Multiplicity of carbohydrate-binding sites in β-prism fold lectins: occurrence and possible evolutionary implications

ALOK SHARMA, DIVYA CHANDRAN, DESH D SINGH and M VIJAYAN*

Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560 012, India

**Corresponding author (Fax, 91-80-23600683; Email, mv@mbu.iisc.ernet.in)*

The β-prism II fold lectins of known structure, all from monocots, invariably have three carbohydrate-binding sites in each subunit/domain. Until recently, β-prism I fold lectins of known structure were all from dicots and they exhibited one carbohydrate-binding site per subunit/domain. However, the recently determined structure of the β-prism fold I lectin from banana, a monocot, has two very similar carbohydrate-binding sites. This prompted a detailed analysis of all the sequences appropriate for two-lectin folds and which carry one or more relevant carbohydrate-binding motifs. The very recent observation of a β-prism I fold lectin, griffithsin, with three binding sites in each domain further confirmed the need for such an analysis. The analysis demonstrates substantial diversity in the number of binding sites unrelated to the taxonomical position of the plant source. However, the number of binding sites and the symmetry within the sequence exhibit reasonable correlation. The distribution of the two families of β-prism fold lectins among plants and the number of binding sites in them, appear to suggest that both of them arose through successive gene duplication, fusion and divergent evolution of the same primitive carbohydrate-binding motif involving a Greek key. Analysis with sequences in individual Greek keys as independent units lends further support to this conclusion. It would seem that the preponderance of three carbohydrate-binding sites per domain in monocot lectins, particularly those with the β-prism II fold, is related to the role of plant lectins in defence.

[Sharma A, Chandran D, Singh D D and Vijayan M 2007 Multiplicity of carbohydrate-binding sites in β-prism fold lectins: occurrence and possible evolutionary implications; *J.Biosci.* **32** 1089–1110]

1. Introduction

Lectins, multivalent carbohydrate-binding proteins of non-immune origin, have the unique ability to decode the information contained in complex carbohydrate structures of glycoproteins and glycolipids by stereo-specifically recognizing and binding to carbohydrates and carbohydrate linkages. Lectins are present in all kingdoms of life. They are involved in various biological processes such as cell–cell communication, host–pathogen interaction, cancer metastasis, embryogenesis, tissue development and mitogenic stimulation (Lis *et al* 1998; Drickamer 1999; Vijayan and Chandra 1999; Loris *et al* 2002). Because of the complex nature and numerous possibilities of glycosidic linkages and stereoisomers, carbohydrates have always been a challenge to structural biologists. The advancement in high-resolution techniques such as X-ray crystallography and nuclear magnetic resonance (NMR), as well as a wealth of biochemical data indicating the importance of carbohydrates in *in vivo* systems have resulted in increased attention being paid to carbohydrates. Thus, the study of protein–carbohydrate interactions and evolution of proteins with stringent affinity towards specific isomers from a pool of equivalent possibilities is of prime importance. Lectins appear to be the ideal candidates for such studies. The biological roles of animal, bacterial and viral lectins are reasonably well understood. However, although thoroughly studied structurally and biochemically, the endogenous roles of plant lectins are yet to be fully elucidated. It is believed that they are involved in root–nodule symbiosis in legume plants and also in plant defence (Chrispeels and Raikhel 1991; Peumans and Van Damme 1995; Hirsch 1999;

Keywords. β-prism fold; carbohydrate-binding; evolution; gene duplication; multiple ligand sites

Navarro-Gochicoa *et al* 2003; Imberty *et al* 2004). The stereo-specific selectivity of plant lectins has been exploited in a wide variety of applications, such as purification of glycoproteins, markers for cancer cells, antimicrobial agents and drug delivery (Lehr and Gabor 2004). Studies on plant lectins have also contributed substantially to the understanding of the structure and assembly of proteins and strategies for generating ligand specificity (Vijayan and Chandra 1999; Delbaere *et al* 1993; Banerjee *et al* 1994; Rini 1995; Elgavish and Shaanan 1998; Jeyaprakash *et al* 2004; Jeyaprakash *et al* 2005).

Based on the structure of their subunit folds, plant lectins themselves have been classified into five groups (http://www.cermav.cnrs.fr/lectines): legume lectins, hevein domain lectins, β-prism I fold lectins (also referred to as jacalin-like lectins), β-prism II fold lectins (also referred to as monocot mannose-binding lectins) and β-trefoil fold lectins. Of these, the last three exhibit threefold symmetry. In particular, the β-prism I and the β-prism II folds have prismoidal arrangements involving a four-stranded β-sheet constituting each side of the prism. The strands are roughly parallel to the threefold axis in the β-prism I fold while they are nearly perpendicular to the axis in the β-prism II fold.

The β-prism I fold was first characterized as a lectin fold in this laboratory through the X-ray analysis of jacalin, one of the two lectins from jackfruit seeds (Sankaranarayanan *et al* 1996). The other lectin from the seeds, artocarpin, also has a β-prism I fold (Pratap *et al* 2002). A jacalin subunit contains two polypeptide chains resulting from post-translational proteolysis. The amino terminus generated by the proteolysis has been shown to be important for the lectin's specificity for galactose at the primary binding site. Artocarpin is a single polypeptide chain and is specific for mannose at the primary binding site. Subsequently, the structural basis of the carbohydrate specificity in the lectin has been thoroughly characterized. Although both the lectins have threefold symmetrical subunits, each subunit binds only one sugar. Also, the symmetry in the three-dimensional structure is not reflected in the sequence. In the meantime, the crystal structures of several other β-prism I fold plant lectins became available (Lee *et al* 1988; Bourne *et al* 1999; Bourne *et al* 2004; Rao *et al* 2004; Gallego *et al* 2005; Rabijns *et al* 2005; Yen-Chieh *et al* 2006). Their subunits share the basic structural and carbohydrate-binding characteristics of jacalin and artocarpin. However, they exhibit a wide variety of quaternary structures. Originally, the β-prism I fold was considered to be characteristic of the *Moraceae* family. However, the fold has been found in lectins from other plant families as well. The widespread occurrence of this fold in lectins from different families has also been confirmed by a detailed sequence analysis (Raval *et al* 2004).

The β-prism II fold was first discovered in snowdrop lectin (Hester *et al* 1996). Snowdrop lectin is tetrameric while the second lectin of the same class to be X-ray analysed, garlic lectin, is dimeric (Chandra *et al* 1999). Since then, the structures of a few other lectins with β-prism II fold have been reported (Chantalat *et al* 1996; Sauerborn *et al* 1999; Wood *et al* 1999). All of them are mannose-specific. Unlike in the case of β-prism I fold lectins, the threefold symmetry of the β-prism II fold lectins is reflected in the sequence as well (Ramachandraiah and Chandra 2000). Further, each subunit contains three carbohydrate-binding sites.

Some features of the recently determined crystal structure of banana lectin went against the conventional wisdom on β-prism I fold lectins in certain respects (Singh *et al* 2005; Meagher *et al* 2005). The threefold symmetry of the subunit structure is reflected, albeit weakly, in the sequence as well. Furthermore, each subunit contains two carbohydrate-binding sites of identical structure situated at two of the three threefold-equivalent positions. It is also interesting that banana is a monocot while all other β-prism I fold plant lectins of known structure are from dicots. When reporting the structure of jacalin, we had hypothesized that the β-prism I fold could have arisen from successive gene duplication and fusion of a primitive carbohydrate-binding motif involving a polypeptide chain containing approximately 40 amino acid residues. The new features observed in banana lectin appear to support this hypothesis. In banana lectin, three components resulting from successive gene duplication and fusion have not diverged enough to obliterate past history, while the components have done so in other β-prism I fold lectins of known structure, all from dicots. This observation led to an analysis of the structure and sequence of β-prism I fold lectins with special reference to the evolution of carbohydrate-binding sites. After the completion of one stage of this analysis, the structure of an algal lectin, griffithsin, containing β-prism I fold domains which bear three carbohydrate-binding sites each, has been reported (Chandra 2006, Ziolkowska *et al* 2006). This adds to the relevance of the analysis. In parallel, a similar analysis was carried out on β-prism II fold lectins also. These analyses, presented here, provide interesting insights into the evolutionary history and the possible common ancestor of the two types of β-prism fold lectins. They also point to a plausible rationale for the presence of a higher number of binding sites per domain in these lectins from monocots, than in those from dicots, in terms of the role of plant lectins in defence.

2. Materials and methods

Sequence homologues of the banana and garlic lectins (accession number AAM48480.1 for banana lectin and 4389040 for garlic lectin) were searched by PSI-BLAST alignment with an e-value cut off of 0.0005 using the NR database available at NCBI (Altschul *et al* 1997; Schaffer *et al* 2001). Alignments with an overlap length of less than

75% were not considered for further study, as they cannot form the complete fold. The search was first carried out on 9 April 2006 when the database size was 3,632,049. A search was again made in December 2006 and sequences deposited after April 2006 were considered for further analysis. Sequences obtained thus were made non-redundant using a Perl script (Li *et al* 2001; Li *et al* 2002). Smaller sequences with more than 90% identity were removed in all versus all pair-wise alignment. Lectin domains in each sequence were searched using the CDD tool available at the NCBI. Domain search was relaxed with an e-value cut-off of 10 and lower stringency cut-offs (Marchel Bauer *et al* 2002).

In both the cases, sequences with at least one carbohydrate-binding site motif were sorted after analysing the pair-wise alignment of all sequences with the corresponding target lectin sequence and profile search using 3of5. Those sequences in which the carbohydrate-binding motif (G… GXXXD or QXDXNXVXY) also aligned were then selected for further analysis. This selection was further cross-checked by the profile search tool in the 3of5 module available on the Expasy server. GX[GAVIYWF] [GAVIYWF][DNEQ] and [QE]X[DENQ][X][DENQ][AVILG]X[YF] were used as search profiles for *β*-prism I and *β*-prism II fold lectins, respectively. All the selected sequences were indicated to have lectin domains. Models were built for this set of sorted sequences using various structure prediction tools for further selection on the basis of the ability of the sequence to fold into a reasonably complete *β*-prism (Rost 1996; Bates and Sternberg 1999; Bates *et al* 2001; Contreras-Moreira and Bates 2002). The sequences for which models could not be predicted or the model did not yield either of the *β*-prism folds lay in the twilight zone of similarity (~15–30%). All the pair-wise and multiple alignments were carried out using Align and CLUSTALW, respectively, both available at www.ebi.ac.uk (Rice *et al* 2000; Thompson *et al* 1994). Corresponding binding-site motifs in both the cases were searched using 3of5 available at http://www.dkfz.de/mga2/3of5/ (Seiler *et al* 2006).

Dot plot analysis was carried out using the DOTMATCHER program available at the EMBOSS server with a window size of 30 and threshold cut-off of 10 (Sonnhammer Erik and Durbin 1995).

Phylogenetic analyses were carried out using the Bayesian method as implemented in MrBayes 3.1 (Huelsenbeck and Ronquist 2001) and maximum parsimony as implemented in the MEGA suite of programs (Kumar *et al* 2004*)*. Both the methods gave similar connectivity. In all the illustrations the Mrbayes output has been used. Protein coordinates were obtained from the Protein Data Bank (PDB) (Berman *et al* 2000). *In silico* mutations for structural studies were carried out using Coot 0.0 (Emsley and Cowtan 2004). Pymol was used for the analysis and for illustrating 3-dimensional structures (http://www.pymol.org).

