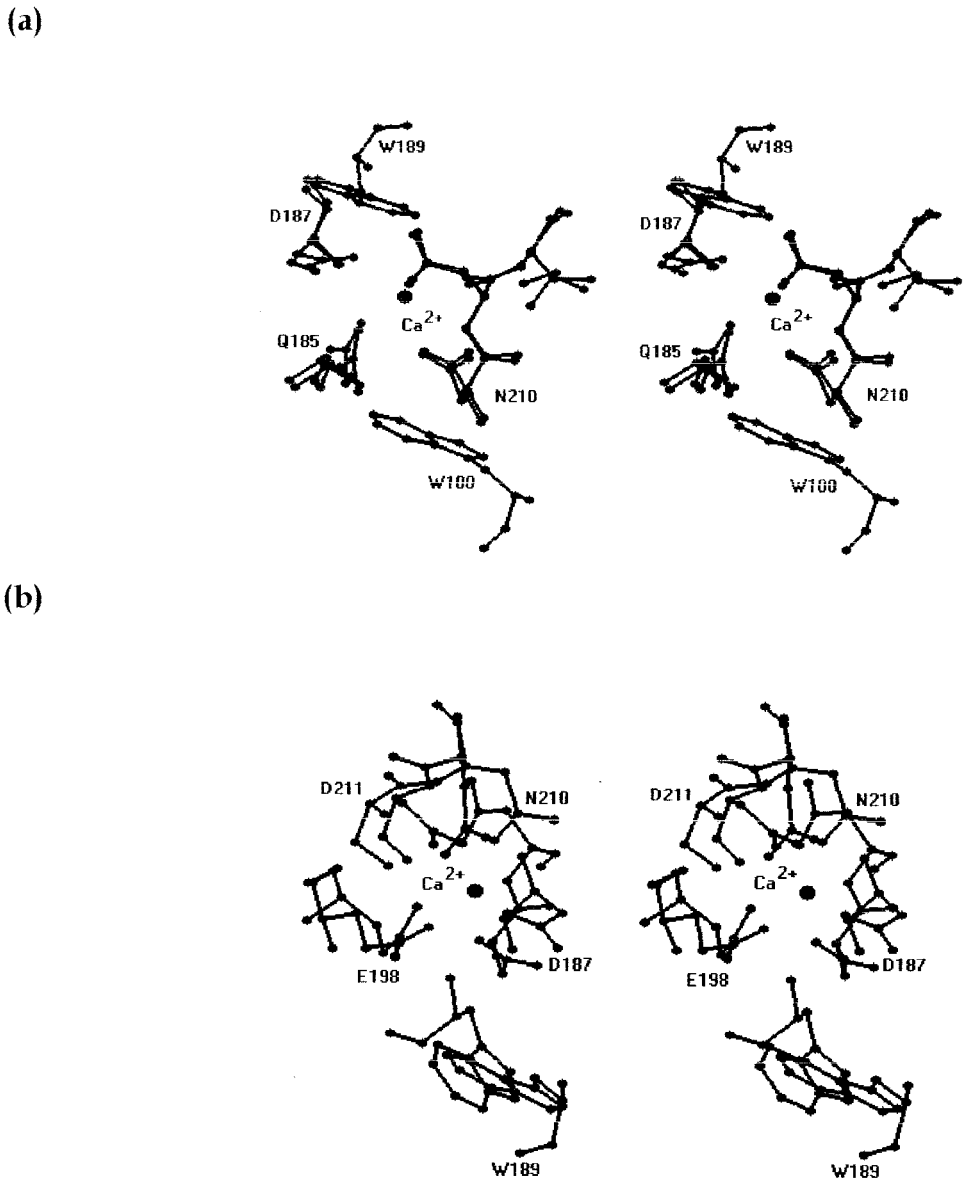


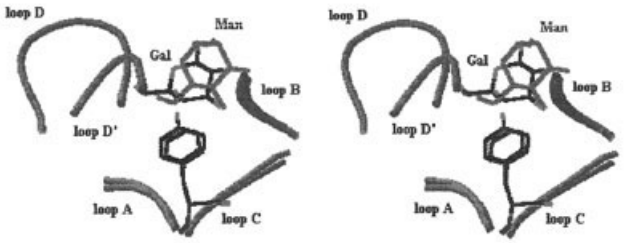
Fig. 6. Stereo diagrams showing the binding site residues of Gal-specific mutant MBP-A (1AFA) superposed over those of tunicate C-type lectin (1TLG). Both the proteins are members of the C-type animal lectin family. The superposition was done with respect to Asp211 (in 1AFA) and Asp108 (in 1TLG) (**a, top:** the bound galactose residues are not shown for clarity). Despite the lack of sequence similarity, the binding site architecture is superposable in the two proteins except for the spatial location of the stacking aromatic residue (Trp189 in 1AFA and Trp100 in 1TLG). The residues that superpose well in space align with each other in the sequence alignment [Fig. 5(a)]. However, when superposition is done with respect to the bound galactose (not shown for clarity), the stacking aromatic residues (Trp189 and Trp100), Glu198 (1AFA)/Glu86 (1TLG) and Ca^{2+} superpose well (**b, bottom**). In this superposition, the residues that superpose in 3D do not align with each other in sequence alignment [Fig. 5(a)]: for example, side-chain carboxyl groups of Asp187 (1AFA) and Asp107 (1TLG); similarly, Asp211 (1AFA) and Asp108 (1TLG). Thus, residues that align with each other in sequence alignment do not interact with the same hydroxyls of galactose.

Visual comparison of the average distance matrices of different families by considering the residues that interact with the same atom of the ligand showed that the distance matrices are distinct. For example, His158, Arg162, and Asn174 of galectin 3 (1A3K) interact with Gal:C4-OH, Gal:Ring O, and Gal:C6-OH; Trp181 is the stacking residue. The corresponding residues in ricin B chain (2AAI) are Asp22 (Gal:C4-OH), Asp25 (Gal:Ring O and Gal:C6-OH), and Trp37. Their relative distances [first row in Figs. 2(b) and 4(b)] are not the same. The distance matrices of heat-labile enterotoxin, of jacalin and of CBD of neuraminidase were also found to be distinct from each other and from those for the other four families. These differences are also borne out by the observed scattering of the interacting atoms around galactose when the binding site residues of all 18 proteins were superimposed (Fig. 8). The

(a)

Protein	Loop A	Loop B	Loop C	Loop D
	*	*	* *	
1AX1	^{83}TRPLPA**D**GLVF^{93}	^{101}KPAQGY**G**-YLGI^{111}	^{126}VEFDT-**FS**----**N**--PW-DPP^{138}	^{211}GLSGATGA----QRDAAE^{224}
1BZW	^{77}KDYDPA**D**GIIF^{86}	^{97}PAGSIG**G**GTLGV^{108}	^{120}VEFDT-**YS**----**N**S-EYNDP-^{133}	^{205}GFSAS-GS----L-GGRQ^{216}
1WBL	^{79}PRPHPA**D**GLVF^{91}	^{99}QTGEGG**G**-YFGI^{109}	^{121}VEFDT-**FR**----**N**--TW-DP-^{132}	^{205}GFSAATGDPSGKQRNATE^{222}
1F9K	^{82}AYPEPA**D**GLTF^{91}	^{100}PQGEDG**G**-NLGV^{110}	^{122}VEFDT-**FQ**----**N**--TW-DP-^{133}	^{208}GLSATTGY----QKNAVE^{221}
1G9F	^{83}TKRL-A**D**GLAF^{92}	^{100}KPQTHA**G**-YLGL^{110}	^{123}VEFDT-**F**----**RN**--SW-DPP^{135}	^{207}GFSAATGL----DIPGES^{220}
1FNZ	^{82}PATT-A**D**GLAF^{91}	^{99}QPLDLG**G**-MLGI^{109}	^{124}VEFDT-**F**----**SN**G-DW-DPK^{137}	^{209}GFSATTGI----DKGYVQ^{222}
9257007	^{82}SDG--V**D**GLAF^{90}	^{100}PSGSSA**G**-MFGL^{110}	^{125}VEFDT-**Y**FGKAY**N**--PW-DPD^{141}	^{213}GFSGGVG-----NAAEFE^{225}
P81371	^{82}ESKT-A**D**GIAF^{91}	^{100}PQ-KDG**G**-FLGL^{109}	^{122}VEFDT-**F**----**SN**--TW-DPS^{134}	^{206}GFSATSGL----SRDHVE^{219}
JQ1981	^{84}PSTAAT**D**GLAF^{94}	^{102}QPQSAG**G**-YLGL^{112}	^{128}VEFDT-**Y**----**YN**S-AW-DPQ^{141}	^{218}GFSATTGQ----TDNYIE^{231}
P05046	^{115}TKRL-A**D**GLAF^{124}	^{132}KPQTHA**G**-YLGL^{142}	^{155}VEFDT-**F**----**RN**--SW-DPP^{167}	^{239}GFSAATGL----DIPGES^{252}
1QMO	^{81}TSR-IA**D**GLAF^{91}	^{100}S--YHG**G**-FLGL^{108}	^{137}VEFDT-**YL**----**N**PD-YGDPN^{151}	^{222}GLSASTG-------QDKE^{232}
5CNA	^{204}-SH-PA**D**GIAF^{213}	^{222}PSGSTG**R**-LLGL^{232}	^{7}VELDT-**Y**P----**N**TD-IGDPS^{21}	^{92}GLSASTG-------LYKE^{102}
1FX5	^{82}PKAA-T**D**GLTF^{92}	^{99}PLRRAG**G**-YFGL^{109}	^{125}VEFDT-**I**-GSPV**N**--FW-DPG^{140}	^{214}GFSGSTY------IGRQA^{225}
2BQP	^{76}SYNV-A**D**GFTF^{86}	^{95}QTG--G**G**-YLGV^{103}	^{118}VEFDT-**F**----**YN**A-AW-DP-^{130}	^{210}GFSATTG-------AEYA^{220}
1LOA	^{76}SYNV-A**D**GFTF^{86}	^{95}QTG--G**G**-YLGV^{103}	^{118}VEFDT-**F**----**YN**T-AW-DP-^{130}	^{210}GFSATTG-------AEFA^{220}
Q39529	^{119}-NN-PG**D**GIAF^{128}	^{138}PPGSSG**G**-LLGL^{147}	^{164}VEFDT-**FV**----**N**NN-W-DPS^{177}	^{251}GFSGSTG-------GYVQ^{261}
S66356	^{119}-QD-PG**D**GFAF^{127}	^{135}-PGYGG**G**-LLGL^{144}	^{170}VEFDT-**YI**----**N**GQ-C-DPK^{183}	^{254}GFSASTG-------QNVE^{264}
P38662	^{71}TSR-IA**D**GLAF^{80}	^{89}S--YHG**G**-FLGL^{97}	^{114}VEFDTD**YL**----**N**PD-YGDPN^{129}	^{200}GLSASTG-------QNIE^{210}
AAF28739	^{81}PYY-AA**D**GFAF^{91}	^{100}PPNSWG**K**-FLGL^{110}	^{138}VEFDT-**FP**----**N**AN-I-DPN^{151}	^{222}GLSASTG-------EEKQ^{232}
AAB36103	^{86}KGGTIA**D**GLTF^{97}	^{106}PSKIEG**E**-YLGV^{116}	^{133}CEFDL-**YK**----**N**G--I-DPS^{145}	^{223}GISGCSG-------LQVS^{233}
AAA74576	^{84}DNG--A**D**GIAF^{93}	^{102}PKNSAG**G**-TLGL^{112}	^{129}VEFDT-**FY**AQDS**N**G--W-DPN^{145}	^{216}GFSAASG-------QQYQ^{226}
Q01806	^{107}SSNV-A**D**GLAF^{117}	^{226}QNIGRA**G**-FLGV^{236}	^{151}VEIDT-**FH**----**N**--TW-DP-^{162}	^{236}GFSAATG-------AEFA^{246}
CAA42938	^{103}TYNV-A**D**GLAF^{113}	^{122}KSIHHG**G**-YLGV^{132}	^{147}VEIDT-**FY**----**NA**-QW-DPN^{160}	^{238}GFSSSTG-------AEYS^{248}
P02874	^{84}INRG-G**D**GITF^{94}	^{101}QPKSGG**G**-YLGI^{111}	^{124}VEFDT-**FS**----**N**--RW-DPA^{136}	^{207}GLSAATG-------DLVE^{217}