3. Results and discussion

3.1 *Occurrence of β-prism fold lectins*

A subunit of banana lectin (figure 1a) was chosen as the search model for proteins with the *β*-prism I fold. A PSI-BLAST search, first made in April 2006, using the sequence of this lectin through the entire non-redundant database using cut-off values and criteria as mentioned in the section on Materials and methods, led to the identification of 194

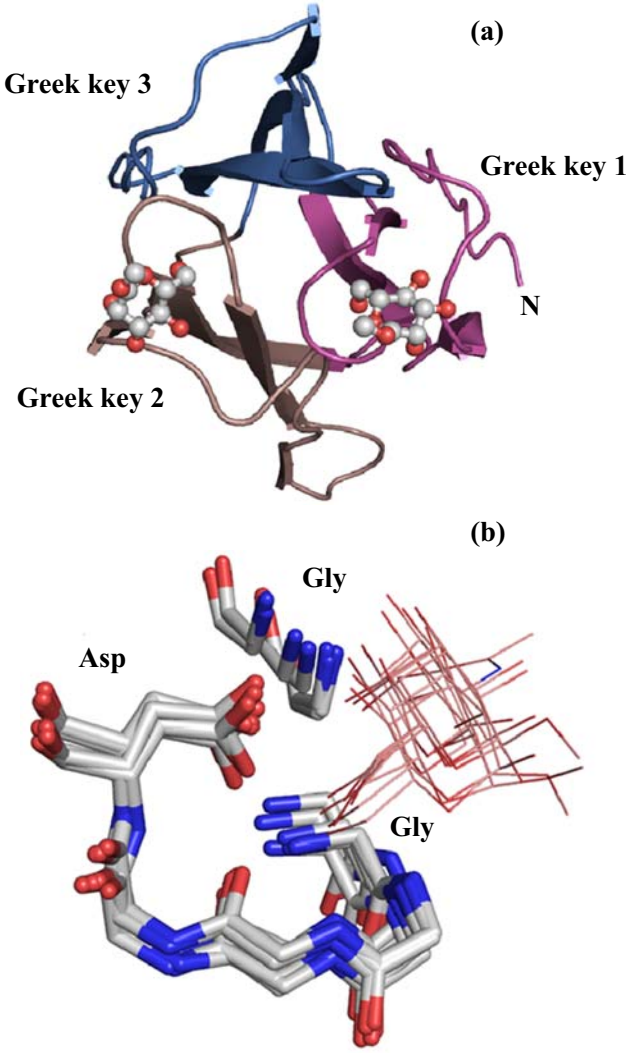


Figure 1. **(a)** Subunit structure of banana lectin (PDB CODE 1X1V) viewed down the pseudo threefold axis. The three Greek keys are shown in different colours. Sugars are represented as ball and stick. **(b)** A structural superposition of the carbohydrate-binding sites of all the *β*-prism I fold lectins with known structure in complex with sugar. For the sake of clarity, side chains of residues XXX of the G…GXXXD motif are not shown. Sugars are shown in the line representation.

Table 1. List of finally selected banana lectin homologues identified from the sequence database, using banana lectin as a search template

Sl. No.	Accession number	Source	I	II	III	IV	V
				Plants			
Algae							
1	P84801	*Griffithsia* sp.	1	121	44	3	Griffithsin (lectin)
Gymnosperm							
2	BAE95375.1	*Cycas revolute*	2	291	(a) 52	1	Lectin
3					(b) 48	2	
Monocots other than *Oryza sativa*							
4	AAR20919.1	*Triticum aestivum*	1	304	56	2	Jasmonate-induced protein
5	AAA87042.1	*Hordeum vulgare* subsp. *vulgare*	1	304	54	2	Jasmonate-induced protein
6**	AAM46813.1	*Triticum aestivum*	1	345	45	2	Hessian fly response gene 1 protein (a lectin-like wheat gene which responds to Hessian fly)
7	AAV39531.1	*Hordeum vulgare*	1	146	56	2	Horcolin
8	AAM48480.1	*Musa acuminata*	1	141	100	2	Lectin
9	AAY41607.1	*Agrostis stolonifera*	1	319	50	1	Crs-1 (meiosis-specific cyclin, i.e. meiotically upregulated protein)
10	AAF71261.2	*Zea mays*	1	306	52	1	Beta-glucosidase aggregating factor precursor
11	AAS20963.1	*Hyacinthus orientalis*	1	161	47	1	OSJNBa0016N04.20-like protein
12	AAC49284.1	*Triticum aestivum*	1	343	45	1	Unknown
13	AAQ07258.1	*Ananas comosus*	1	145	59	1	Jacalin-like lectin
14	AAP87359.1	*Hordeum vulgare*	1	160	50	1	High light protein
15**	ABI24164.1	*Sorghum bicolor*	1	305	52	2	Beta-glucosidase aggregating factor
Oryza sativa							
16	NP_918855.1	*O. sativa*	1	144	51	2	Putative mannose-binding rice lectin
17	AAU90197.1	*O. sativa*	1	152	54	2	Unknown protein
18	XP_471663.1	*O. sativa*	1	150	52	2	OSJNBa0041M21.2
19	BAD67983.1	*O. sativa*	1	209	51	2	Putative GOS9 (rice-specific gene)
20	XP_465120.1	*O. sativa*	1	1072	54	2	Putative LZ-NBS-LRR class RGA (stripe rust resistance protein)
21	ABA96667.1	*O. sativa*	1	307	55	2	Jasmonate-induced protein
22**	ABA97248.1	*O. sativa*	1	306	55	2	Expressed protein
23	BAD67976.1	*O. sativa*	1	161	52	2	GOS9 (root-specific rice gene)
24	ABA93998.1	*O. sativa*	3	1384	(a) 55	1	Stripe rust resistance protein Yr10
25					(b) 44	1	
26**					(c) 49	1	
27	ABA94721.1	*O. sativa*	2	734	(a) 53	1	Jacalin-like lectin domain containing protein
28					(b) 53	1	
29	XP_472139.1	*O. sativa*	1	477	54	1	OSJNBa0016N04.16
30	NP_916350.1	*O. sativa*	3	925	(a) 48	1	P0413G02.3
31					(b) 56	1	

Sl. No.	Accession number	Source	I	II	III	IV	V
32					(c) 46	1	
33	ABB46687.1	*O. sativa*	1	154	49	1	Jacalin-like lectin domain containing protein
34	BAD52750.1	*O. sativa*	1	271	51	1	Putative salT
35	XP_474804.1	*O. sativa*	1	1269	48	1	OSJNBa0014F04.15
36	ABA96835.1	*O. sativa*	1	260	55	1	Jacalin homologue/Jjasmonate-induced protein
37	NP_908901.1	*O. sativa*	1	145	50	2	Mannose-binding rice lectin
38	BAD37295.1	*O. sativa*	1	146	54	1	Putative salT protein precursor
39**	AAP12924.1	*O. sativa*	1	191	44	2	Putative salt-induced protein
40	XP_475665.1	*O. sativa*	2	604	(a) 53	1	Unknown protein
41					(b) 48	1	
42	ABA94002.1	*O. sativa*	2	1386	(a) 55	1	NBS-LRR resistance protein/Jacalin-like lectin
43**					(c) 51	1	
44	ABA94728.1	*O. sativa*	3	837	(a) 50	1	Jacalin-like lectin domain containing protein
45**					(b) 53	1	
46	NP_001042976.1	*O. sativa*	1	145	50	1	Japonica-cultivar group
47	NP_001044410.1	*O. sativa*	1	349	55	2	Japonica-cultivar group
48	NP_001046624.1	*O. sativa*	1	141	53	2	Japonica-cultivar group
49	NP_001050311.1	*O. sativa*	1	343	44	2	Japonica-cultivar group
50	NP_001052399.1	*O. sativa*	1	150	52	2	Japonica-cultivar group
51	NP_001052560.1	*O. sativa*	3	770	51	1	Japonica-cultivar group
52					54	1	Japonica-cultivar group
53					53	2	Japonica-cultivar group
54	NP_001054618.1	*O. sativa*	1	152	54	2	Japonica-cultivar group
Dicots other than *Arabidopsis thaliana*							
55	1XXR	*Morus nigra*	1	161	50	1	Mannose-specific jacalin-related lectin
56	AAD11577.1	*Helianthus tuberosus*	1	151	50	1	Lectin HE17
57	AAL09163.1	*Morus nigra*	1	216	46	1	Galactose-binding lectin
58	AAA32678.1	*Artocarpus heterophyllus*	1	217	47	1	Jacalin
59	P83304	*Parkia platycephala*	3	447	(a) 50	1	Mannose/glucose-specific lectin
60					(b) 53	1	
61					(c) 50	1	
62	1J4S	*Artocarpus heterophyllus*	1	149	46	1	Artocarpin: mannose-specific lectin

Sl. No.	Accession number	Source	I	II	III	IV	V
63	ABC70328.1	*Castanea crenata*	1	310	46	1	Agglutinin isoform
64	S15825	*Maclura pomifera*	1	133	46	1	Agglutinin alpha chain
65	1TP8	*Artocarpus hirsuta*	1	133	47	1	*Artocarpus hirsuta*: galactose-specific lectin
66	AAB23126.1	*Artocarpus heterophyllus*	1	133	48	1	Jacalin
67**	AAC08051.1	*Brassica napus*	1	552	44	1	Myrosinase-binding protein
68	AAC08050.1	*Brassica napus*	1	331	44	1	Myrosinase-binding protein
69	BAB18761.1	*Helianthus tuberosus*	1	143	59	1	Lectin
70	AAG10403.1	*Convolvulus arvensis*	1	152	48	1	Mannose-binding lectin
71	BAA14024.1	*Ipomoea batatas*	1	154	44	1	Ipomoelin
72	AAC49564.1	*Calystegia sepium*	1	153	50	1	Lectin
73	AAB22274.1	*Artocarpus heterophyllus*	1	133	48	1	Jacalin heavy chain
74	CAJ38387.1	*Plantago major*	1	197	53	1	Jacalin-like domain protein
Arabidopsis thaliana							
75	NP_177447.1	*A. thaliana*	1	176	52	1	Unknown protein
76	NP_849691.1	*A. thaliana*	2	595	(a) 50	1	Unknown protein
77					(b) 48	1	
78	NP_974324.2	*A. thaliana*	2	705	(a) 46	1	Unknown protein
79**					(b) 44	1	
80	NP_175618.1	*A. thaliana*	2	293	(a) 43	1	Unknown protein
81**					(b) 43	1	
82	AAD12681.1	*A. thaliana*	2	303	(a) 44	1	Putative myrosinase-binding protein
83**					(b) 42	1	
84	AAD12684.1	*A. thaliana*	1	445	44	1	Putative myrosinase-binding protein
85	NP_198444.1	*A. thaliana*	3	444	42	1	Unknown protein
				Animals			
86	XP_510910.1	*Pan troglodytes*	1	1242	49	1	PREDICTED: similar to kinesin-like protein
87	Q8CJD3	*Rattus norvegicus*	1	167	47	1	Zymogen granule membrane protein
88	XP_536909.1	*Canis familiaris*	1	167	50	1	PREDICTED: similar to zymogen granule protein
89	XP_871351.1	*Bos taurus*	1	167	49	1	PREDICTED: similar to zymogen granule protein
				Fungi			
90	CAG90055.1	*Debaryomyces hansenii*	1	735	28	1	Unnamed protein product
91	XP_506051.1	*Yarrowia lipolytica*	1	702	45	1	Hypothetical protein
92	NP_012158.1	*Saccharomyces cerevisiae*	1	696	32	1	Putative metalloprotease
93	XP_445234.1	*Candida glabrata*	1	683	34	1	Unnamed protein product
94	CAB63793.1	*Schizosaccharomyces pombe*	2	612	47	2	SPAC607.06c

Sl. No.	Accession number	Source	I	II	III	IV	V
95	BAE57820.1	*Aspergillus oryzae*	1	785	40	2	Unknown protein
96	EAT80432.1	*Phaeosphaeria nodorum*	1	788	41	1	Hypothetical protein
			Monera				
97	ZP_00591571.1	*Prosthecochloris aestuarii DSM 271*	1	171	46	3	Jacalin-related lectin
98	ZP_00532662.1	*Chlorobium phaeobacteroides BS1*	1	171	45	1	Zymogen granule protein

In the sequences marked with **, either GXXXE or GXXXN has been considered as a possible carbohydrate-binding motif.

I: Number of Jacalin-related lectin domains with carbohydrate-binding motif(s).
II: Total length of the polypeptide.
III: Similarity (%) of each domain with banana lectin (AAM48480.1).
IV: Number of carbohydrate-binding motif(s) in each domain.
V: Predicted or known function of the protein.

non-redundant sequences. These sequences exhibited a similarity in the range of 28–64% (identity 16–42%) with that of banana lectin. They were then searched for carbohydrate-binding motifs. A superposition of the binding sites in *β*-prism I fold lectins is shown in figure 1b. The binding site involves the motif G…GXXXD. Although not contiguous in sequence with the rest of the motif, it is important to take into account the distal glycine as well. Not only does it occur in all relevant structures, but it also occurs at a position in conformational space, which can be occupied only by glycine. The φ and ψ values for the residue in the relevant lectins of known structure vary between 51 and 88º and –154 and 163º (–197º), respectively. Furthermore, in the three-dimensional structure, the distal glycine comes close to the rest of the carbohydrate-binding site. If each of the relevant sequences is circularly arranged such that the N- and C-termini are in close proximity, the separation between this glycine and aspartic acid is around 20 residues in all cases. The second glycyl residue in the motif also has φ, ψ values appropriate only for a glycyl residue. Thus, the two glycines appear to be important for maintaining the desired geometry of the binding site. The aspartate side chain is crucial for lectin–carbohydrate interactions. In a few instances (ten), motifs G…GXXXE and G…GXXXN were also accepted as carbohydrate-binding motifs. It was verified through modelling that the presence of E or N instead of D is consistent with the observed lectin–sugar interactions.