(b)



(c)

	D89	G107	F131	N133
D89		5.4	5.4	8.0
G107	0.2		7.6	6.2
F131	0.2	0.1		5.1
N133	0.1	0.3	0.1	

Fig. 7. **a:** Multiple-sequence alignment corresponding to the four loop regions of the sugar-binding site in proteins obtained following Scheme 1 for the legume lectin family. The first five entries were used for PSI-Blast. Block I (above the line) contains 10 sequences that are specific to Gal, and Block II (below the line) contains 14 Glc/Man/GlcNAc specific proteins. Conserved binding site residues are shown in bold and marked by an asterisk (*). Alignment in loop D has been manually altered based on 3D structure information. The PDB code, if available, or the GenBank accession number is given: Gal-specific lectins: 1AX1, 1BZW, 1WBL, 1F9K, and 1G9F as in Table II. 1FNZ, bark Lectin from *Robinia Pseudoacacia;* 9257007, lectin Uea-Ii; P81371, seed lectin (*VML*); JQ1981, lectin II-scotch broom; P05046, lectin precursor (Agglutinin) (SBA); Man/Glc/GlcNAc-specific lectins: 1QMO, fril; 5CNA, concanavalin A; 1FX5, *Ulex europaeus* lectin I; 2BQP, pea lectin; 1LOA, legume lectin (isolectin I); Q39529, agglutinin II precursor (ClAII); S66356, mannose/glucose-binding lectin CLAI precursor-*Cladrastis lutea;* P38662, *Dolichos lablab* lectin; AAF28739, mannose lectin FRIL (*Phaseolus vulgaris*); AAB36103, insecticidal N-acetylglucosamine-specific lectin (*Griffonia simplicifolia*); AAA74576, mannose/glucose-binding lectin precursor; Q01806, lectin I precursor; CAA42938, lectin (LEC2) (*Medicago truncatula*); P02874, lectin. **b:** Stereo representation of loops that form the binding site in legume lectins. The binding site regions of Gal-specific *Erythrina corallodendron lectin* (1AX1) is superposed on Man/Glc-specific concanavalin (5CNA). Loops A, B, and C superpose on each other very well. Large differences can be seen in the specificity-determining loop D region (D in 1AX1, D′ in 5CNA) of the two proteins. **c:** Distance matrix of conserved binding site residues are shown, averaged for the first six proteins in the multiple-sequence alignment (Block I) for which 3D structure data are available. Residue numbering corresponds to *Erythrina corallodendron* lectin (1AX1). The upper triangle of the matrix shows the average distance, and the lower triangle shows the corresponding standard deviation. The distance matrix of 1FNZ (which does not have bound galactose) is very similar to those of other five proteins (which have bound galactose).

distances between the interacting atoms/residue (Table IV) within each protein were calculated. The relative distances between any two pairs of interacting atoms show significant variation (Table V); for example, the distance between the atoms hydrogen bonding with C2-OH and the stacking residue varies from 4.8 to 10.1 Å; similar variations were observed in the distance between most of the atom pairs. Only the distance between the stacking residue and the atom hydrogen bonding with ring oxygen atom is very nearly the same in all the proteins (varies from 7.3 to 8.0 Å). This finding indicated that the spatial disposition of binding site residues is not conserved among the 18 galactose-binding proteins (Fig. 8). However, the variation in the relative distances was quite low for proteins within the same family (data not shown).

Stacking Aromatic Amino Acid Residue Is Solvent Accessible

A different approach was adopted to elucidate the common features of the galactose-binding sites because the relative distances were found to be similar only within the families. One of the common features shared by all the proteins is the presence of an aromatic residue stacking against the *b* face of galactose (Fig. 1). Hence, the characteristics of the stacking aromatic amino acid residue were analyzed in the 18 proteins of the nonredundant data set. Although, in general, aromatic residues are buried in proteins, the residue that stacks against galactose was found to have an average absolute solvent accessibility of 98 $Å^2$; the values in the 18 proteins range from 50 to 144 $Å^2$ (Table VI). This is not surprising because all these proteins bind the sugar in a shallow surface groove.[16] It was also observed that the stacking aromatic residue is part of a strand, a coil, or a bend (Table VI). The residues, whose secondary structure type is a strand, are found to be toward the end of the strand and close to a coil.

Galactose Can Slide Along the Plane of the Stacking Residue to Establish Optimal Interactions With Binding Site Residues

The side-chain of the stacking aromatic residue is planar and in principle can provide stacking interactions on either of the two sides. However, it was noticed from visual inspection that the bound galactose and the main-chain atoms of the stacking residue lie on opposite sides/faces of the aromatic ring in all the 18 proteins of the nonredundant data set (Fig. 1). To establish this correlation quantitatively, the position of galactose relative to the stacking aromatic residue was determined (Table VI) in polar coordinates in a frame of reference defined as shown in Figure 9. The polar coordinate θ of the C4 atom of galactose was found to be correlated to the dihedral angle χ_2 (Cα-Cβ-Cγ-Cδ1): θ is <90 (i.e., galactose in the positive *z*-axis direction) when χ_2 is negative and θ is >90 (i.e., galactose in the negative *z*-axis direction) when χ_2 is positive; the correlation coefficient is 0.94. The correlation between θ and χ_2 can be rationalized by viewing the stacking aromatic residue as forming the base of the binding pocket. The amino acid residues that form hydrogen-bonding interactions with galactose will form the rest of the binding site, leaving one side open for the entry of the ligand (Fig. 10). Such an arrangement will not be possible if galactose were to bind on the side of the main-chain atoms of the aromatic residue.

The values of ϕ of Gal:C4 atoms show large variations within the data set of 18 proteins (Table VI). These variations reflect the variability in the position of galactose with reference to the plane of the aromatic ring in different proteins. Galactose can stack on top of either of the two ring systems of the tryptophan side-chain or can be some where in between also. Variations in ϕ are also due to a different set of nonpolar hydrogen atoms of galactose being used for stacking in different proteins. The H3, H4, and H5 atoms are above Trp181 in human galectin-3, and the H61 and H62 atoms are pointing away; in contrast, the H5 and H6 atoms are above Trp542 in neuraminidase (1EUU), and the H3 and H4 atoms are pointing away (Fig. 1). This finding illustrates the differences in the modes of binding of galactose relative to the aromatic amino acid residue in these proteins.

The value of the polar coordinate *r* representing the distance of the C4 atom of galactose from the stacking aromatic residue varies from 3.5 to 5.4 Å (Table VI). The magnitude of the nonbonded interactions, which is dependent on distance, between galactose and the stacking aromatic residue may thus be expected to vary in different proteins. Taken together, the variations observed in *r*, θ, and ϕ suggest that galactose has sufficient freedom to slide along the plane of the stacking aromatic residue to establish optimal interactions with other residues constituting the binding site. Such a freedom resulting in optimal interactions with the binding site residues would have been restricted or absent if the main-chain atoms and the galactose-binding site were to be on the same side of the stacking aromatic residue.

The Atom That Forms Hydrogen Bond With Gal:O4 Is in a Solvent-Shielded Environment

The presence of hydrogen-bonding groups around Gal:O4 is the other characteristic feature of galactose-binding proteins. The O4-hydroxyl group is axial in galactose, whereas it is equatorial in glucose and mannose and is thus the most important recognition point for determining specificity. It is invariably stabilized by hydrogen-bonding interactions with the side-chain of a polar residue in all the proteins (Table IV). The total absolute solvent-accessible area for the residue that hydrogen bonds with Gal:O4 was found to vary between 2 and 72 $Å^2$, indicating that this residue, as a whole, can be either buried or solvent exposed. However, the absolute solvent-accessible area for the specific hydrogen-bonding atom in this residue was found to be <10 $Å^2$ in all 18 proteins of the data set (minimum: 0 in 1JAC; maximum: 8.6 in 2GAL; average 3.7; median 3.0), indicating that this atom is in a solvent-shielded environment. Gal:O4 is the specificity-determining group; solvent shielding of the atom that hydrogen bonds with this group probably enhances the contribution of this interaction to the binding energy.