Of the 194 sequences considered, 36 and 51 were from *Oryza sativa* and *Arabidopsis thaliana*, respectively. The availability of their entire genomes probably accounts for these large numbers. Of these, 13 sequences in *O. sativa* and 44 in *A. thaliana* did not contain any carbohydrate-binding motif. These sequences were omitted from further consideration. Sequences from other sources which do not contain carbohydrate-binding motifs were also omitted from further consideration. Sequences that failed to fold into a *β*-prism I fold on model building were also not considered further. There were five such sequences which exhibited low sequence similarity. The remaining domains/subunits, which may be considered as homologues of banana lectin in structure and function, are listed category-wise in table 1. The second search, made in December 2006, following the same protocol, added 10 more sequences, which are also given in the table. In view of the large number of sequences from *O. sativa* and *A. thaliana*, they have been separately grouped in the table.

A similar search, first made in April 2006, for *β*-prism II fold using a garlic lectin subunit (figure 2a) as the search model, resulted in the identification of 452 *β*-prism II fold sequences. Of these, 123 are from *O. sativa*, 77 from *A. thaliana* and 106 from *Brassica* spp, all organisms with sequenced genomes. The motif QXDXNXVXY (figure 2b) was used to search for the carbohydrate-binding sites. In a few instances, motifs with one or two conservative changes were also accepted as those involved in carbohydrate binding. In each such instance, all the rotamers of the changed side chain, available in the Coot 0.0 rotamer library, were examined in the garlic lectin structure and it was ensured that there was no unacceptable steric contact. Only such changes were accepted in which the lectin–sugar hydrogen bonds were substantially maintained. In particular, it was ensured that all interactions involving O2, which are crucial for mannose recognition, were present even when the residue was changed.

It turns out that none of 77 and 106 sequences identified in *A. thaliana* and *Brassica* spp, respectively, contain any carbohydrate-binding motif. In the case of *O. sativa*, only 1 of the 116 sequences contains carbohydrate-binding motifs. Therefore, there was no need to treat the sequences from the whole genomes of these organisms separately. The single

sequence from *O. sativa* was grouped along with those from other monocots in table 2, which lists all the lectin domains with *β*-prism II fold containing one or more mannose-binding motifs. A second search made in December 2006 added 9 more sequences, which are also given in the table.

3.2 *Distribution of β-prism fold lectins*

A majority of *β*-prism fold lectins of both types occur in plants. They are also found in animals, fungi and bacteria. Among plants, *β*-prism I fold lectins occur in monocots as well as dicots. In dicots, each domain invariably carries only one carbohydrate-binding site. In monocots, domains with one and two carbohydrate-binding sites occur with almost equal frequency. In animals, *β*-prism I fold lectins with only one binding site have so far been identified. Domains with one or two binding sites are seen in fungi. The rare examples of a *β*-prism I fold lectin with three binding sites are seen in bacteria and algae.

In plants, *β*-prism II fold lectins occur overwhelmingly in monocots. In most cases, they carry three carbohydrate-binding sites each. At least one monocot *β*-prism II fold lectin has been identified with two carbohydrate-binding sites in it. There are a few which carry only one carbohydrate-binding site each. Three dicots containing *β*-prism II fold lectins have been identified. They carry one to three carbohydrate-binding sites. It is also interesting to note that most of the domains containing one carbohydrate-binding site in monocots form a part of sequences containing multiple domains. The only gymnosperm lectin with a *β*-prism II fold domain carries three carbohydrate-binding motifs. *β*-prism II fold lectins from non-plants carry one to three binding sites each. Most of the bacterial domains (28 out of 32) contain two or three carbohydrate-binding motifs. All protists have two carbohydrate-binding motifs. Fungal domains have one or two whereas animal domains have two or three carbohydrate-binding motifs. The *β*-prism II fold with three carbohydrate-binding sites predominantly appears to be a monocot phenomenon. Also, the sequence similarities and sources as listed in tables 1 and 2 indicate that *β*-prism II fold lectins are more widespread but less diverse in terms of carbohydrate-binding sites than *β*-prism I fold lectins.

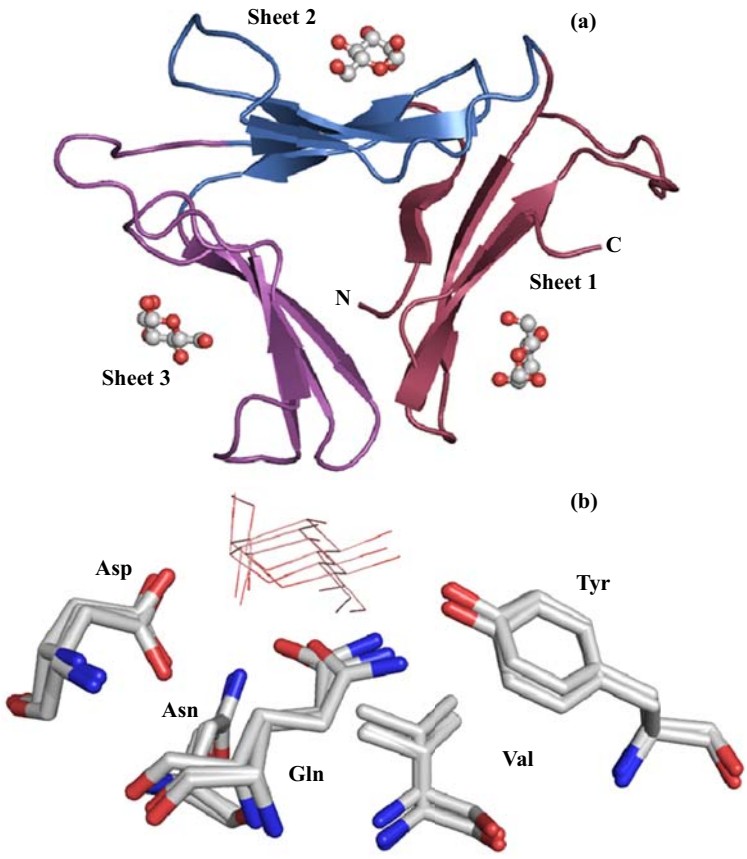


Figure 2. **(a)** Subunit structure of garlic lectin (PDB CODE 1BWU) viewed down the pseudo threefold axis. The three sheets are shown in different colours. Sugars are represented as ball and stick. **(b)** A structural superposition of the carbohydrate-binding sites of all the β-prism II fold lectins with known structure in complex with sugar. Sugars are shown in line representation.

Table 2. List of finally selected garlic lectin homologues identified from the sequence database, using garlic lectin as a search template

Sl. No.	Accession number	Source	I	II	III	IV	V
			Plants				
Gymnosperm							
1	AAT73201.1	*Taxus* x *media*	3	144	59	3	Mannose-binding lectin
Monocots							
2	AAL07478.1	*Galanthus nivalis*	1	157	64	3	Lectin
3	AAW22055.1	*Lycoris* sp.	1	162	63	3	Agglutinin
4	AAB64238.1	*Allium sativum*	1	181	85	3	Mannose-specific lectin
5	AAA33546.1	*Narcissus hybrid* cultivar	1	171	66	3	Mannose-specific lectin precursor
6	AAP37975.1	*Zephyranthes grandiflora*	1	163	67	3	Agglutinin
7	BAD98798.1	*Lycoris radiata* var. *pumila*	1	156	62	3	Lectin
8	AAM44412.1	*Zephyranthes candida*	1	169	65	3	Agglutinin
9	1NPL	*Narcissus pseudonarcissus*	1	109	64	3	Agglutinin
10	AAP57409.1	*Amaryllis vittata*	1	158	67	3	Agglutinin
11	AAP20877.1	*Lycoris radiata*	1	158	62	3	Lectin
12	AAM28277.1	*Ananas comosus*	1	164	67	3	Mannose-binding lectin
13	BAD67183.1	*Dioscorea polystachya*	1	149	65	3	Mannose-specific lectin
14	AAP04617.1	*Amorphophallus konjac*	1	158	59	3	3DAKA precursor
15	AAV70492.1	*Zingiber officinale*	1	169	60	3	Mannose-binding lectin precursor
16	AAB64239.1	*Allium sativum*	2	303	(a) 71	3	Lectin-related protein
17**					(b) 59	2	
18	AAR82848.1	*Crinum asiaticum*	1	175	61	3	Mannose binding lectin
19	AAV66418.1	*Dendrobium officinale*	1	165	61	3	Mannose-binding lectin precursor
20	AAG52664.2	*Gastrodia elata*	1	179	62	3	Antifungal protein precursor
21	AAQ55289.1	*Typhonium divaricatum*	1	197	63	3	Lectin precursor
22	AAQ18904.1	*Zephyranthes grandiflora*	1	191	64	3	Mannose-binding lectin
23	AAK59994.1	*Gastrodia elata*	1	169	59	3	Antifungal protein
24	AAA16281.1	*Allium ursinum*	1	185	86	3	Mannose-specific lectin
25	AAA32643.1	*Allium sativum*	1	155	92	3	Lectin
26	JE0136	*Galanthus nivalis*	1	160	66	3	Lectin precursor
27	AAA16280.1	*Allium ursinum*	1	176	86	3	Mannose-specific lectin
28	AAA19911.1	*Clivia miniata*	1	169	65	3	Lectin
29	AAA33347.1	*Galanthus nivalis*	1	154	63	3	Lectin
30	AAA19913.1	*Clivia miniata*	1	166	62	3	Lectin
31	AAC37360.1	*Allium ascalonicum*	1	177	85	3	Mannose-specific lectin
32	AAC49387.1	*Tulipa hybrid cultivar*	1	183	66	3	Mannose-binding lectin precursor
33	AAA19577.1	*Epipactis helleborine*	1	172	63	3	Lectin
34	AAA19578.1	*Cymbidium hybrid*	1	176	60	3	Lectin

Sl. No.	Accession number	Source	I	II	III	IV	V
35	AAA20899.1	*Listera ovata*	1	175	62	3	Lectin
36	AAC48927.1	*Epipactis helleborine*	1	168	59	3	Lectin
37	AAC37423.1	*Listera ovata*	1	167	63	3	Mannose-binding protein
38	1XD6	*Gastrodia elata*	1	112	61	3	Mannose-binding lectin
39	AAC37422.1	*Listera ovata*	1	176	63	3	Lectin
40**	AAA33345.1	*Galanthus nivalis*	1	161	65	3	Lectin
41**	AAC49858.1	*Allium ursinum*	1	166	82	3	Mannose-specific lectin precursor
42**	AAW82332.1	*Polygonatum roseum*	1	159	61	3	Mannose/sialic acid binding lectin
43**	AAQ75079.1	*Zantedeschia aethiopica*	1	138	65	3	Mannose binding lectin
44**	AAM77364.1	*Polygonatum cyrtonema*	1	160	60	3	Mannose/sialic acid-binding lectin
45**	AAA32646.1	*Allium sativum*	1	313	91	3	Lectin
46**	AAC49413.1	*Polygonatum multiflorum*	1	160	60	3	Mannose-specific lectin precursor
47**	P49329	*Aloe arborescens*	1	109	65	3	Mannose-specific lectin precursor
48	AAD16403.1	*Hyacinthoides hispanica*	1	155	65	2	Lectin SCA man precursor
49	AAP20876.1	*Pinellia ternata*	2	269	(a) 51	1	Lectin
50					(b) 54	1	
51	CAA53717.1	*Colocasia esculenta*	1	253	51	1	Tarin (storage protein)
52	ABC69036.1	*Alocasia macrorrhizos*	2	270	(a) 49	1	Mannose-binding lectin
53					(b) 52	1	
54	BAA03722.1	*Colocasia esculenta*	2	268	(a) 53	1	Storage protein
55					(b) 51	1	
56	AAP50524.1	*Arisaema heterophyllum*	2	258	(a) 55	1	Agglutinin
57					(b) 49	1	
58	AAS66304.1	*Arisaema lobatum*	2	258	(a) 51	1	Mannose-binding lectin
59					(b) 50	1	
60	AAC48998.1	*Arum maculatum*	2	260	(a) 46	1	Lectin precursor
61					(b) 55	1	
62	AAC49384.1	Tulipa hybrid *cultivar*	1	275	51	1	Complex specificity lectin precursor
63	ABA00714.1	*Allium triquetrum*	1	173	81	3	Agglutinin
64	BAD98797.1	*Lycoris radiata*	1	156	62	3	Lectin
65	NP_910000.1	*Oryza sativa*	1	797	26	1	Putative protein kinase
		Dicots					
67	AAD45250.1	*Hernandia moerenhoutiana* subsp. *samoensis*	1	133	64	3	Seed lectin
68**	AAZ30387.1	*Helianthus tuberosus*	1	118	51	2	Mannose-binding lectin
69**	ABE91586.1	*Medicago truncatula*	1	825	42	1	Protein kinase; curculin-like (Mannose-binding) lectin
		Animals					
70	CAI91574.1	*Lubomirskia baicalensis*	1	120	50	3	Mannose-binding lectin