The residue that forms hydrogen bond with Gal:O4 was found to be part of either a strand or a coil region in the 18

TABLE V. Variation of Distance Between Protein Atoms That Interact With Galactose[†]

Protein atoms interacting with	Protein atoms interacting with				
	-OH at C-3	-OH at C-4	Ring O	-OH at C-6	Stacking residue
-OH at C-2	3.4–5.5	5.1–8.7	4.7–9.9	8.6–12.0	4.8–10.1
-OH at C-3		2.1–5.5	4.0–8.5	6.2–10.9	5.2–8.0
-OH at C-4			4.0–7.5	3.5–10.5	4.7–8.2
Ring O				3.4–7.8	7.3–8.0
-OH at C-6					4.6–8.7

[†]Atoms that form hydrogen bonds to or stack against galactose were identified (tabulated in Table IV). The distances (in Å) between each pair of such atoms were calculated for each protein and the lowest and highest values are tabulated here. When more than one atom interacts with the same hydroxyl group/ring oxygen atom of galactose, a pseudoatom was used to represent the interacting atoms. The aromatic residue was also represented by a pseudoatom, defined as the arithmetic average of the ring atoms.

TABLE VI. Some Characteristics of the Stacking Aromatic Amino Acid Residues

PDB code[a]	Stacking aromatic residue	Absolute/relative solvent accessibility[b]	Secondary structure type[c]	χ_2[d]	Gal:C4 position[e]		
					r	θ	ϕ
1AX1	Phe131	64/32	Bend	−88	3.9	13	63
1G9F	Phe128	89/45	Strand	86	4.2	153	23
1WBL	Phe126	50/25	Bend	74	4.5	145	10
1BZW	Tyr125	101/48	Bend	−83	3.8	16	52
1F9K	Phe127	54/27	Bend	86	4.8	130	5
1A3K	Trp181	120/48	Strand	107	4.2	149	15
1C1L	Trp70	98/39	Coil	115	4.2	151	10
1GAN	Trp69	130/52	Strand	105	4.3	148	20
1SLT	Trp68	123/50	Strand	117	4.4	142	22
1HLC	Trp65	120/48	Strand	107	4.6	138	31
2GAL	Trp69	134/54	Strand	110	4.7	139	32
1AFA	Trp189	127/51	Coil	41	3.6	158	172
1TLG	Trp100	69/28	Strand	21	4.6	147	66
2AAI	Trp37	102/41	Strand	−109	4.0	34	−30
1HWM	Trp39	88/35	Strand	−96	3.7	9	88
1JAC	Tyr78	144/67	Strand	−82	4.0	23	116
1DJR	Trp88	58/23	Strand	−75	3.5	5	70
1EUU	Trp542	65/26	Coil	−92	5.3	18	12

[a]The names of the protein are given in Table II.
[b]The absolute solvent accessibility values are in Å^2. The relative solvent accessibility values are calculated as the percent accessibility compared to the accessibility of that residue type in an extended ALA-x-ALA tripeptide (199.48 Å^2 for Phe. 212.76 Å^2 for Tyr and 249.36 Å^2 for Trp).
[c]The secondary structure type was identified following the DSSP assignments (E, strand: C, coil; and S. bend) given in the protein data bank.
[d]The dihedral angle χ_2 was calculated for the four atoms Cα-Cβ-Cγ-Cδ1.
[e]The position of the C4 atom of galactose with reference to the stacking aromatic residue is given in polar coordinates. The definition of the coordinate system used is given in Figure 9.

proteins taken for analysis. The position of the atom that hydrogen bonds with Gal:O4 with reference to the stacking aromatic residue was determined by using polar coordinates and it was observed that r, θ, and ϕ vary over a range of values (Table VII), the variations arising because of the variability in the galactose-aromatic residue-stacking interactions (vide supra). The spatial positions of the atoms that hydrogen bond with other hydroxyl groups of galactose were also determined with reference to the stacking aromatic residue (Table VIII); these also vary over a range for a similar reason.

The Deduced Features Fairly Uniquely Characterize the Galactose-Binding Sites in Lectins

To determine if the features inferred are unique to galactose-binding sites, a C computer program *COTRAN* was developed to incorporate these features. *COTRAN* uses the PDB file along with solvent-accessibility values and DSSP secondary structure assignments as input, identifies potential galactose-binding sites and outputs the corresponding stacking aromatic residue. For convenience, the program uses different frames of reference, depending on whether χ_2 is positive or negative. Different sets of criteria were used for Trp and Phe/Tyr because the orientation of galactose with reference to these residues is not exactly identical. The steps followed in *COTRAN* are as follows:

1. Identify the secondary structure type of aromatic residue that has absolute solvent accessibility > 50 Å^2 (Table VII). This criterion alone was able to eliminate nearly 75% of the aromatic residues present in the test proteins (see below) as not part of a Gal-binding site.

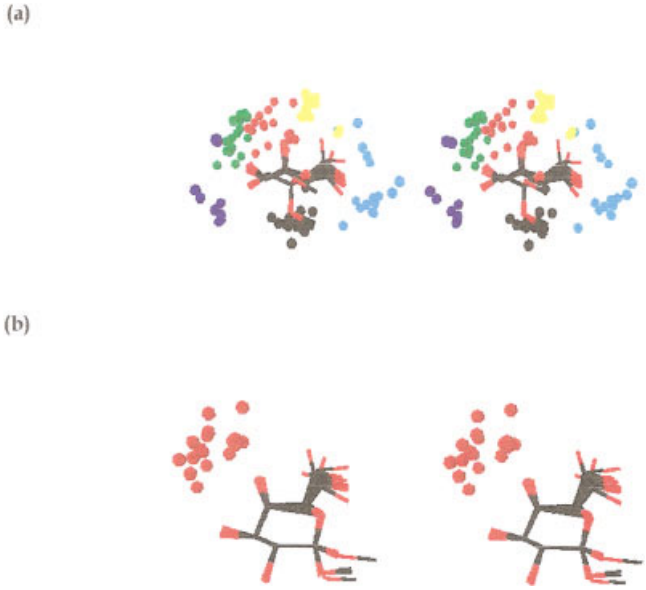


Fig. 8. Stereo diagrams showing the superposition of atoms hydrogen bonding with galactose (Table IV) in the 18 proteins of the nonredundant data set (Table II). Superposition was done with reference to the bound galactose (residue numbering shown in Table II). All the hydrogen-bonding atoms are shown in (**a, top**). The color code used in (a) are as follows: atoms hydrogen bonding to Gal:O2, blue; Gal:O3, green; Gal:O4, red; Gal:O5, yellow; Gal:O6, cyan. Pseudoatom representing the stacking residue is shown as white sphere. The scattering of the atoms indicates that the spatial disposition of interacting atoms is not conserved in the proteins. The scattering can be more clearly seen in (**b, bottom**) where only those atoms that hydrogen bond with Gal:O4 are shown. Similar scattering of atoms was observed (figure not shown) even when the stacking residue is used as reference for superposition.

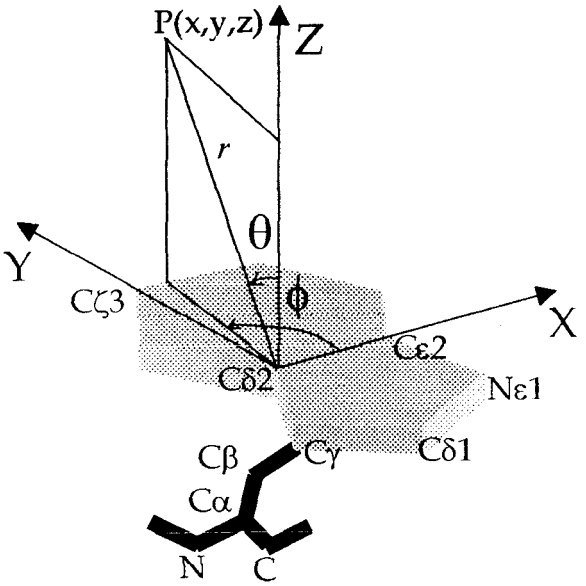


Fig. 9. Definition of the frame of reference used for calculating the polar coordinates (r, θ, and ϕ) with reference to the stacking tryptophan residue. The atom Cδ2 is used as the origin and the bond Cδ2-Cε2 is along the x axis. The y axis has been defined by using the atom Cη2 in such a way that the plane of the tryptophan ring lies in the xy plane.

2. Check for the presence of a cavity for accommodating galactose. For this, no protein atom should be present within the following ranges of polar coordinates: $4.0 < r < 6.5$, $30 < \theta < 55$, $0 < \phi < 65$ for Trp and $4.2 < r < 6.2$, $19 < \theta < 48$, $-15 < \phi < 65$, and $\phi > 11\tilde{0}$ for Phe/Tyr. These limits were derived from the positions of galactose in the 18 proteins of the data set used for analysis.
3. Check for the presence of at least one atom that can potentially hydrogen bond with Gal:O4 to satisfy the secondary structure type, solvent accessibility, and relative position with respect to the stacking residue as in Table VII. At this stage, nearly 95% of the aromatic residues were eliminated as not part of a Gal-binding site.
4. Check for the presence of at least one atom that can form hydrogen bonds with Gal:O6 and at least one atom that can hydrogen bond with either Gal:O3 or Gal:O5 to satisfy the spatial position criteria both with respect to stacking aromatic residue (Table VIII) and to the atom that can form hydrogen bond with Gal:O4.

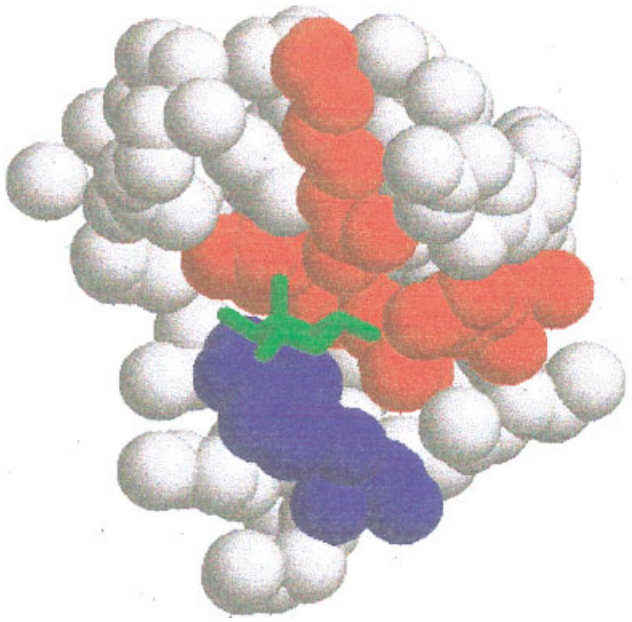


Fig. 10. Space filling diagram showing the residues within 10 Å from bound galactose (shown in green) in human galectin 3 (1A3K). The stacking aromatic residue is shown in blue, whereas the residues that form hydrogen bonds (His158, Arg162, Asn174, Glu 184) with galactose are shown in red. The binding site can be visualized as having the stacking residue as its base, one side of the binding site being lined with the hydrogen-bonding atoms and the other side being open for ligand entry.

COTRAN was first run for the nonredundant data set of 18 proteins (Table II). These proteins collectively have 351 aromatic amino acid residues (285 Phe/Tyr and 66 Trp). Of these, 13 Trp and 8 Phe/Tyr were identified as stacking aromatic residues of potential Gal-binding sites. This included all 18 (12 Trp and 6 Phe/Tyr) expected binding sites. The additional binding sites are two in ebulin (1HWM; stacking residue Trp162 and Phe 249) and one in ricin B (2AAI; stacking residue Tyr248). Ricin does have two binding sites for galactose,[35] even though only one (stacking residue Trp37) was used for analysis and Tyr248

TABLE VII. Characteristic Features of the Galactose-Binding Site

Feature		Tryptophan	Tyrosine or phenylalanine
Stacking aromatic residue			
Absolute solvent accessibility		>50	
Secondary structure type		Strand. coil, or bend	
Galactose and main-chain atoms		Opposite sides of the planar ring	
Atom hydrogen bonding to Gal:O4			
Side-chain atom		Carboxyl oxygen; amide, imidazole, and guanidine nitrogen	
Absolute solvent accessibility		<10	
Secondary structure type		Strand or coil	
Position relative to stacking residue (range)[a]	r	4.3 to 9.4	4.5 to 8.0
	θ	28 to 70	15 to 40
	ϕ	−125 to 37	−100 to −56

[a]In view of the correlation of θ with χ_2, the frames of reference for determining the polar coordinates were defined differently depending on whether χ_2 is positive or negative for convenience of analysis. This ensures that the bound ligand is always in the positive Z-direction (i.e., $\theta < 90$). Such a change in the frame of reference has no effect on the inferences drawn from the results. When $\chi_2 < 0$: Origin at Cζ3 of Trp (Cγ for Phe/Tyr), X-axis along Cζ3-Cγ (Cγ-Cδ1 for Phe/Tyr), and Y-axis such that Cε2 (Cε1 for Phe/Tyr) is in the first quadrant. When $\chi_2 > 0$: Origin at Cγ of Trp/Phe/Tyr, X-axis along Cγ-Cζ3 (Cγ-Cδ2 for Phe/Tyr), and Y-axis such that Cη2 (Cε2 for Phe/Tyr) is in the first quadrant.

TABLE VIII. Characterization of Atoms That Form Hydrogen Bond With Galactose

Position relative to stacking residue	Tryptophan	Tyrosine or phenylalanine
Atom hydrogen bonding to Gal:O2		
Atom/molecule	Nε2, Nδ2, Nζ, HOH	
r	7.5 to 12.5	4.3 to 6.5
θ	38 to 75	69 to 88
ϕ	−65 to 105	−27 to −8
Atom hydrogen bonding to Gal:O3		
Atom/molecule	Oδ1, Oδ2, Oε1, Oε2, Nε2, Nδ2, Nη1, Nη2, N, Nζ, HOH	
r	4.0 to 9.7	6.4 to 8.6
θ	35 to 87	36 to 57
ϕ	−95 to 65	−56 to −10
Atom hydrogen bonding to Gal:O5		
Atom/molecule	N, Nη1, Nη2, Oγ, Oδ1, Oδ2, HOH	
r	6.8 to 9.3	6.9 to 8.5
θ	5 to 34	0 to 20
ϕ	−34 to 110	33 to 127
Atom hydrogen bonding to Gal:O6		
Atom/molecule	Oη, Oδ1, Oδ2, Oε1, Oε2, Nε2, Nδ2, Nη1, Nη2, HOH	
r	3.2 to 7.9	5.1 to 10.3
θ	0 to 75	32 to 77
ϕ	−155 to 143	−84 to 125

is the stacking residue in the second binding site. In ebulin, Phe249 is the residue corresponding to Tyr248 of ricin and the second binding sites of ricin and ebulin are superimposable. The third binding site involving Trp162 of ebulin is a false positive.

At least one atom capable of forming hydrogen bond with Gal:O6 should be present in the putative Gal-binding site as per step 4 of *COTRAN* (see above). In the crystal structure of the CBD of neuraminidase (1EUU), Gal:O6 is forming hydrogen bond neither with protein atom nor with solvent (Table IV). However, *COTRAN* identifies 1EUU as having Gal-binding site because the r, θ, and ϕ values of His539:Nϵ2 (7.451, 62.224, 11.432) and of Glu522:Oϵ2 (6.140, 74.896, 45.786) are within the ranges expected for the atom hydrogen bonding to Gal:O6 (Table VIII). Identification of 1EUU as having a Gal-binding site by *COTRAN* despite Gal:O6 being not hydrogen bonded in the crystal structure can be attributed to the large range observed in r, θ, and ϕ values in the 18 proteins of the data set.

The 18 proteins constituting the data set and used for inferring the features of Gal-binding site (Table II) belong to six *fold* types (seven nonhomologous protein families; Tables II and III). It can be seen from the SCOP database that a large number of proteins share this *fold* type, many of which have no known ability to bind galactose. Some of the protein families that share these *fold* types are lectins, glucanases, toxins, xylanases, superantigens, tRNA synthetases, proteins involved in replication and translation, EGFs, interleukins, plant cytotoxins, and so forth. *COTRAN*, when run on all these proteins as input (total of 757 proteins; a redundant data set), identified the Gal-binding sites with very high specificity and sensitivity (Table IX). Together, there are 3,677 Trp and 16,606 Phe/Tyr residues in these proteins, and close to 75% of these do not have the characteristic solvent accessibility and secondary structure type of the stacking aromatic residue of the galactose-binding site. A literature scrutiny was undertaken to find if the proteins identified by *COTRAN* have been experimentally shown to bind to galactose to assess true/false hits. Some of the false positives and false negatives are redundant entries, and the sensitivity and specificity values given in Table IX will be better if *COTRAN* is run on a nonredundant data set.

TABLE IX. Statistics of Galactose-Binding Sites Identified by *COTRAN*

Fold	Total PDB files	Hits		Non-hits		Specificity	Sensitivity
		TP[a]	FP[b]	TN[c]	FN[d]		
Concanavalin A-like lectins/glucanases	216	41	8	163	4	0.84	0.91
C-type lectin-like	73	11	1	61	0	0.92	1
OB fold	305	23	12	268	2	0.66	0.92
Galactose-binding domain-like	52	2	2	44	4	0.5	0.33
Beta-trefoil[e]	104	16	3	90	1	0.84	0.94
Beta-prism 1	10	1	1	7	1	0.5	0.5

[a]TP: True-positive proteins that bind to galactose.
[b]FP: False-positive proteins that have been identified as galactose binding by the program, but they do not bind galactose as per the literature.
[c]TN: True-negative proteins that do not bind galactose and also have not been identified by the program as galactose binding.
[d]FN: False-negative proteins that have been reported to bind galactose but not being identified by the program as galactose binding.
[e]The total number of hits and non-hits (110) are more than the total number of PDB files (104) because some of the proteins have more than one Gal-binding site.

Minor Violations in the Criteria Used by *COTRAN*, Mainly in Steps 3 and 4, Lead to the Identification of Known Gal-Binding Sites as Non-Hits (False Negatives)

All the Gal-binding sites present in proteins having the C-type lectin-like *fold* were identified by *COTRAN* (sensitivity = 1; Table IX). In proteins belonging to other five *fold* types, some Gal-binding proteins were not identified by *COTRAN* (false negatives) resulting in sensitivity being <1. All the four false negatives (1FAY, 1QF3, 1WBF, and 2TEP) in the Concanavalin A-like lectins/glucanases *fold* (Table X) are redundant entries (of 1F9K, 1BZW, 1WBL, and 1BZW, respectively, which are identified as true positives) solved either at a lower resolution or with a different ligand. In all four cases, the atom hydrogen bonding to Gal:O4 is marginally outside the range expected by *COTRAN* (Table VIII) in step 3. Both the false negatives for the OB fold (Table X) are mutants of heat-labile enterotoxin, one of the 18 proteins used for Gal-binding site analysis (1DJR; Table II). Both are identified as non-hits because *COTRAN* could not find an atom that can hydrogen bond with Gal:O3/Gal:O5 (step 4). The four false negatives for the Galactose-binding domain-like *fold* (Table X) are galactose oxidases (redundant entries). *COTRAN* identifies these four as non-hits either because it finds one/two atoms within the expected cavity region (step 2; atom violation by <1 Å; 1GOG, 1GOF, and 1GOH) or for the lack of an atom that can hydrogen bond with Gal:O4 (1K3I). *COTRAN* identifies a potential Gal-binding site with Trp265 as the stacking aromatic residue in *Amaranthus caudatus* agglutinin (1JLX; with β-trefoil *fold*). However, in the crystal structure, Trp265 is not in the immediate neighborhood of the bound Gal (~6 Å from Gal); hence, this was considered as a false negative (Table X). It is of interest that galactose is bound to this protein (1JLX) in an unusual *upside down* orientation with its pyranose ring nonpolar hydrogen atoms facing the solvent, instead of stacking against an aromatic residue as observed in all other Gal-specific proteins. The lone false negative identified for the β-prism I *fold* (Table X) corresponds to *Maclura pomifera* lectin Mpa (1JOT); in this case, the solvent accessibility of the Gal:O4 hydrogen-bonding atom (13 $Å^2$) is slightly more than that expected by *COTRAN* (<10 $Å^2$; step 3). From this analysis, it is clear that small changes in the conformation of binding site residues are likely to make these as true positives. Alternatively, if additional high-resolution structures of Gal-specific proteins become available, they can be included in the data set used to derive parameters for *COTRAN,* thereby increasing the sensitivity.