Sl. No.	Accession number	Source	I	II	III	IV	V
71**	BAD90686.1	*Lophiomus setigerus*	1	111	52	2	Skin mucus lectin
72	BAE79275.1	*Leiognathus nuchalis*	1	113	49	2	Lily-type lectin
73	AAU14874.1	*Oncorhynchus mykiss*	1	111	50	2	Lectin
74	CAG10253.1	*Tetraodon nigroviridis*	1	116	50	2	Unnamed protein product
75	NP_001027736.1	*Takifugu rubripes*	1	116	48	2	Skin mucus lectin
		Fungi					
76	BAE55557.1	*Aspergillus oryzae*	1	114	50	2	Unnamed protein product
77**	XP_383865.1	*Gibberella zeae*	1	183	48	2	Hypothetical protein product
78	EAS27517.1	*Coccidioides immitis*	1	114	45	1	Hypothetical protein product
79	BAE63462.1	*Aspergillus oryzae*	1	129	44	1	Unnamed protein product
80	BAE63461.1	*Aspergillus oryzae*	1	138	43	1	Unnamed protein product
		Protista					
81	XP_636121.1	*Dictyostelium discoideum*	1	185	46	2	Comitin (membrane-associated protein)
82	XP_641612.1	*Dictyostelium discoideum*	1	135	47	2	Hypothetical protein
83**	EAR96445.1	*Tetrahymena thermophila*	2	413	(a) 46	2	Conserved hypothetical protein
84**					(b) 46	2	
85**	EAR80561.1	*Tetrahymena thermophila*	2	295	(a) 46	2	Conserved hypothetical protein
86**					(b) 46	2	
		Monera					
87	ZP_00462266.1	*Burkholderia cenocepacia*	2	298	(a) 57	3	Curculin-like lectin
88					(b) 56	3	
89	ZP_00413163.1	*Arthrobacter* sp.	2	226	(a) 49	3	Curculin-like lectin
90					(b) 45	3	
91	ZP_00687583.1	*Burkholderia ambifaria*	2	788	(a) 46	3	Peptidase, subtilisin Kexin, sedolisin: curculin like lectin
92					(b) 57	3	
93	YP_258360.1	*Pseudomonas fluorescens*	1	316	49	2	Putidacin L1 (Plant lectin-like bacteriocin)
94**	AAX31574.1	*Streptomyces filamentosus*	1	338	51	2	Unknown
95	YP_586686.1	*Ralstonia metallidurans*	2	852	(a) 53	2	Curculin-like lectin
96					(b) 51	1	
97	ZP_00520232.1	*Solibacter usitatus*	1	228	39	3	Curculin-like lectin
98	AAL73547.1	*Ruminococcus albus*	1	339	48	2	Bacteriocin (an antibacterial substance)
99	AAM95702.1	*Pseudomonas* sp.	2	276	(a) 46	2	Putidacin (plant lectin-like bacteriocin)
100**					(b) 38	2	
101	ABB23888.1	*Pelodictyon luteolum*	1	388	42	2	Hypothetical protein
102	AAM35756.1	*Xanthomonas axonopodis*	2	269	(a) 38	1	Hypothetical protein
103					(b) 45	1	

Sl. No.	Accession number	Source	I	II	III	IV	V
104	NP_440485.1	*Synechocystis* sp.	1	3972	43	1	Integrin alpha-subunit domain-like protein
105	ABK70862.1	*Mycobacterium smegmatis*	1	208	51	3	Mannose-binding lectin
106	YP_620971.1	*Burkholderia cenocepacia*	3	298	55	3	Curculin-like lectin
107					56	3	
108					40	3	
109	YP_620972.1	*Burkholderia cenocepacia*	3	270	54	3	Curculin-like lectin
110					48	2	
111					48	2	
112**	YP_772659.1	*Burkholderia cenocepacia*	3	788	46	3	Curculin-like lectin
113					58	3	
114					40	3	
115	YP_827995.1	*Solibacter usitatus*	1	228	39	3	Curculin domain protein
116	YP_829274.1	*Arthrobacter* sp.	2	226	49	3	Curculin-like protein
117					45	3	
118**	ZP_01463094.1	*Stigmatella aurantiaca*	1	513	51	3	Aqualysin-1

In the sequences marked with **, at least one ambiguous motif (other than QXDXNXVXY) has been considered as a possible carbohydrate-binding motif.

I: Number of bulb lectin domains with carbohydrate-binding motif(s).
II: Total length of the polypeptide.
III: Similarity (%) of each domain with garlic lectin (4389040).
IV: Number of carbohydrate-binding motif(s) in each domain.
V: Predicted or known function of the protein.

3.3 *Interrelationship among the three faces of the prism*

On account of the approximate internal threefold symmetry, each subunit of a β-prism I fold lectin has an appropriate loop in each one of the three Greek keys, irrespective of whether the loop carries a carbohydrate-binding motif or not. This is illustrated in figure 3a, b, c, in which the three Greek keys are superposed in artocarpin, banana lectin and griffithsin, three lectins containing one, two and three carbohydrate-binding sites, respectively, on a subunit. A similar representation has been shown for garlic lectin also (figure 3d). In artocarpin, only loop 1 (the loop in Greek key 1) binds sugar. Loop 2 (on Greek key 2) has nearly the same geometry as loop 1, but it does not contain the motif and hence does not bind sugar. The longer loop 3 (on Greek key 3) has a different geometry; it also does not carry the carbohydrate-binding motif. In banana lectin, loops 1 and 2 contain the motif and bind sugar. The longer loop 3 again has a different geometry. In both the cases, this loop functions as the secondary-binding site when oligosaccharides bind to the lectin. The same is true in the case of heltuba, a β-prism I fold lectin with the known crystal structure of an oligosaccharide complex. Griffithsin has three loops of similar structure, each carrying a carbohydrate-binding motif resulting in three binding sites on each subunit. Thus, the ability of each loop to bind sugar is determined by the structure (geometry of the loop) as well as the presence or absence of the sequence motif.

The similarity among the three loops appears to be a reflection of that among the Greek keys that carry them. For example, the percentage similarity (identity) between keys 1 and 2, keys 2 and 3, and keys 1 and 3 in artocarpin are low at 14.7 (10.7), 25.8 (16.7) and 34.5 (20.7), respectively. The corresponding values in banana lectin are higher at 38.2 (23.6), 42.0 (22.0) and 35.3 (23.5), respectively. The values, on an average, are still higher in griffithsin at 37.8 (31.1), 46.8 (27.7) and 42.3 (26.9), respectively. The extent of relatedness among the three lectins becomes even more striking when the dot plots of their sequences, with a window size of 30 and a stringency cut-off of 10, are examined (figure 4).

It may be mentioned that in terms of sequence and structure, the integrity of the Greek keys is maintained even when differences occur in multimerization. Also, the level of multimerization is in no way correlated with the number of carbohydrate-binding sites in each subunit. β-prism fold lectins, as indeed other type of lectins (Prabu *et al* 1999), exhibit a variety of quaternary structures. However, in a majority of β-prism I fold lectins of known structure, the

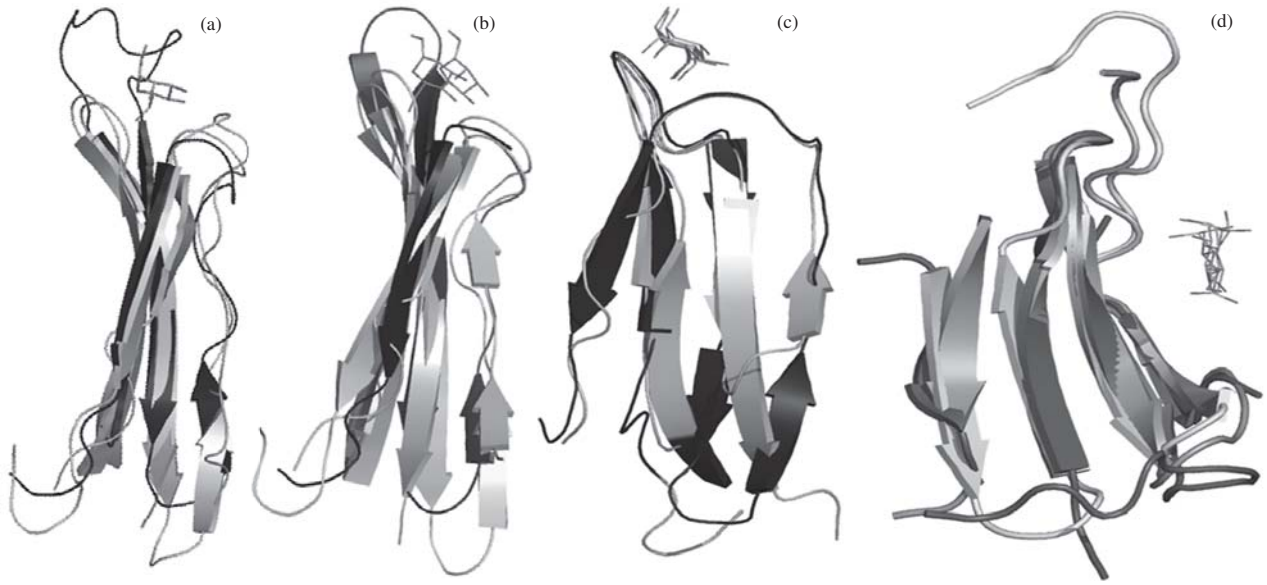


Figure 3. Structural superposition of individual sheets in **(a)** artocarpin, **(b)** banana lectin, **(c)** griffithsin, **(d)** garlic lectin. For **(a)**, **(b)** and **(c)** the longer loop from Greek key 3 is shown in the darker shade. Sugars are shown in line representation.

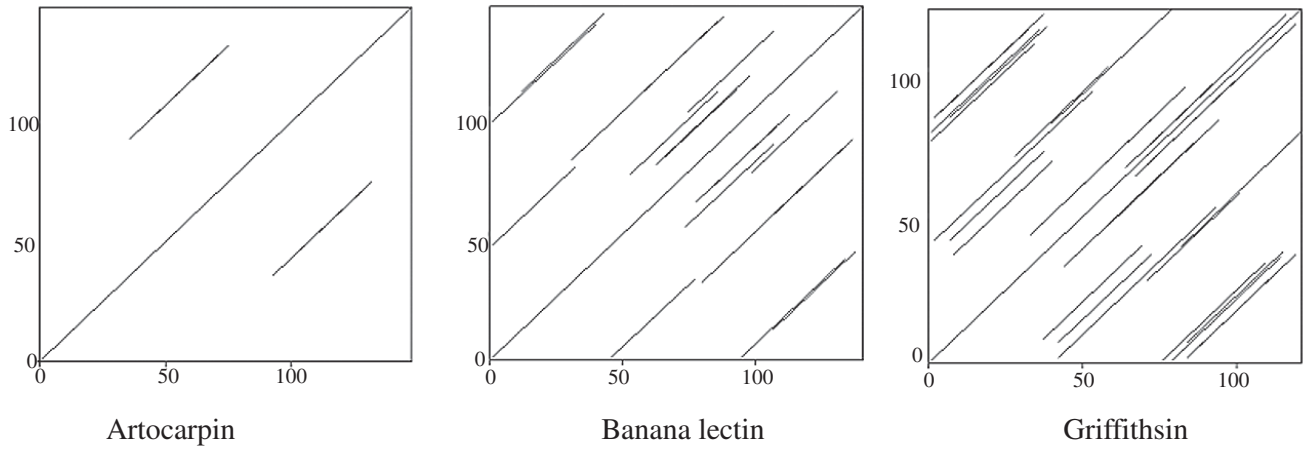


Figure 4. Dot plot representation of artocarpin, banana lectin and griffithsin sequences. In all cases window size 30 and threshhold cut-off 10 were used.

position of one subunit is not restricted by that of another subunit except through normal non-bonded interactions. However, in one instance, the outer strand of one Greek key and the same strand of a neighbouring key swap during dimerization. Thus, the sequence of the strands constituting each key remains the same. The same is true of all oligomeric *β*-prism II fold lectins. Therefore, the analysis of sequences is unaffected by strand swapping.