The Putative Gal-Binding Site in Several False Positives Appears to Have Good Similarity to Known Gal-Binding Sites

COTRAN identified a total of 27 proteins as having a Gal-binding site from all six fold families, but these have not been reported as Gal-binding proteins in the literature. Hence, these are treated as false positives (Tables IX and X). Of these, only 20 proteins are unique; the other seven are either the same protein crystallized/studied under different conditions or a close homologue. Among these 27 proteins, 7 are xylanases (1ENX, 1QH7, 1XND, 1F5J, 1UKR, Concanavalin A-like lectins/glucanases *fold*; 1GMM and 1GNY, Galactose-binding domain-like *fold*). It is of interest that xylanases belonging to the *fold* type β-trefoil (1XYF and 1ISV) have been shown to bind galactose.[24a–c,25] The putative galactose-binding site identified by *COTRAN* in 13 of the 27 false positives appears by visual inspection to have good similarity to genuine galactose-binding sites. Such false positives are aspartyl tRNA synthetase (1EFW, stacking aromatic residue Trp24; 1EQR, Trp23; 1G51, Trp24), superantigen from *Streptococcus pyogenes* (1EU3, Trp108), shiga-like toxin (1QOH, Trp130), various xylanases/xylan-inding proteins (1GMM, Trp92; 1F5J, Trp18; 1UKR, Trp44; 1ENX, Trp18; 1XND, Trp18; 1GNY, Trp176; 1QH7, Trp19),and toxic shock syndrome toxin-1 from *S. aureus* (3TSS, Trp80). It remains to be experimentally verified if these proteins do indeed bind galactose because unrelated proteins can have similar binding sites: a 3D cluster of side-chains implicated in drug binding in influenza sialidase has been found to be similar to the side-chains involved in isocitrate binding in *Escherichia coli* isocitrate dehydrogenase.[36]

TABLE X. PDB IDs of Proteins Used for Validating *COTRAN*†

1. Concanavalin A-like lectins/glucanases fold
   True positives (41)
   1A3K 1A78 1AX0 1AX1 1AX2 1AXY 1AXZ 1BKZ 1BZW 1C1F 1CIL 1CIW 1CR7 1F9K 1FNZ
   1FYU 1G9F 1GAN 1HDK 1HLC 1HQL 1LCL 1LEC 1LED 1LTE 1QKQ 1QMJ 1SBD 1SBE 1SBF
   1SLA 1SLB 1SLC 1SLT 1WBL 2GAL 2PEL 2SBA 3GAL 4GAL 5GAL
   False positives (8)
   1AXK (Trp346) 1CPN (Trp161) 1ENX (Trp18) 1F5J (Trp18) 1GLH (Trp94) 1QH7 (Trp19)
   1UKR (Trp44) 1XND (Trp18)
   True negatives (163)
   1A39 1A8D 1AF9 1AJK 1AJO 1APN 1AVB 1AZD 1B09 1BCX 1BJQ 1BK1 1BQP 1BVV 1BXH 1BYH 1C4R 1C57 1C5H
   1C5I 1CEL 1CES 1CJP 1CN1 1CON 1CPM 1CVN 1D0H 1D2S 1DBN 1DFQ 1DGL 1DHK 1DIW 1DLL 1DQ0 1DQ1 1DQ2
   1DQ4 1DQ5 1DQ6 1DY4 1DYK 1DYM 1DYP 1DZQ 1EG1 1EGN 1ENQ 1ENR 1ENS 1EPW 1F3L 1F5F 1FAT 1FNY 1FV2
   1FV3 1FX5 1G7Y 1G8W 1GBG 1GIC 1GKB 1GNH 1GNZ 1GPI 1GSL 1H8V 1H9P 1H9W 1HIX 1HV0 1HVL 1I3H 1IKP
   1IKQ 1ILE 1IOA 1JBC 1JHN 1JOJ 1JW6 1KIT 1LEM 1LEN 1LES 1LGB 1LGC 1LGN 1LOA 1LOB 1LOC 1LOD 1LOE
   1LOF 1LOG 1LU1 1LU2 1LUL 1MAC 1NLR 1NLS 1ONA 1OVW 1PVX 1QDC 1QDO 1QGL 1QH6 1QMO 1QNW 1QNY 1QOO
   1QOS 1QOT 1QU0 1RED 1REE 1REF 1RIN 1SAC 1SCR 1SCS 1SLI 1SLL 1TEI 1VAL 1VAM 1VIW 1VLN 1XNB 1XNC
   1XYN 1XYO 1XYP 1YNA 2A39 2AYH 2BQP 2BVV 2CEL 2CNA 2CTV 2ENR 2LAL 2LTN 2NLR 2OVW 2SLI 3BTA 3CEL
   3CNA 3ENR 3OVW 3SLI 4CEL 4OVW 4SLI 5CEL 5CNA 6CEL 7CEL
   False negatives (4)
   1FAY 1QF3 1WBF 2TEP
2. C-type lectin-like fold
   True positives (11)
   1AFA 1AFB 1AFD 1BCH 1BCJ 1BCP 1BYF 1DV8 1FIF 1FIH 1TLG
   False positive (1)
   1PRT (Trp26)
   True negatives (61)
   1B08 1B6E 1BJ3 1BNL 1BUU 1BV4 1C3A 1CWV 1DY0 1DY1 1DY2 1E5U 1E87 1E8I 1EGG 1EGI 1ESL 1F00 1F02
   1FM5 1FVU 1G1Q 1G1R 1G1S 1G1T 1H8U 1HQ8 1HTN 1HUP 1HYR 1IOD 1IXX 1JSK 1JWI 1K9I 1K9J 1KCG 1KMB
   1KOE 1LIT 1MSB 1PRE 1PTO 1QDD 1QO3 1RDI 1RDJ 1RDK 1RDL 1RDM 1RDN 1RDO 1RTM 1TN3 1TSG 1YTT 2AFP
   2KMB 2MSB 3KMB 4KMB
   False negatives (0)
3. OB fold
   True positives (23)
   1CHQ 1CT1 1DJR 1EEF 1EEI 1EFI 1FD7 1FGB 1G8Z 1HTL 1LT3 1LT4 1LT5 1LT6 1LTA
   1LTB 1LTI 1LTR 1LTS 1LTT 2BOS 2CHB 3CHB
   False positives (12)
   1BCP (Trp26) 1EFW (Trp24) 1EQR (Trp23) 1EU3 (Trp108) 1G51 (Trp24) 1I3Q (Phe942)
   1JMC (Trp212) 1KAW (Trp88) 1PRT (Trp26) 1PYS (Trp270) 1QOH (Trp130) 3TSS (Trp80)
   True negatives (268)
   117E 1A0I 1A1D 1A2T 1A2U 1A3T 1A3U 1A3V 1A62 1A63 1A8V 1AE2 1AE3 1AEX 1AH9 1AN8 1ASY 1ASZ 1AW7
   1B1Z 1B3Q 1B70 1B7Y 1B8A 1B9M 1B9N 1BBU 1BBW 1BDX 1BKB 1BOS 1BOV 1BQQ 1BR9 1BUV 1BVS 1BXT 1C0A
   1C48 1C4Q 1C7Y 1C9O 1CHP 1CKM 1CKN 1CKO 1CQF 1CQV 1CSP 1CSQ 1CUK 1CZG 1CZW 1D1K 1D2B 1D5M 1D5X
   1D5Z 1D6E 1D7Q 1D8L 1DGS 1DGT 1DLI 1DM0 1DYQ 1E1O 1E1T 1E22 1E24 1E3P 1E6A 1E7Z 1E9G 1EIF 1EIY
   1ENA 1ENC 1ENF 1EOV 1ESF 1ET6 1ET9 1EU4 1EUJ 1EWC 1EWI 1EY0 1EY4 1EY5 1EY6 1EY7 1EY8 1EY9 1EYA
   1EYC 1EYD 1EYG 1EZ6 1EZ8 1F77 1FAJ 1FGU 1FL0 1FNU 1FNV 1FNW 1FR3 1FVI 1G29 1G6P 1GD7 1GKH 1GM5
   1GO3 1GOZ 1GPC 1GUG 1GUN 1GUO 1GUS 1GUT 1GVP 1H95 1H9J 1H9K 1H9M 1H9R 1H9S 1HH2 1HJP 1HQR 1HUK
   1HXY 1HZ9 1HZA 1HZB 1HZC 1I40 1I4G 1I4H 1I4P 1I4Q 1I4R 1I4X 1I50 1I5F 1I6H 1I6T 1IGP 1IL2 1INO
   1IPW 1JB3 1JB7 1JC7 1JCK 1JE5 1JFD 1JJC 1JOK 1JOO 1JOQ 1JOR 1JT8 1K0R 1K0S 1K28 1K83 1K8G 1KAA
   1KAB 1KDA 1KDB 1KDC 1KIX 1KRS 1KRT 1LYL 1MJC 1MJW 1MJX 1MJY 1MJZ 1NMF 1NMG 1NSN 1NUC 1OBW 1OTC
   1PFS 1PTO 1QEZ 1QIL 1QNU 1QUQ 1QVC 1RIP 1RL2 1SBB 1SE2 1SE3 1SE4 1SEB 1SND 1SNO 1SNP 1SNQ 1SRO
   1STA 1STB 1STE 1STG 1STH 1STN 1STY 1SXT 1SYB 1SYC 1SYD 1SYE 1SYF 1SYG 1TII 1TS2 1TS3 1TS4 1TS5
   1UEA 1VQA 1VQB 1VQC 1VQD 1VQE 1VQF 1VQG 1VQH 1VQI 1VQJ 1WGI 1WGJ 1XTC 1YHA 1YHB 1YPP 2A8V 2EIF
   2EIP 2ENB 2GN5 2GVB 2NUC 2PRD 2QIL 2SEB 2SNM 2SOB 2TMP 2TSS 3MEF 3NUC 3SEB 3ULL 4TSS 4ULL 5NUC
   5TSS 8PRK
   False negatives (2)
   1B44 1LTG
4. Galactose-binding domain-like fold
   True positives (2)
   1EUT 1EUU
   False positives (2)
   1GMM (Trp92) 1GNY (Trp176)
   True negatives (44)