The analysis of the sequence relationship among the three loops in a *β*-prism I fold lectin was extended to all plant lectins listed in table 1. In those from dicots, which invariably have only one carbohydrate-binding site per subunit, the maximum sequence divergence is between keys 1 and 2, with an average similarity of only 13.5 %. That between keys 1 and 3, and 2 and 3 is in the range of 20.0–22.7%. Thus, Greek key 3, which carries the secondary-binding site in artocarpin and heltuba, has a sequence intermediate between that carrying the primary binding site (Greek key 1) and that having no binding site at all (Greek key 2). The situation in monocots is somewhat different. Irrespective of whether the subunit has one or two binding motifs, the maximum similarity among them is between

Greek keys 1 and 2, with an average similarity of 33.0% in the case of lectins with one binding motif, and 37.2% in those with two binding motifs. The average similarities between keys 2 and 3, and 1 and 3 range between 25.3% and 29.4%. This difference between β-prism I fold lectins from dicots and monocots presumably reflects the difference in the evolutionary paths the two groups followed.

There are 10 β-prism I fold lectins, galactose specific or mannose/glucose specific, with known structure. The sequences of the three Greek keys in them were individually aligned. The commonality in the 10 sequences is not very striking in each case. However, it is the highest in Greek keys 1 and 3. Greek key 1 carries a carbohydrate-binding site in all the lectins. Greek key 3 carries a secondary binding site in several of these lectins. Greek key 2 has a carbohydrate-binding site only in griffithsin and banana lectin, and it displays the lowest level of sequence conservation. Incidentally, griffithsin, which exhibits very high anti-HIV activity, presents an interesting case with three carbohydrate-binding sites forming an equilateral triangle of side about 15 Å in length. Modelling indicates that a tridentate oligosaccharide, commonly found in viral glycoprotein, simultaneously makes use of the three binding sites (Ziolkowska and Wlodawer 2006).

The β-prism II fold lectins present an altogether simpler picture. There are 7 such lectins of known structure, all from monocots. The threefold symmetry of the subunit in each lectin is reflected in the sequence as well. The average sequence similarity (identity) among the three Greek key-like motifs in them range from 46.4 to 59.0% (28.5 to 40.7%) Again, as in the case of β-prism I fold lectins, in these lectins also individual Greek keys from different lectins align better than the whole sequences do. If all the sequences with three carbohydrate-binding motifs (table 2) from monocots are taken into account, the average sequence similarity among the sheets ranges from 42.1 to 50.1%. The corresponding range in monocot lectins with only one carbohydrate-binding motif is 27.6–36.5%. The same trend in correlation between the number of carbohydrate-binding motifs and sequence identity among the keys is observed in β-prism II fold dicot lectins also, although the number of such lectins is too small for any firm statistical inference to be drawn.

Most of the β-prism II fold lectins carry three carbohydrate-binding sites, one per Greek key-like sheet, and the sheets are also related to one another through sequence similarity. This indicates the probability of gene duplication and fusion in the generation of the lectin. The situation is less obvious in β-prism I fold lectins. Therefore, a phylogenetic tree for β-prism I fold lectins of known structure was constructed using the sequence of each Greek key in each subunit as an individual unit. Nearly the same tree was obtained irrespective of the method, lending credence to the result obtained. In the tree (figure 5), individual Greek keys cluster together. Deviation from this behaviour is exhibited only by banana lectin and griffithsin, which have two and three carbohydrate-binding sites, respectively, in each subunit.

3.4 *Phylogenetic analysis of β-prism fold lectins*

The numbers of binding sites in each lectin domain do not follow strict taxonomical classification. They are, however, closely related to the similarity in sequences within the domain. This can be clearly seen in figure 6a, which gives a phylogenetic classification of β-prism II fold domains, each made up of three Greek key-like sheets, on the basis of their sequences. Although the phylogenetic tree for β-prism II fold lectins shows five major clusters, only three of them indicate clustering of sequences based on strict taxonomic positions. Most of the sequences from *Amaryllidaceae*, *Alliaceae* and *Orchidaceae* families form two independent clusters. Sequences from the *Araceae* family also constitute a major part of other clusters. Almost all the sequences in each of these clusters have the same number of carbohydrate-binding motifs. It is interesting to examine the distribution of sequences from sources other than monocots and sequences that do not contain three carbohydrate-binding motifs. The single dicot with three carbohydrate-binding motifs (AAD45250.1), although not part of any cluster, shares the same origin with the taxonomic cluster of *Amaryllidaceae* where all the sequences have three carbohydrate-binding motifs. Interestingly, two sequences from the *Araceae* family (AAP04617.1 and AAQ55289.1), which have three carbohydrate-binding motifs, do not form part of the taxonomic cluster of *Araceae* sequences with a single carbohydrate-binding motif each. Instead, they cluster with other monocot lectins with three carbohydrate-binding motifs. Similarly, the dicot sequence ABE91586.1, with a single carbohydrate-binding site, does not align with any other dicot sequence; instead, it clusters with the *Araceae* family where most of the sequences have a single carbohydrate-binding site. This clustering of sequences containing the same number of carbohydrate-binding motifs across taxonomic positions is interesting. An analysis of internal symmetry in the sequences clearly indicates that the decrease in number of carbohydrate-binding motifs is a manifestation of the decrease in sequence similarity among the three Greek key-like sheets in the sequence. Thus, the numbers of binding sites provide a reasonable explanation for the heterogeneity in the sequence-based phylogenetic analysis.

β-prism I fold lectins, which appear to be more divergent in sequence than β-prism II fold lectins, do not form well-defined taxonomic clusters (figure 6b). Sequences from *Poaceae* and *Brassicaceae* form reasonable, though not exclusive, clusters. This is probably because of the large

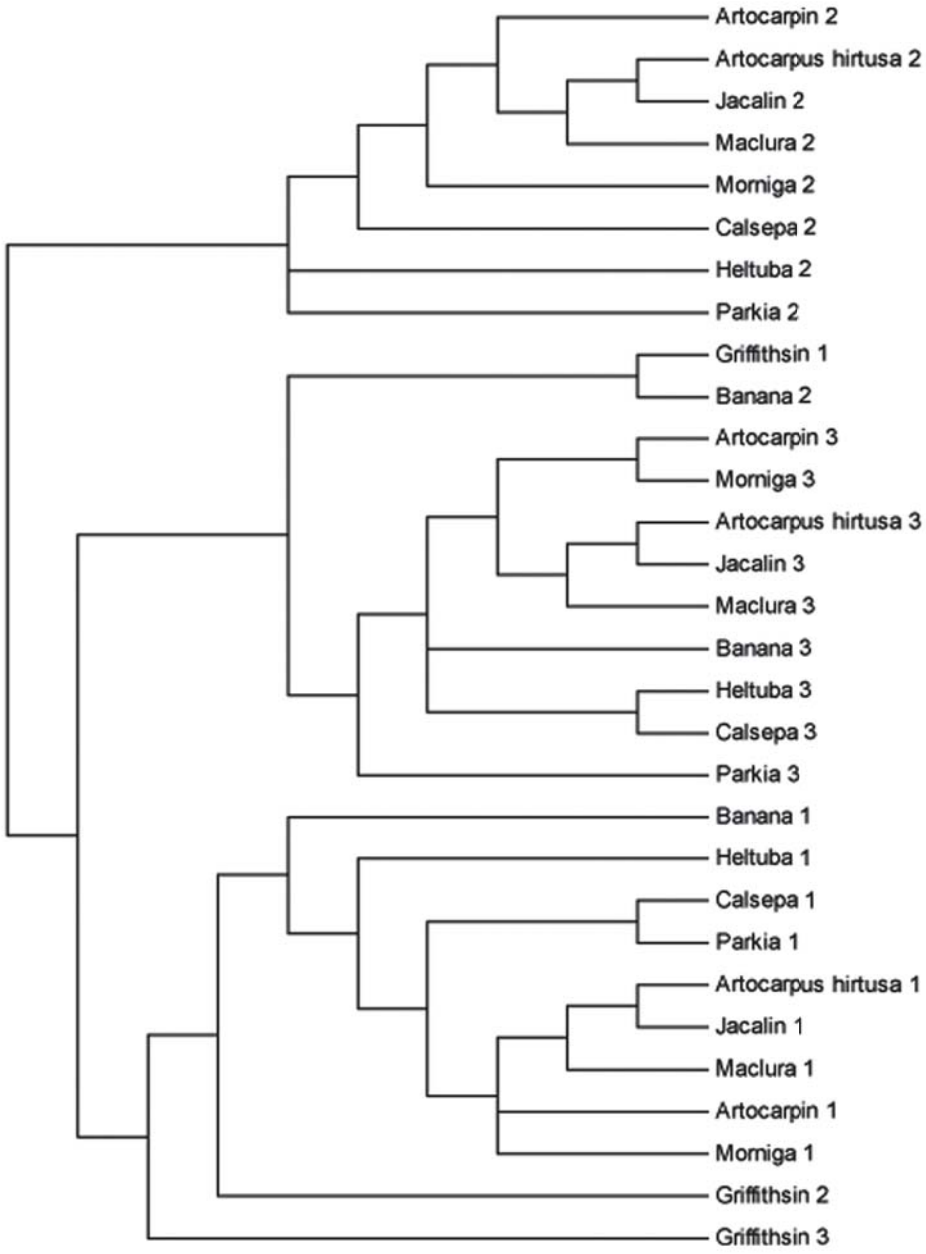
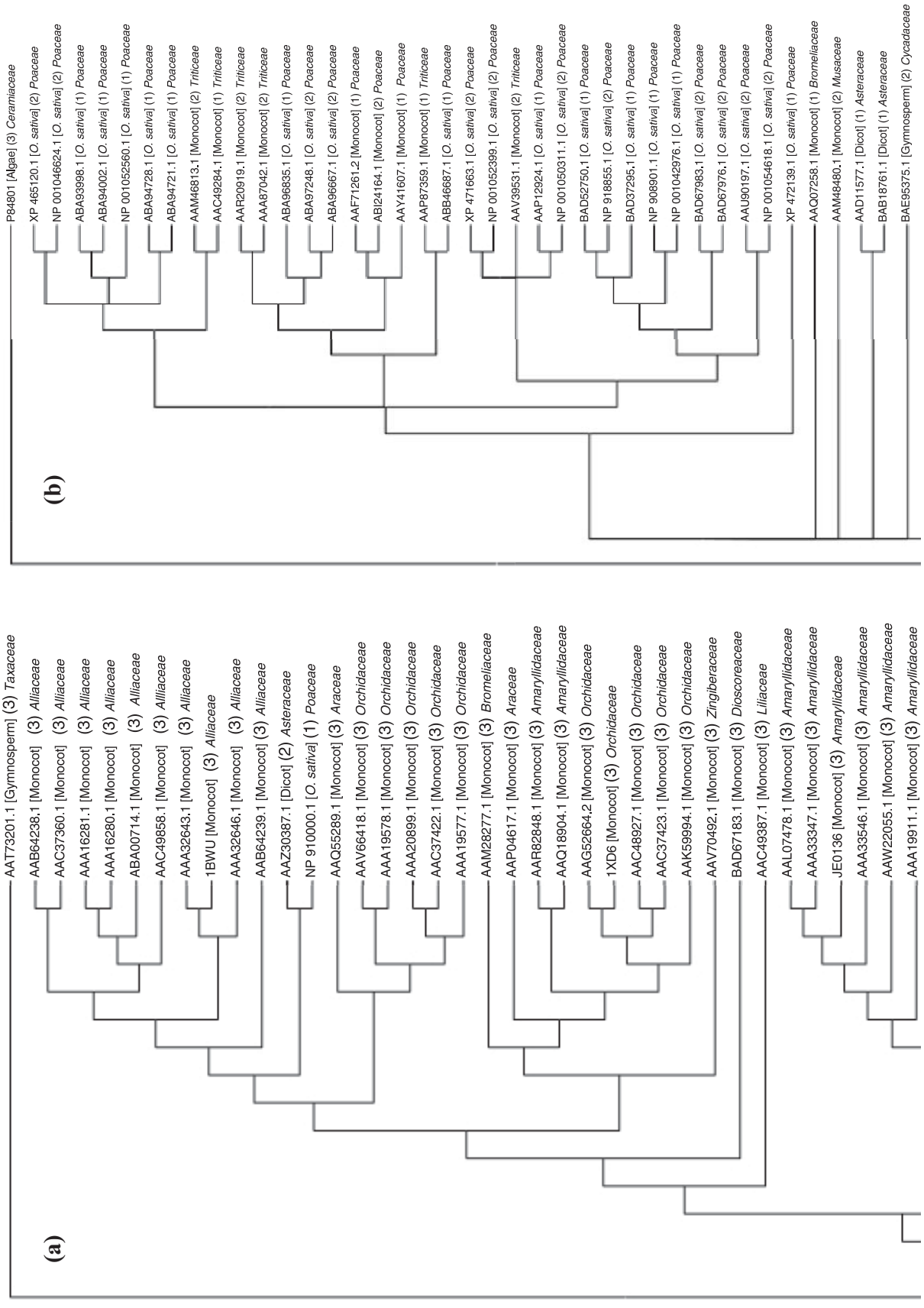


Figure 5. Phylogenetic tree produced from the multiple sequence alignment of all Greek key sequences in *β*-prism I fold lectins of known structures.

number of sequences obtained from *O. sativa* and *A. thaliana* genomes belonging to the *Poaceae* and *Brassicaceae* families, respectively. The only well-defined taxonomic cluster is made up of sequences from the *Moraceae* family. Although the tendency is not as clear as in the *β*-prism II fold lectins, there appears to be a correlation between clustering and number of carbohydrate-binding sites in the *β*-prism I fold lectins also. On the basis of number of carbohydrate-binding sites, the tree can be roughly divided in two halves; the lower half made up of all the sequences, except one, with a single carbohydrate-binding site and the upper half with sequences containing one or two carbohydrate-binding sites. The sequence from *O. sativa* (XP_47804.1) which contains a single carbohydrate-binding motif clusters in the lower half of the phylogenetic tree with dicot sequences containing a single carbohydrate-binding motif each.


(a)
AAT73201.1 [Gymnosperm] (3) Taxaceae
AAB64238.1 [Monocot] (3) Alliaceae
AAC37360.1 [Monocot] (3) Alliaceae
AAA16281.1 [Monocot] (3) Alliaceae
AAA16280.1 [Monocot] (3) Alliaceae
ABA00714.1 [Monocot] (3) Alliaceae
AAC49858.1 [Monocot] (3) Alliaceae
AAA32643.1 [Monocot] (3) Alliaceae
1BWU [Monocot] (3) Alliaceae
AAA32646.1 [Monocot] (3) Alliaceae
AAB64239.1 [Monocot] (3) Alliaceae
AAZ30387.1 [Dicot] (2) Asteraceae
NP 910000.1 [O. sativa] (1) Poaceae
AAQ55289.1 [Monocot] (3) Araceae
AAV66418.1 [Monocot] (3) Orchidaceae
AAA19578.1 [Monocot] (3) Orchidaceae
AAA20899.1 [Monocot] (3) Orchidaceae
AAC37422.1 [Monocot] (3) Orchidaceae
AAA19577.1 [Monocot] (3) Orchidaceae
AAM28277.1 [Monocot] (3) Bromeliaceae
AAP04617.1 [Monocot] (3) Araceae
AAR82848.1 [Monocot] (3) Amaryllidaceae
AAQ18904.1 [Monocot] (3) Amaryllidaceae
AAG52664.2 [Monocot] (3) Orchidaceae
1XD6 [Monocot] (3) Orchidaceae
AAC48927.1 [Monocot] (3) Orchidaceae
AAC37423.1 [Monocot] (3) Orchidaceae
AAK59994.1 [Monocot] (3) Orchidaceae
AAV70492.1 [Monocot] (3) Zingiberaceae
BAD67183.1 [Monocot] (3) Dioscoreaceae
AAC49387.1 [Monocot] (3) Liliaceae
AAL07478.1 [Monocot] (3) Amaryllidaceae
AAA33347.1 [Monocot] (3) Amaryllidaceae
JE0136 [Monocot] (3) Amaryllidaceae
AAA33546.1 [Monocot] (3) Amaryllidaceae
AAW22055.1 [Monocot] (3) Amaryllidaceae
AAA19911.1 [Monocot] (3) Amaryllidaceae
(b)
P84801 [Algae] (3) Ceramiaceae
XP 465120.1 [O. sativa] (2) Poaceae
NP 001046624.1 [O. sativa] (2) Poaceae
ABA93998.1 [O. sativa] (1) Poaceae
ABA94002.1 [O. sativa] (1) Poaceae
NP 001052560.1 [O. sativa] (1) Poaceae
ABA94728.1 [O. sativa] (1) Poaceae
ABA94721.1 [O. sativa] (1) Poaceae
AAM46813.1 [Monocot] (2) Triticeae
AAC49284.1 [Monocot] (1) Triticeae
AAR20919.1 [Monocot] (2) Triticeae
AAA87042.1 [Monocot] (2) Triticeae
ABA96835.1 [O. sativa] (1) Poaceae
ABA97248.1 [O. sativa] (2) Poaceae
ABA96667.1 [O. sativa] (2) Poaceae
AAF71261-2 [Monocot] (1) Poaceae
ABI24164.1 [Monocot] (2) Poaceae
AAY41607.1 [Monocot] (1) Poaceae
AAP87359.1 [Monocot] (1) Triticeae
ABB46687.1 [O. sativa] (1) Poaceae
XP 471663.1 [O. sativa] (2) Poaceae
NP 001052399.1 [O. sativa] (2) Poaceae
AAV39531.1 [Monocot] (2) Triticeae
AAP12924.1 [O. sativa] (1) Poaceae
NP 001050311.1 [O. sativa] (2) Poaceae
BAD52750.1 [O. sativa] (1) Poaceae
NP 918855.1 [O. sativa] (2) Poaceae
BAD37295.1 [O. sativa] (1) Poaceae
NP 908901.1 [O. sativa] (1) Poaceae
NP 001042976.1 [O. sativa] (1) Poaceae
BAD67983.1 [O. sativa] (2) Poaceae
BAD67976.1 [O. sativa] (2) Poaceae
AAU90197.1 [O. sativa] (2) Poaceae
NP 001054618.1 [O. sativa] (2) Poaceae
XP 472139.1 [O. sativa] (1) Poaceae
AAQ07258.1 [Monocot] (1) Bromeliaceae
AAM48480.1 [Monocot] (2) Musaceae
AAD11577.1 [Dicot] (1) Asteraceae
BAB18761.1 [Dicot] (1) Asteraceae
BAE95375.1 [Gymnosperm] (2) Cycadaceae

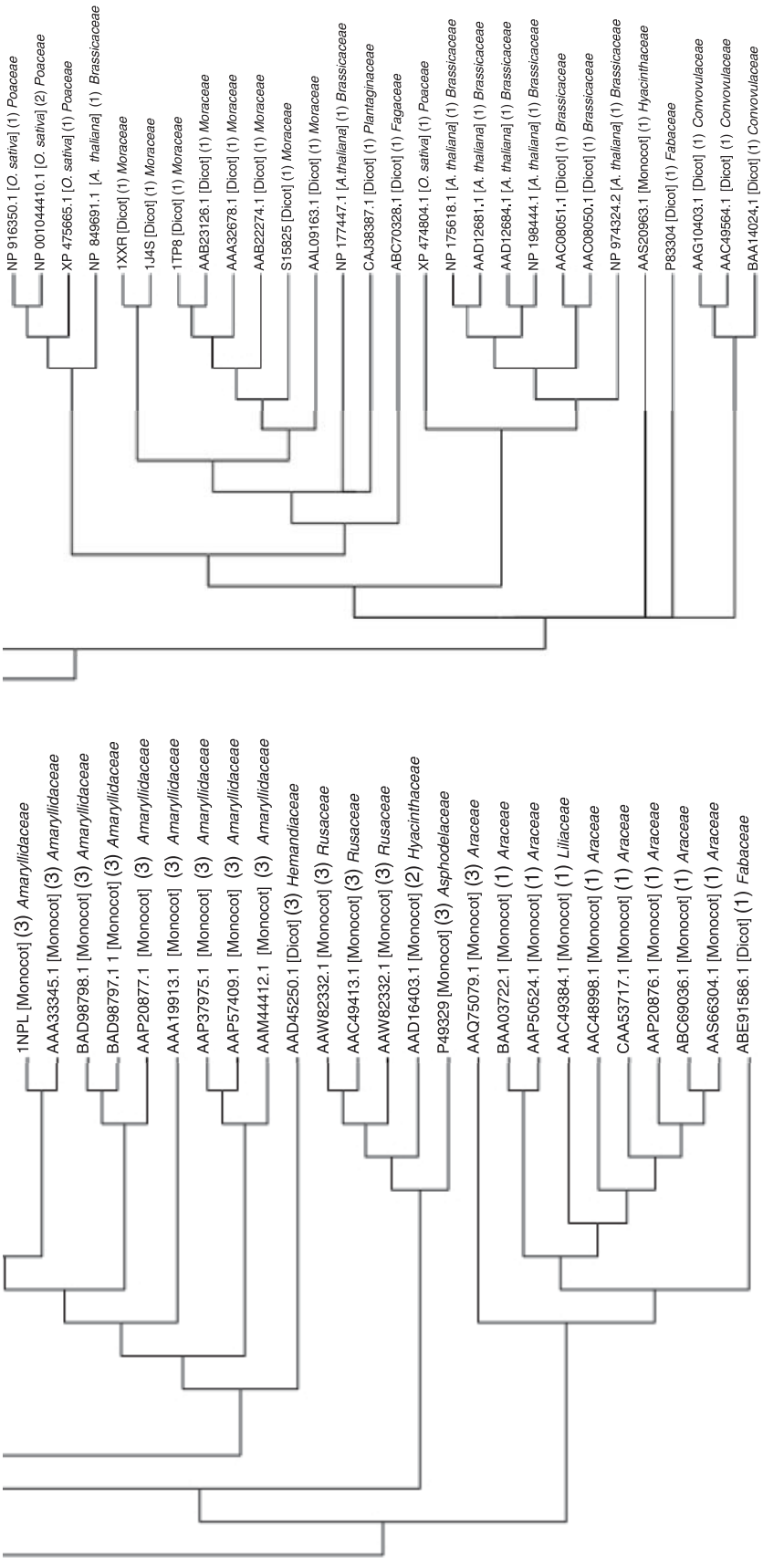


Figure 6. Phylogenetic tree produced from the multiple sequence alignment of **(a)** garlic lectin homologues and **(b)** banana lectin homologues. Accession numbers are followed by organism class, as listed in tables 1 and 2. The number of carbohydrate-binding motifs per domain for each accession number is shown in brackets. Corresponding taxonomic family names have been given in italics.

3.5 *Possible evolutionary relationship between the two β-prism folds*

Interestingly, carbohydrate-binding motifs are exhibited by only a subset of domains from among those that are suggested to have the apparent *β*-prism fold by sequence comparison. This is particularly true in the case of the *β*-prism II fold. It is therefore difficult to address the evolutionary relatedness between the proteins which exhibit carbohydrate-binding motifs and those which do not.