TABLE X. (Continued)

1BGL 1BGM 1BHG 1CIY 1CX1 1CZS 1CZT 1CZV 1D7P 1DLC 1DP0 1DYO 1F49 1F4A 1F4H 1GHO 1HNL 1I5P 1IQD
1J83 1J84 1JHJ 1JI6 1JU3 1JU4 1JYN 1JYW 1JYX 1JYY 1JYZ 1JZ0 1JZ1 1JZ2 1JZ3 1JZ4 1JZ5 1JZ6 1JZ7
1JZ8 1NUK 1ULO 1ULP 1XNA 1XNT

False negatives (4)

1GOF 1GOG 1GOH 1K3I

5. Beta-trefoil fold

True positives (16)[b]

1CE7 1HWM 1HWN 1HWO 1HWP 2AAI 2MLL 1ISV 1ISX 1ISY 1ISZ

False positives (3)

1ILR (Trp16) 1IRP (Trp17) 1JLX (Trp265)

True negatives (90)

1A8D 1ABR 1AF9 1AFC 1AVA 1AVU 1AVW 1AVX 1AXM 1BA7 1BAR 1BAS 1BFB 1BFC 1BFF 1BFG 1BLA 1BLD 1CVS
1D0H 1DFC 1DFQ 1DIW 1DJS 1DLL 1DQG 1DQO 1DZC 1DZD 1E0O 1EPW 1EV2 1EVT 1EYL 1F31 1FGA 1FMM 1FMZ
1FN0 1FQ9 1FV2 1FV3 1FWU 1FWV 1G82 1HCD 1HCE 1HIB 1IHK 1II4 1IIL 1IJT 1ILB 1ILE 1ILT 1IOB 1IRA
1ISW 1IT0 1ITB 1JLY 1JQZ 1JT3 1JT4 1JT5 1JT7 1JTC 1K5U 1K5V 1QQK 1QQL 1RML 1TIE 1WBA 1WBC 1XYF
2AFG 2AXM 2BFH 2FGF 2ILA 2ILB 2IRT 2MIB 2WBC 3ILB 3BTA 4FGF 4WBC 9ILB

False negative (1)

1JLX

6. Beta-prism I fold

True positive (1)

1JAC

False positive (1)

1VMO (Trp270)

True negatives (7)

1C3K 1C3M 1C3N 1CIY 1DLC 1I5P 1JI6

False negative (1)

1JOT

[†]The statistics are given in Table IX.

[b]The total number of true positives is counted as 16 because 1HWM, 1HWN, 1HWO, 1HWP, and 2AAI have two Gal-binding sites.

DISCUSSION

A large number of biochemical and structural studies have been conducted to characterize lectin-carbohydrate interactions.[33,37–40] Several attempts have been made to characterize the sugar-binding site features and to understand the origin of carbohydrate specificity in lectins. On the basis of a detailed analysis of the structure/specificity relationship within the whole group of plant lectins, it has been inferred that some carbohydrates (mannose, chitin, Gal/GalNAc) are recognized by multiple structurally different carbohydrate-binding motifs.[40a] While noting that each family of sugar-binding proteins has evolved a unique stereochemistry at the binding site to achieve specificity, Elgavish and Shaanan observed that ligand-dependent stereochemistry of the hydrogen-bonding pattern around the C4-OH group, together with the preferential disposition of aromatic residues, plays a key role in eliciting primary specificity.[13] With the aim of providing a framework for understanding the molecular basis of sugar specificity and to arrive at a rationale for the redesign of ligand-binding propensities, Sharma and Surolia[32] conducted an extensive analysis of sequences and 3D structures of several legume lectins; from this, they showed that the size of the binding site loop D is possibly a primary determinant of saccharide specificity in these proteins. Similar conclusions were also arrived at by computer-modeling studies; in addition, it was observed that loop B of the binding site is important in discriminating between Gal and GalNAc in legume lectins.[15] Recently, from an analysis of the characteristic properties of sugar-binding sites in a set of 19 sugar-binding proteins, it was observed that certain amino acids (aromatic residues, Arg, Asp, and Glu) show a strong propensity to be in the sugar-binding site; it was also observed that no single recognition template exists for binding carbohydrates because proteins were found to bind to sugars in many different ways.[40]

With the assumption that common recognition principles exist for common ligand recognition, a nonredundant data set of 18 proteins (Table II) was analyzed in the present study to determine the common features of galactose-binding sites. This data set included proteins belonging to seven nonhomologous protein families (i.e., with no detectable sequence similarity across the families). Even the overall folds of the protein families are different (Table III). Comparison of the family-specific multiple-sequence alignments clearly showed the dissimilarity in the nature of amino acid residues and functional groups that constitute the binding sites in these proteins. This dissimilarity was confirmed by the absence of any *different family* hits when the sequence database was scanned with PROSITE-type signature sequences inferred for the galectin and C-type animal lectin families. The distance matrices are also different for the different families (Figs. 2, 4, 5, and 7).

However, the common principles of the binding site became apparent when solvent accessibility and secondary structure types of binding site residues were characterized (Table VII). Furthermore, the reason for the differences in the distance matrices became obvious when the position

TABLE XI. Comparison of Ligand-Bound and Ligand-Free Structures of Lectins

Protein	Ligand specificity	PDB code		Largest difference in the distance matrix (Å)
		Ligand-bound	Ligand-free	
Peanut lectin	Galactose	1BZW	1CQ9	0.2
Congerin I	Galactose	1C1L	1C1F	0.1
Tunicate C-type lectin	Galactose	1TLG	1BYF	0.1
Human galectin 7	Galactose	2GAL	1BKZ	0.1
Gal-specific mutant MBP-A	Galactose	1AFA	1AFD	0.1
Endoglucanase Cel5A	Glucose	1E5J	1A3H	0.1
Carbohydrate-binding module of xylanase 10A	Glucose	1I8A	1I8U	0.2
Cyclodextrin glycosyltransferase	Glucose	1A47	1CIU	0.2
Concanavalin A	Mannose/Glucose	5CNA	1GKB	0.6

and orientation of galactose were represented in polar coordinates: variations in ϕ (Table VI) are suggestive of the stacking aromatic residue serving a platform on which the ligand slides to optimize its interactions with the hydrogen-bonding groups of the binding pocket. Because of this variability in the relative positions of galactose and stacking residue, the distances of the latter with respect to other binding site residues also vary.

The program *COTRAN,* which incorporated the deduced features to search for potential galactose-binding sites, displays very high sensitivity and specificity (Table IX). The search for the presence of hydrogen-bonding groups relative to the stacking residue included a range of r, θ, and ϕ values in *COTRAN.* In principle, it is possible to have a distance matrix representation with a similar range of values to characterize the binding site. However, such a representation will have low specificity compared to *COTRAN:* because distance is a scalar quantity, some of the information content regarding the relative spatial disposition is lost in distance matrix representation. Representation in terms of the polar coordinates preserves the information regarding the spatial arrangement, and this probably is a reason for the high sensitivity and specificity achieved by *COTRAN.* Thus, representing binding site features in the form of polar coordinates and combining other structural features, such as secondary structure type and solvent accessibility, although simplistic, seems to be an elegant diagnostic approach.

The binding sites of ligand-free and ligand-bound forms were compared for some of the lectins, for which 3D structure data are available, to determine the extent of conformational changes caused by ligand binding. The residues that constitute the binding site were identified by using the ligand-bound form, and the relative distances of these residues were calculated in both ligand-bound and ligand-free forms. The deviation in the distance between any pair of equivalent binding site residues was found to be very small (Table XI), suggesting that ligand-induced conformational changes are negligible in these proteins. In fact, comparison of the 3D structures of the *Erythrina corallodendron* lectin (*Ecor*L) and of its complexes with Gal, GalNAc, lactose, and N-acetyllactosamine showed that galactose is bound in an identical way in all four complexes and that no conformational change occurs in the protein on binding the ligand.[9] Thus, *COTRAN* is able to predict galactose-binding sites with a very high specificity and sensitivity. The absence of conformational changes on ligand binding has been observed even in other protein families: comparison of the bound and ligand-free structures of proteins belonging to the lysozyme, desthiobiotin synthase, Cyt P450-CAM, papain, trypsin, D-xylose isomerase, chymotrypsin, and thymidine kinase families showed that the structures of the binding sites are preserved on ligand binding.[41] However, it should be remembered that some proteins do undergo significant conformational changes on ligand binding.[42–44] The nature and extent of such ligand-induced conformational changes are not known a priori and also vary from protein to protein. This finding will be a major limitation for the development of knowledge-based methods for identifying ligand-binding sites.