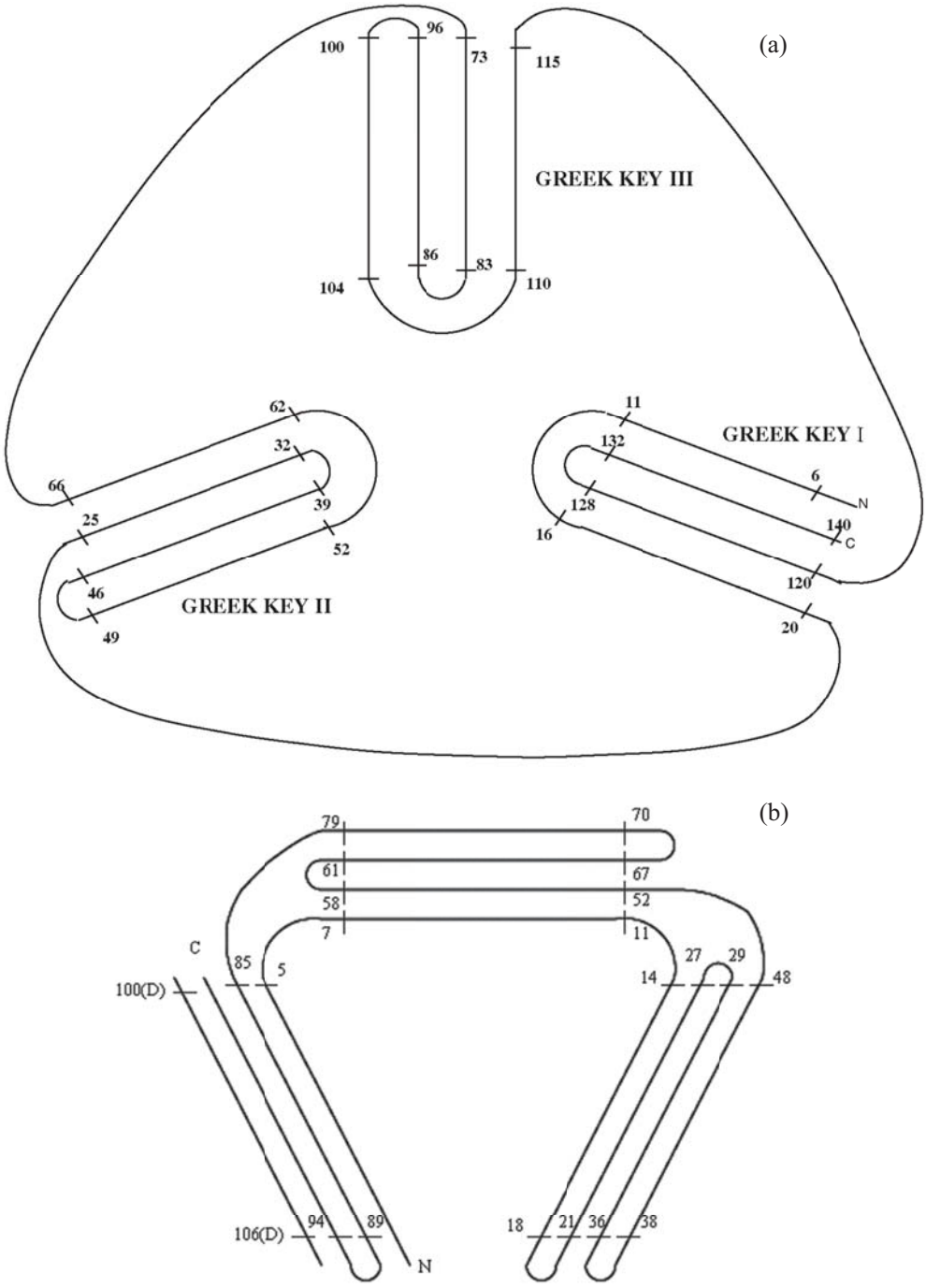


Figure 7. Schematic representations of the topology of **(a)** the banana lectin fold and **(b)** the garlic lectin fold. In **(b)**, the strand from the other subunit involved in swapping to form the complete third Greek key has been labelled as 100(D) to 106(D).

Instances of proteins unrelated in sequence but sharing the same fold are in any case very well known. Both the *β*-prism folds can be considered to have been generated by different types of fusion of the three Greek keys. The Greek key is such a well-known and stable super-secondary structural element that proteins resulting from their fusion could have originated primarily for reasons of structural viability. Therefore, it is reasonable to expect the existence of *β*-prism folds without any sequence or functional similarity with the *β*-prism fold lectins. Hence, the discussion here is confined to proteins listed in tables 1 and 2, which exhibit sequence similarity with the *β*-prism fold lectins and also carry the characteristic lectin carbohydrate-binding motifs.

In the absence of independent experimental confirmation, it is difficult to assert that all the *β*-prism fold proteins which carry the appropriate carbohydrate-binding motif are lectins. Such definitive confirmation is not available in most cases. However, indications are that most such proteins in plants have the lectin function. Furthermore, the evolutionary pressures that plants undergo are perhaps different from those that other groups of organisms are subjected to. In any case, plants form the majority of sources of the sequences listed in tables 1 and 2. Therefore, perhaps plants constitute the best group to explore the possible evolutionary implications of the observations summarized in the tables.

Interestingly, a *β*-prism I fold lectin, griffithsin is found in a red alga (Chandra 2006); and it has been suggested that the appearance of the red algae in evolution was prior to that of the leading to modern plants, animals and fungi (Stiller and Hall 1997). This lectin has three carbohydrate-binding sites. In the course of divergent evolution, *β*-prism I fold lectins in plants appear to have lost one or more carbohydrate-binding sites. Dicots almost invariably lost two while loss of one site and that of two sites appear to have been nearly equally distributed in monocots. A somewhat analogous situation exists in the case of *β*-prism II fold lectins in plants. It exists with three carbohydrate-binding sites in the lectin from gymnosperm, a primitive plant. Lectins from most of the monocots retain the three sites, but there are some with one binding site and at least one with two binding sites. *β*-prism II fold carbohydrate-binding domains occur only sparingly in dicots. Of the three that have been identified, one has three binding motifs, another two motifs and the third one motif.

Successive gene duplication and fusion, and the extent of divergent evolution are reflected in the internal symmetry of the sequence containing the three Greek keys. Griffithsin appears to be the earliest plant or plant-like organism in which a *β*-prism I fold lectin has been identified. Its structure has also been reported very recently (Chandra 2006, Ziolkowska *et al* 2006). The organism in the case of *β*-prism II fold lectin is *Taxus* x *media*. These possible ancestors not only have three carbohydrate-binding motifs, but they also share high sequence similarity among the three Greek key motifs. The internal sequence similarity is retained in all the lectins containing three carbohydrate-binding motifs. The similarity decreases with the decrease in the number of carbohydrate-binding motifs; although the fold remains the same, the carbohydrate–lectin stoichiometry decreases.

The information presented in tables 1 and 2 also appears to suggest that the differences between *β*-prism I fold and *β*-prism II fold lectins are not as clear cut as normally believed. *β*-prism fold I lectins occur in dicots and monocots. *β*-prism II fold lectins, originally christened as monocot lectins, appear at least in three dicots. Both types of lectins

```
Artocarpin      -SQTITVGPWGGPGGNGWDDGSYTGIRQIELSYKEAIGSFSVIYDLNGDPFSGPKHTSKL 60
Banana          MNGAIKVGAWGGNGGSAFDMGPAYRIISVKIFSGDVVDAVDVTFTYYG--KTETRHFGGS 58
Griffithsin     ----LTHRKFGGSGGS-----PFSGLSSIAVRSGSYLDAIIIDGVHHG---------GSG 42
Garlic          --RNILRNDEGLYGGQSLDVNPYHFIMQEDCNLVLYDHSTSVWASNTG----------IL 48
                    :     *  **.     .   : .          :  :     *

Artocarpin      PYKNVKIELKFPDEFLESVSGYTGPFSALATPTPVVRSLTFKTNKGRTFGPYGDEEGTYF 120
Banana          GGTPHEIVLQE-GEYLVGMKGEFGNYHG----VVVVGKLGFSTNK-KSYGPFGNTGGTPF 112
Griffithsin     GNLSPTFTFGS-GEYISNMTIRSGDYID---------NISFETNMGRRFGPYGGSGGSAN 92
Garlic          GKKGCRAVLQSDGNFVVYDAE---------------GRSLWASHSVRGNGNY-------- 85
                       :   .:::           :     ::    : ::  :  * :

Artocarpin      NLPIENGLIVGFKGRTGDLLDAIGIHMSL- 149
Banana          SLPIAAGKISGFFGRGGDFIDAIGVYLEP- 141
Griffithsin     --TLSNVKVIQINGSAGDYLDSLDIYYEQY 120
Garlic          --------VLVLQEDGNVVIYRSDIWSTN- 106
                 : :    :  :    .  :   .:
```

Figure 8. A multiple sequence alignment of artocarpin, banana lectin, griffithsin and garlic lectin. Carbohydrate-binding motif(s) are underlined.

simultaneously occur at least in one of them (*Helianthus tuberosus*). The topology of the two folds, as exemplified by banana lectin and garlic lectin, are shown in figures 7a and b. Both can be considered to have evolved through successive gene duplication and fusion from a primitive gene corresponding roughly to a Greek key, except that the Greek keys are assembled in different ways in the folds. The analysis presented here is in conformity with this hypothesis. Both the folds might have existed even before monocots and dicots originated. Monocots and dicots use both the folds, but perhaps with different frequencies. In relation to the elimination of one or two binding sites also, evolution seems to have proceeded in them at different rates and perhaps in different ways. Interestingly, when the sequences corresponding to the three Greek keys in garlic lectin, artocarpin, banana lectin and *Griffithsia* lectin are aligned, as in figure 8, the carbohydrate motifs, where present, occur at similar locations in the sequences. Garlic lectin has a β-prism II fold with three carbohydrate-binding sites; the remaining three have β-prism I fold with one, two and three binding sites, respectively. Also, when the sequence of garlic lectin was threaded on a banana backbone and vice versa, and the models thus generated were energy minimized, the resulting structure not only does not show any steric clash among the residues but all the residues also lie within the Ramachandran allowed area. This suggests that sequences of β-prism fold lectins are at least structurally compatible with both the folds.

4. Conclusion

Among the β-prism I fold lectins of known three-dimensional structure, all except banana and griffithsin lectin carry one carbohydrate-binding site per domain/subunit; banana lectin carries two; griffithsin, the structure of which has been determined recently, carries three. Three binding sites are found in each domain/subunit of β-prism II fold lectins of known structure. The analysis presented here shows that β-prism fold lectins display a much higher diversity in carbohydrate-binding than indicated by the above data. This diversity does not follow any taxonomic pattern in a convincing manner. However, the number of carbohydrate-binding sites correlates reasonably well with the symmetry within the sequence. In the case of both the structural families of lectins, the lectin from the earliest plant or plant-like organism known to harbour it has three carbohydrate-binding sites and a substantially threefold symmetrical sequence. In the course of divergent evolution, the symmetry and the number of binding sites tended to decrease simultaneously, at different rates and probably in different ways, in the two families. Both the folds are produced essentially by combining the three Greek keys in different ways. It is possible that both of them arose through successive gene duplication, fusion and divergent evolution of the same primitive carbohydrate-binding motif involving a Greek key.

Plant lectins are believed to be involved in defending the organisms to which they belong from predators and infectious pathogens. There are obviously other modes of defence as well. For example, and with particular reference to the present discussion, dicots have a cork cambium layer protecting them while such a layer does not exist in monocots. Thus, in the absence of this defence system, monocots perhaps have a higher dependence on other defence mechanisms including those involving lectins. In this context, it is probably significant that almost all monocot β-prism II fold lectins have three carbohydrate-binding sites. β-prism II fold lectins rarely occur in dicots. β-prism I fold lectins occur in dicots and monocots. Those from dicots invariably carry only one carbohydrate-binding motif while the number varies between one and two in monocot β-prism I fold lectins. It is reasonable to consider the number of carbohydrate-binding sites per domain as an indication of the strength and extent of carbohydrate-binding. Larger numbers also give an additional edge to multivalency (Ramachandraiah *et al* 2003), which is important in agglutination. It is interesting to note that in the rare instances where a monocot β-prism II fold domain has only one binding site, two or three such domains tend to occur in the same protein. This could conceivably be so to compensate in some manner for the loss of binding sites in the single domain.

Acknowledgements

Financial assistance from the Department of Science and Technology is acknowledged. MV is supported by a Distinguished Biotechnologist award of the Department of Biotechnology. AS is a senior research fellow of the Council of Scientific and Industrial Research.

References

Altschul S F, Thomas L M, Alejandro A S, Jinghui Z, Zheng Z, Webb M and Lipman D J 1997 Gapped BLAST and PSI-BLAST: a new generation of protein database search programs; *Nucleic Acids Res.* **25** 3389–3402

Banerjee R, Mande S C, Ganesh V, Das K, Dhanaraj V, Mahanta S K, Suguna K, Surolia A and Vijayan M 1994 Crystal structure of peanut lectin, a protein with an unusual quaternary structure; *Proc. Natl. Acad. Sci. USA* **91** 227–231

Bates P A, Kelley L A, MacCallum R M and Sternberg M J E 2001 Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM; *Proteins: Struct. Funct. Genet. (Suppl.)* **5** 39–46

Bates P A and Sternberg M J E 1999 Model building by comparison at CASP3: using expert knowledge and computer automation; *Proteins Struct. Funct. Genet. (Suppl.)* **3** 47–54

Berman H M, Westbrook J, Feng Z, Gilliland G, Bhat T N, Weissig H, Shindyalov I N and Bourne P E 2000 The Protein Data Bank; *Nucleic Acids Res.* **28** 235–242

Bourne Y, Roig-Zamboni V, Barre A, Peumans W J, Houles Astoul C, Van Damme E J and Rouge P 2004 The crystal structure of the *Calystegia sepium* agglutinin reveals a novel quaternary arrangement of lectin subunits with a *β*-prism fold; *J. Biol. Chem.* **279** 527–533

Bourne Y, Zamboni V, Barre A, Peumans W J, Van Damme E J M and Rouge P 1999 *Helianthus tuberosus* lectin reveals a widespread scaffold for mannose-binding lectins; *Structure* **7** 1473–1482

Chandra N R, Ramachandraiah G, Bachhawat K, Dam T K, Surolia A and Vijayan M 1999 Crystal structure of a dimeric mannose-specific agglutinin from garlic: quaternary association and carbohydrate specificity; *J. Mol. Biol.* **285** 1157–1168

Chandra N 2006 Common scaffolds, diverse recognition profiles; *Structure* **14** 1093–1094

Chantalat L, Wood S D, Rizkallah P and Reynolds C D 1996 X-ray structure solution of *Amaryllis* lectin by molecular replacement with only 4% of the total diffracting matter; *Acta Crystallogr. Sect. D* **52** 1146–1152

Chrispeels M J and Raikhel N V 1991 Lectins, lectin genes, and their role in plant defense; *Plant Cell* **3** 1–9

Contreras-Moreira B and Bates P A 2002 Domain fishing: a first step in protein comparative modeling; *Bioinformatics* **18** 1141–1142

Delbaere L T J, Vandonselaar M, Prasad L, Quail J W, Wilson K S and Dauter Z 1993 Structure of the lectin IV of *Griffonia simplicifolia* and its complex with the Lewis b human blood group determinant at 2.0 Å resolution; *J. Mol. Biol.* **230** 950–965

Drickamer K 1999 C-type lectin-like domains; *Curr. Opin. Struct. Biol.* **9** 585–590

Elgavish S and Shaanan B 1998 Structures of the *Erythrina corallodendron* lectin and of its complexes with mono- and disaccharides; *J. Mol. Biol.* **277** 917–932

Emsley P and Cowtan K 2004 Coot: model-building tools for molecular graphics; *Acta Crystallogr. Sect. D* **60** 2126–2132

Gallego del Sol F, Nagano C, Cavada B S and Calvete J J 2005 The first crystal structure of a *Mimosoideae* lectin reveals a novel quaternary arrangement of a widespread domain; *J. Mol. Biol.* **353** 574–583

Hester G, Kaku H, Goldstein I J and Wright C S 1996 Structure of mannose-specific snowdrop (*Galanthus nivalis*) lectin is representative of a new plant lectin family; *Nat. Struct. Biol.* **2** 472–479

Hirsch A M 1999 Role of lectins and rhizobial exopolysaccharides in legume nodulation; *Curr. Opin. Plant Biol.* **2** 320–326

Huelsenbeck J P and Ronquist F 2001 MRBAYES: Bayesian inference of phylogeny; *Bioinformatics* **17** 754–755

Imberty A, Wimmerova M, Mitchell E P and Gilboa-Garber N 2004 Structures of the lectins from *Pseudomonas aeruginosa*: insight into the molecular basis for host glycan recognition; *Microbes Infect.* **6** 221–228

Jeyaprakash A A, Jayashree G, Mahanta S K, Swaminathan C P, Sekar K, Surolia A and Vijayan M 2005 Structural basis for the energetics of jacalin–sugar interactions: promiscuity versus specificity; *J. Mol. Biol.* **347** 181–188

Jeyaprakash A A, Srivastav A, Surolia A and Vijayan M 2004 Structural basis for the carbohydrate specificities of artocarpin: variation in the length of a loop as a strategy for generating ligand specificity; *J. Mol. Biol.* **338** 757–770

Kumar S, Tamura K and Nei M 2004 MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment; *Brief Bioinform.* **5** 150–163

Lee X, Thompson A, Zhang Z, Ton-that H, Biesterfeldt J, Ogata C Xu L, Johnston R A and Young N M 1998 Structure of the complex of *Maclura pomifera* agglutinin and the T-antigen disaccharide, Gal*β*1,3GalNAc; *J. Biol. Chem.* **273** 6312–6318

Lehr C M and Gabor F 2004 Lectins and glycoconjugates in drug delivery and targeting; *Adv. Drug. Deliv. Rev.* **56** 419–420

Li W, Jaroszewski L and Godzik A 2001 Clustering of highly homologues sequences to reduce the size of large protein databases; *Bioinformatics* **17** 282–283

Li W, Jaroszewski L and Godzik A 2002 Tolerating some redundancy significantly speeds up clustering of large protein databases; *Bioinformatics* **18** 77–82

Lis H and Sharon N 1998 Lectins: carbohydrate specific proteins that mediate cellular recognition; *Chem. Rev.* **98** 637–664

Loris R 2002 Principles of structures of animal and plant lectins; *Biochim. Biophys. Acta* **238** 777–793

Marchel-Bauer A, Panchenko A R, Shoemaker B A, Thiessen P, A, Geer L Y and Bryant S H 2002 CDD: a database of conserved domain alignments with links to domain three-dimensional structure; *Nucleic Acids Res.* **30** 281–283

Meagher J L, Winter H C, Ezell P, Goldstein I J and Stuckey J A 2005 Crystal structure of banana lectin reveals a novel second carbohydrate binding site; *Glycobiology* **15** 1033–1042

Navarro-Gochicoa M, Camut S, Timmers C J T, Niebel A, Herve C, Boutet E, Bono J, Imberty A and Cullimore J V 2003 Characterization of four lectin like receptor kinases expressed in roots of *Medicago truncatula.* Structure, location, regulation of expression, and potential role in the symbiosis with *Sinorhizobium meliloti*; *Plant Physiol.* **133** 1893–1910

Peumans W J and Van Damme E J M 1995 Lectins as plant defense proteins; *Plant Physiol.* **109** 347–352

Prabu M M, Suguna K and Vijayan M 1999 Variability in quaternary association of proteins with the same tertiary fold: a case study and rationalization involving legume lectins; *Proteins* **35** 58–69

Pratap J V, Jeyaprakash A A, Rani P G, Sekar K, Surolia A and Vijayan M 2002 Crystal structures of artocarpin, a Moraceae lectin with mannose specificity, and its complex with methyl-α-D-mannose: implications to the generation of carbohydrate specificity; *J. Mol. Biol.* **317** 237–247

Rabijns A, Barre A, Van Damme E J, Peumans W J, De Ranter C J and Rouge P 2005 Structural analysis of the jacalin-related lectin Morniga M from the black mulberry (*Morus nigra*) in complex with mannose; *FEBS J.* **272** 3725–3732

Ramachandraiah G, Chandra N R, Surolia A and Vijayan M 2003 Computational analysis of multivalency in lectins: structures of

garlic lectin–oligosaccharide complexes and their aggregates; *Glycobiology* **13** 765–775

Ramachandraiah G and Chandra N R 2000 Sequence and structural determinants of mannose recognition; *Proteins: Struct. Funct. Genet.* **39** 358–364

Rao K N, Suresh C G, Katre U V, Gaikwad S M and Khan M I 2004 Two orthorhombic crystal structures of a galactose-specific lectin from *Artocarpus hirsuta* in complex with methyl-*α*-D-galactose; *Acta Crystallog. Sect. D* **60** 1404–1412

Raval S, Gowda S B, Singh D D and Chandra N R 2004 A database analysis of jacalin-like lectins: sequence–structure–function relationships; *Glycobiology* **14** 1–17

Rice P, Longden I and Bleasby 2000 EMBOSS: the European Molecular Biology Open Software Suite; *Trends Genet.* **16** 276–277

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

Rost B 1996; PHD: predicting one-dimensional protein structure by profile based neural networks; *Methods Enzymol.* **266** 525–539

Sankaranarayanan R, Sekar K, Banerjee R, Sharma V, Surolia A and Vijayan M 1996 A novel mode of carbohydrate recognition in jacalin; a Moraceae plant lectin with a *β*-prism fold; *Nat. Struct. Biol.* **3** 596–602

Sauerborn M, Wright L M, Reynolds C D, Grossmann J G and Rizkallah P J 1999 Insights into carbohydrate recognition by *Narcissus pseudonarcissus* lectin: the crystal structure at 2Å resolution in complex with α 1-3 mannobiose; *J. Mol. Biol.* **290** 185–199

Schaffer, Alejandro A, Aravind L, Madden T L, Shavirin S, John L S, Yuri I W, Koonin E V and Altschul S F 2001 Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements; *Nucleic Acids Res.* **29** 2994–3005

Seiler M, Mehrle A, Poustka A and Wiemann S 2006 The 3of5 web application for complex and comprehensive pattern matching in protein sequences; *BMC Bioinformatics* **7** 144

Singh D D, Saikrishnan K, Kumar P, Surolia A, Sekar K and Vijayan M 2005 Unusual sugar specificity of banana lectin from *Musa paradisiaca* and its probable evolutionary origin: crystallographic and modelling studies; *Glycobiology* **15** 1025–1032

Sonnhammer Erik L L and Durbin R 1995 A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis; *Gene* **167** GC1–10

Stiller J W and Hall B D 1997 The origin of red algae: Implications for plastid evolution; *Proc. Natl. Acad. Sci. USA* **94** 4520–4525

Thompson J D, Higgins D G and Gibson T J 1994 CLUSTALW: improving the sensitivity of progressive multiple sequence align-ment through sequence weighting, position-specific gap penalties and weight matrix choice; *Nucleic Acids Res.* **22** 4673–4680

Vijayan M and Chandra N 1999 Lectins; *Curr. Opin. Struct. Biol.* **9** 707–714

Wood S D, Wright L M, Reynolds C D, Rizkallah P J, Allen A K, Peumans W J and Van Damme E J M 1999 Structure of the native (unligated) mannose-specific bulb lectin from *Scilla campanulata* (bluebell) at 1.7 Å resolution; *Acta Crystallogr. Sect. D* **55** 1264–1272

Yen-Chieh H, Yi-Hung L, Chia-Hao S, Chun-Liang S, Tschining C and Chun-Jung C 2006 Purification, crystallization and preliminary X-ray crystallographic analysis of rice lectin from *Oryza sativa*; *Acta Crystallogr. Sect. F* **62** 94–96

Ziolkowska N E, O'Keefe B R, Mori T, Zhu C, Glomarelli B, Vojdani F, Palmer K E, McMahon J B and Wlodawer A 2006 Domain-swapped structure of the potent antiviral protein griffithsin and its mode of carbohydrate binding; *Structure* **14** 1127–1135

Ziolkowska N E and Wlodawer A 2006 Structural studies of algal lectins with anti-HIV activity; *Acta Biochim. Pol.* [Epub ahead of print]

MS received 19 January 2007; accepted 26 April 2007

*e*Publication: 7 July 2007

Corresponding editor: Vidyanand Nanjundiah