The mode of binding of Gal to legume lectins is different from that of Man/Glc, even though both the ligands bind in the same pocket.[15,32] The binding site residues present in loops A, B, and C are conserved in both Gal- and Man/Glc-specific proteins (Fig. VII). The distance matrices are very nearly identical, and multiple-sequence alignment cannot distinguish between Gal- and Man/Glc-specific proteins. However, the size of the binding site loop D is small in Man/Glc-specific lectins compared with those that are specific to Gal and are thus responsible for ligand specificity.[15,32] In view of small size, loop D is in proximity of the stacking aromatic residue in Man/Glc-specific proteins; thus, relative to the stacking residue, within a specific range of r, θ, and ϕ (vide supra), there is no cavity (step 2 of *COTRAN*) for binding the sugar in these proteins. Man/Glc, because they bind in a different mode, are placed differently relative to the stacking residue [Fig. 7(b)]. This is the reason why *COTRAN* very effectively distinguishes the Gal- and Man/Glc-specific proteins of the legume lectin family.

D-Fucose (6-deoxygalactose), L-arabinose, and D-galactose are homomorphous sugars differing only in the nature of exocyclic group at C-5 atom: $-CH_3$ in D-fucose, $-H$ in L-arabinose, and $-CH_2OH$ in D-galactose. The presence of an atom that can hydrogen bond with Gal:O6 is checked by *COTRAN* (step 4). Proteins that bind to L-arabinose and D-fucose are not expected to have such an atom and hence, will not be identified as a hit by *COTRAN.* Only the 3D structures of L-arabinose specific proteins are known

(5ABP and 2ARC). Expectedly, *COTRAN* does not find any Gal-binding site in these proteins (non-hits).

From an analysis of the protein-adenylate complexes, it was found that no recognition motif in terms of specific residue/ligand interactions exists for adenylate binding; however, certain properties of the protein/adenylate interactions were found to be common.[45] These common properties were related to the shape and polarity of the environment around the ligand; these were used to create a composite description of the adenylate-binding site. This was termed as a fuzzy recognition template because these proteins displayed many different specific ways to recognize adenylate. It was also observed that such fuzzy recognition of ligands can be highly discriminatory even among very similar ligands.[46] The results obtained from the analysis of protein-galactose complexes in the present study are strongly suggestive of a similar fuzzy recognition template for recognition of galactose also. Even in this instance, the features are able to discriminate between two closely related monosaccharides (i.e., galactose and mannose/glucose).

The ability of a protein to recognize, bind, and differentiate between different ligands lies in the nature of the binding site it possesses rather than its overall structure. Hence, knowledge of the binding site features will enable assigning functions at the biochemical level to proteins with known 3D structures. Such knowledge can also be used to model the 3D structure of a protein known from biochemical studies to have such a binding site and to design new ligand-binding sites into a protein of known 3D structure.[47] Hence, active site characterization studies, such as the present one, are quite handy in designing new drugs. They are also useful in generating enzymes with altered sugar specificity for the chemoenzymatic synthesis of carbohydrates.

CONCLUSIONS

The 3D structures of a set of 18 nonredundant galactose-specific proteins belonging to seven distinct families have been analyzed, and the common features shared by the binding sites have been inferred. These features have been found to fairly uniquely characterize the galactose-binding sites. Family-specific distance matrices show that relative distances between the binding site residues of different members of the family are well conserved. The matrix generated for each family could be used individually to identify galactose-binding sites in other proteins with known protein structure; they can also be used to model new galactose-binding sites in proteins. A PROSITE-type signature sequence has been inferred for the galectin and C-type lectin-like *fold* family proteins.

Materials and Methods

Databases and Web Tools

The 3D structures of proteins and protein-ligand complexes were retrieved from the Protein Data Bank (October 2002 release).[48] The SCOP database was used to identify the folds of the proteins.[49] Secondary structure assignments were from DSSP.[50] Absolute and relative solvent-accessible surface areas were computed by using NACCESS2.1.1 on a Sun Solaris platform with a probe radius of 1.4 Å.[51] ClustalW was used for multiple-sequence alignment.[52] The NCBI server was used for PSI-Blast analysis and pairwise sequence alignment.[53] Default parameters were used for PSI-Blast and sequence alignments. PSI-Blast analysis was performed against the *nr* database and was iterated until no new hits were obtained. Swiss-PDBviewer 3.7 and RasMol were used for 3D structure visualization, superposition, and other such manipulations.[54,55] C programs and shell scripts were developed in-house for all other analyses and were run under Linux environment.

Choosing a Nonredundant Data Set for 3D Binding-Site Analysis

A total of 151 protein-galactose complex structures were obtained from the protein databank with use of the key words HET:GAL, HET:GLA, HET:GLB, HET:MGA, and HET:AMG for the search. All the structures so obtained were solved by X-ray crystallography. The hits included structures of the same protein determined at different resolutions or with different ligands. Such redundant entries were excluded by considering the entries that correspond to higher resolution. This resulted in the exclusion of 80 hits, and literature scrutiny showed that only 20 of the remaining 71 hits have been experimentally shown to be specific to galactose; the others were proteins that have been crystallized with a galactose containing oligosaccharide. The set of 20 proteins included arabinose-binding protein and Glc/Gal-binding protein. These two proteins envelop the carbohydrate ligand in a deep pocket in contrast to other proteins, which bind the saccharide ligand in a shallow surface groove.[16] Hence, these two were excluded from the data set. The final data set contains 18 proteins belonging to 7 nonhomologous protein families (Table II). One of the C-type animal lectins is a mannose-binding protein, which has been genetically engineered by insertion and site-specific substitution mutations to specifically bind galactose but not mannose.[27] Jacalin belongs to a family of mannose-binding proteins but is specific to galactose.[56] Pairwise sequence comparison of these 18 proteins showed that they share no more than 63% sequence similarity among them (Table XII). The pairwise sequence similarity varied from 38 to 63% among legume lectins, from 26% to 50% among galectins, and was 48% between the two ricin B-like proteins. There was no detectable sequence similarity among other protein pairs. Thus, this set of 18 proteins constituted a nonredundant data set belonging to 7 nonhomologous protein families.

Distance Matrix Construction

The relative distances of the galactose-binding site residues were represented in the form of a distance matrix. Each binding site residue was represented by a pseudoatom, defined as the arithmetic average of all the atoms of the residue. The distances between all pairs of pseudoatoms representing the binding site residues within a

TABLE XII. Pairwise Sequence Similarity Among the Proteins Considered for Analysis†

PDB codes of the two proteins		Percent sequence similarity	PDB codes of the two proteins		Percent sequence similarity
1AX1	1G9F	52	1A3K	1HLC	32
1AX1	1WBL	63	1A3K	2GAL	36
1AX1	1BZW	40	1C1L	1GAN	33
1AX1	1F9K	59	1C1L	1SLT	34
1G9F	1WBL	46	1C1L	1HLC	35
1G9F	1BZW	41	1C1L	2GAL	27
1G9F	1F9K	46	1GAN	1SLT	50
1WBL	1BZW	46	1GAN	1HLC	37
1WBL	1F9K	63	1GAN	2GAL	41
1BZW	1F9K	42	1SLT	1HLC	43
1A3K	1C1L	26	1SLT	2GAL	34
1A3K	1GAN	29	1HLC	2GAL	30
1A3K	1SLT	29	2AAI	1HWM	48

†There was no detectable sequence similarity among other protein pairs.

protein were calculated, and the average distances over all the proteins belonging to a family constituted the distance matrix for that family. 3D structures of even uncomplexed proteins were used for the distance matrix computation.

Sensitivity and Specificity

Sensitivity, a parameter that reflects the ability of a method to detect true positives (TP), has been defined as

$$\text{Sensitivity} = TP/(TP + FN)$$

where FN denotes false negatives. A method that has the highest sensitivity (i.e., 1) will identify all true positives and will have no false negatives. The definition of sensitivity does not include false positives.

Specificity, a parameter that reflects the ability of a method to reject false positives (FP), has been defined as

$$\text{Specificity} = TP/(TP + FP).$$

This definition excludes false negatives. By definition, the values of sensitivity and specificity range between 0 and 1.

ACKNOWLEDGMENTS

We thank Profs. P. Jayadeva Bhat and Y.U. Sasidhar for helpful discussions throughout the course of this work. The authors also thank Prof. S. Durani for discussions and critical reading of the manuscript. MSS is grateful to the Indian Institute of Technology Bombay for teaching assistantship.

Availability of *COTRAN*

COTRAN will be available on request from the authors for academic use.

REFERENCES

1. Todd AE, Orengo CA, Thornton JM. Evolution of function in protein superfamilies, from a structural perspective. J Mol Biol 2001;307:1113–1143.
2. Nagano N, Orengo CA, Thornton JM. One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. J Mol Biol 2002;321:741–765.
3. Makarova KS, Grishin NV. Thermolysin and mitochondrial processing peptidase: how far structure–functional convergence goes. Protein Sci 1999;8:2537–2540.
4. Krem MM, Cera ED. Molecular markers of serine protease evolution. EMBO J 2001;20:3036–3045.
5. Kimber MS, Pai EF. The active site architecture of *Pisum sativum* β-carbonic anhydrase is a mirror image of that of α-carbonic anhydrases. EMBO J 2000;19:1407–1418.
6. Akahani S, Hidenori I, Nangia-Makker P, Raz A. Galectin-3 in tumor metastasis. Trends Glycosci Glycotech 1997;9:69–75.
7. Kaltner H, Stierstorfer B. Animal lectins as cell adhesion molecules. Acta Anat 1998;161:162–179.
8. Perillo NL, Marcus ME, Baum LG. Galectins: versatile modulators of cell adhesion, cell proliferation, and cell death. J Mol Med 1998;76:402–412.
9. Elgavish S, Shaanan B. Structures of the Erythrina corallodendron lectin and its complexes with mono- and di-saccharides. J Mol Biol 1998;277:917–932.
10. Lobsanov YD, Gitt MA, Leffler H, Barondes SH, Rini JM. 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 1993;268: 27034–27038.
11. Liao DI, Kapadia G, Ahmed H, Vasta GR, Herzberg O. Structure of S-lectin, a developmentally regulated vertebrate beta-galactoside-binding protein. Proc Natl Acad Sci USA 1994;91:1428–1432.
12. Poget SF, Legge GB, Proctor MR, Butler PJ, Bycroft M, Williams RL. The structure of a tunicate C-type lectin from *Polyandrocarpa misakiensis* complexed with D-galactose. J Mol Biol 1999;290:867–879.
13. Elgavish S, Shaanan B. Lectin—carbohydrate interactions: different folds, common recognition principles. Trends Biochem Sci 1997;22:462–467.
14. Quiocho FA, Vyas NK. Atomic interactions between proteins/enzymes and carbohydrates. In: Hecht SM, editor. Bioorganic chemistry: carbohydrates. New York: Oxford University Press; 1999. p 441–457.
15. Rao VSR, Lam K, Qasba PK. Architecture of the sugar binding sites in carbohydrate binding proteins—a computer modeling study. Int J Biol Macromol 1998;23:295–307.
16. Rini JM. Lectin structure. Annu Rev Biophys Biomol Struct 1995;24:551–577.
17. Sundari CS, Balasubramanian D. Hydrophobic surfaces in saccharide chains. Prog Biophys Mol Biol 1997;67:183–216.
18. Iobst ST, Drickamer K. Binding of sugar ligands to a Ca^{2+}-dependent animal lectins. J Biol Chem 1994;269:15512–15519.
19. van Damme EJM, Hao Q, Charels D, Barre A, Rouge P, van Leuven F, Peumans WJ. Characterization and molecular cloning of two different type 2 ribosome-inactivating proteins from monocotyledonous plant *Polygonatum multiflorum*. Eur J Biochem 2000;267:2746–2759.
20. Hamelryck TW, Loris R, Bouckaert J, Dao-Thi M-H, Strecker G, Imberty A, Fernandez E, Wyns L, Etzler ME. Carbohydrate binding, quarternary structure and a novel hydrophobic binding site in two legume lectin oligomers from *Dolichos biflorus*. J Mol Biol 1999;286:1161–1177.
21. Falquet L, Pagni M, Bucher P, Hulo N, Sigrist CJ, Hofmann K, Bairoch A. The PROSITE database, its status in 2002. Nucleic Acids Res 2002;30:235–238.
22. Dodd RB, Drickamer K. Lectin-like proteins in model organisms: implications for evolution of carbohydrate-binding activity. Glycobiology 2001;11:71R–79R.
23. Swaminathan GJ, Leonidas DD, Savage MP, Ackerman SJ, Acharya KR. Selective recognition of mannose by the human eosinophil Charcot-Leyden-Crystal protein protein (galectin-10): a crystallographic study at 1.8 Å resolution. Biochemistry 1999;38: 13837–13843.

24a. Kuno A, Kaneko S, Ohtsuki H, Ito S, Fujimoto Z, Mizuno H, Hasegawa T, Taira K, Kusakabe I, Hayashi K. Novel sugar-binding specificity of the type XIII xylan-binding domain of a family F/10 xylanase from *Streptomyces olivaceoviridis* E-86. FEBS Lett. 2000;482:231–236.

b. Fujimoto Z, Kuno A, Kaneko S, Kobayashi H, Kusakabe I, Mizuno H. Crystal structures of the sugar complexes of *Streptomyces olivaceoviridis* E-86 xylanase: sugar binding structure of

the family 13 carbohydrate binding module. J Mol Biol 2002;316: 65–78.
c. Notenboom V, Boraston AB, Williams SJ, Kilburn DG, Rose DR. High-resolution crystal structures of the lectin-like xylan binding domain from *Streptomyces lividans* xylanase 10A with bound substrates reveal a novel mode of xylan binding. Biochemistry 2002;41:4246–4254.
25a. Fujimoto Z, Kuno A, Kaneko S, Yoshida S, Kobayashi H, Kusakabe I, Mizuno H. Crystal structure of *Streptomyces olivaceoviridis* E-86 beta-xylanase containing xylan-binding domain. J Mol Biol 2000;300:575–585.
b. Hirabayashi J, Dutta SK, Kasai K. Novel galactose-binding proteins in Annelida. Characterization of 29-kDa tandem repeat-type lectins from the earthworm *Lumbricus terrestris*. J Biol Chem 1998;273:14450–14460.
26. Drickamer K. Engineering galactose-binding activity into a C-type mannose-binding protein. Nature 1992;360:183–186.
27. Kolatkar AR, Weis WI. Structural basis of galactose recognition by C-type animal lectins. J Biol Chem 1996;271:6679–6685.
28. Suzuki T, Takagi T, Furukohri T, Kawamura K, Nakauchi M. A calcium-dependent galactose-binding lectin from the tunicate *Polyandrocarpa misakiensis*. Isolation, characterization, and amino acid sequence. J Biol Chem 1990;265:1274–1281.
29. Iobst ST, Drickamer K. Selective sugar binding to the carbohydrate recognition domains of the rat hepatic and macrophage asialoglycoprotein receptor. J Biol Chem 1996;271:6686–6693.
30. Weis WI, Drickamer K. Structural basis of lectin-carbohydrate recognition. Annu Rev Biochem 1996;65:441–473.
31. Young NM, Oomen RP. Analysis of sequence variation among legume lectins. A ring of hypervariable residues forms the perimeter of the carbohydrate-binding site. J Mol Biol 1992;228:924–934.
32. Sharma V, Surolia A. Analyses of carbohydrate recognition by legume lectins: size of the combining site loops and their primary specificity. J Mol Biol 1997;267:433–445.
33. Bouckaert J, Hamelryck T, Wyns L, Loris R. Novel structures of plant lectins and their complexes with carbohydrates. Curr Opin Struct Biol 1999;9:572–577.
34. Gaskell A, Crennell S, Taylor G. The three domains of a bacterial sialidase: a beta-propeller, an immunoglobulin module and a galactose-binding jelly-roll. Structure1995;3:1197–1205.
34a. Ito N, Phillips SE, Yadav KD, Knowles PF. Crystal structure of a free radical enzyme, galactose oxidase. J Mol Biol 1994;238:794–814.
35. Rutenber E, Robertus JD. Structure of ricin B-chain at 2.5 Å resolution. Proteins 1991;10:260–269.
36. Poirette AR, Artymiuk PJ, Grindley HM, Rice DW, Willett P. Structural similarity between binding sites in influenza sialidase and isocitrate dehydrogenase: implications for an alternative approach to rational drug design. Protein Sci 1994;3:1128–1130.
37. Poveda A, Asensio JL, Espinosa JF, Martin-Pastor M, Canada J, Jimenez-Barbero J. Applications of nuclear magnetic resonance spectroscopy and molecular modeling to the study of protein-carbohydrate interactions. J Mol Graph Model 1997;15:9–17, 53.
38. Qasba PK. Involvement of sugars in protein-protein interactions. Carbohydr Polymers 2000;41:293–309.
39. Srinivas VR, Reddy GB, Ahmad N, Swaminathan CP, Mitra N, Surolia A. Legume lectin family, the "natural mutants of the quaternary state," provide insights into the relationship between protein stability and oligomerization. Biochim Biophys Acta 2001; 1527:102–111.
40. Taroni C, Susan J, Thornton JM. Analysis and prediction of carbohydrate binding sites. Protein Eng 2000;13:89–98.
40a. Peumans WJ, Barre A, Hao Q, Rouge P, van Damme EJM. Higher plants developed structurally different motifs to recognize foreign glycans. Trends Glycosci Glycotech 2000;12: 83–101.
41. Fradera X, de la Cruz X, Silva CHTP, Gelpi JL, Luque FJ, Orozco M. Ligand-induced changes in the binding sites of proteins. Bioinformatics 2002;18:939–948.
42. Ramakrishnan B, Qasba PK. Crystal structure of lactose synthase reveals a large conformational change in its catalytic component, the beta1,4-galactosyltransferase-I. J Mol Biol 2001;310:205–218.
43. Sharff AJ, Rodseth LE, Spurlino JC, Quiocho FA. Crystallographic evidence of a large ligand-induced hinge-twist motion between the two domains of the maltodextrin binding protein involved in active transport and chemotaxis. Biochemistry 1992; 31:10657–10663.
44. Varrot A, Schulein M, Davies GJ. Insights into ligand-induced conformational change in Cel5A from *Bacillus agaradhaerens* revealed by a catalytically active crystal form. J Mol Biol 2000;297: 819–828.
45. Moodie SL, Mitchell JBO, Thornton JM. Protein recognition of adenylate: an example of a fuzzy recognition template. J Mol Biol 1996;263:486–500.
46. Nobeli I, Laskowski RA, Valdar WSJ, Thornton JM. On the molecular discrimination between adenine and guanine by proteins. Nucleic Acids Res 2001;29:4294–4309.
47. Hellinga HW, Caradonna JP, Richards FM. Construction of new ligand binding sites in proteins of known structure. II. Grafting of a buried transition metal binding site into *Escherichia coli* thioredoxin. J Mol Biol 1991;222:787–803.
48. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242. http://www.rcsb.org
49. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1995;247:536–540. http://scop.mrc-lmb.cam.ac.uk/scop/
50. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637. http://www.cmbi.kun.nl/gv/dssp
51. Lee B, Richards FM. The interpretation of protein structures: estimation of static accessibility. J Mol Biol 1971;55:379–400. http://wolf.bms.umist.ac.uk/naccess/
52. Thompson JD, Higgins DG, Gibson TJ. ClustalW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994;22:4673–4680. http://www.ebi.ac.uk/clustalw/
53. Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25: 3389–3402. http://www.ncbi.nlm.nih.gov/BLAST/
54. Guex N, Peitsch MC. SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. Electrophoresis 1997;18:2714–2723. http://www.expasy.org/spdbv
55. Sayle R. RASMOL molecular visualization program. Greenford, Middlesex, UK: Biomolecular Structure Group, Glaxo Research and Development; 1994. http://www.bernstein-plus-sons.com/software/rasmol/
56. Sankaranarayanan R, Sekar K, Banerjee R, Sharma V, Surolia A, Vijayan M. A novel mode of carbohydrate recognition in jacalin, a *Moraceae* plant lectin with a beta-prism fold. Nat Struct Biol 1996;3:596–603.