PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Archeal lectins: An identification through a genomic search

K. V. Abhinav, Ebenezer Samuel, and M. Vijayan*

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

ABSTRACT

Forty-six lectin domains which have homologues among well established eukaryotic and bacterial lectins of known three-dimensional structure, have been identified through a search of 165 archeal genomes using a multipronged approach involving domain recognition, sequence search and analysis of binding sites. Twenty-one of them have the 7-bladed β-propeller lectin fold while 16 have the β-trefoil fold and 7 the legume lectin fold. The remainder assumes the C-type lectin, the β-prism I and the tachylectin folds. Acceptable models of almost all of them could be generated using the appropriate lectins of known three-dimensional structure as templates, with binding sites at one or more expected locations. The work represents the first comprehensive bioinformatic study of archeal lectins. The presence of lectins with the same fold in all domains of life indicates their ancient origin well before the divergence of the three branches. Further work is necessary to identify archeal lectins which have no homologues among eukaryotic and bacterial species.

Proteins 2015; 00:000–000.

Key words: lectin homologues; microbial lectins; domain identification; 7-bladed β-propeller fold; β-trefoil; evolutionary implications.

INTRODUCTION

Lectins, commonly described as multivalent carbohydrate binding proteins of nonimmune origin, were first identified in plants and their best known property was the ability to agglutinate erythrocytes. Now they are known to occur in all kingdoms of life.[1] The vast repertoire of functions carried out by lectins include cell–cell and host–pathogen interactions, serum glycoprotein turnover, cellular signalling, differentiation, pathogen recognition, complement activation, lectinophagocytosis, cell–cell, and self-nonself recognition, opsonisation, and innate immune responses.[2–8] All of them are based on the ability of lectins to specifically bind to different sugar structures. Of the known lectins, those from eukaryotes, particularly animals and plants, have been well characterized.[9,10] They assume widely different folds, the only common property shared by them being the ability to specifically bind carbohydrates. For instance, in terms of fold, plant lectins belong to six structural classes (http://www.glyco3d.cermav.cnrs.fr/home.php), including the one discovered in this laboratory. Even those belonging to the same structural class assume widely different quaternary structures[11,12] which probably serve to generate different kinds of multivalency.[13] Lectins are also known to employ widely different strategies for generating ligand specificity.[2,14]

Studies on microbial lectins have not been as extensive as on those from plants and animals, although there have been some outstanding individual investigations on bacterial[15–19] and viral lectins.[20] One focussed effort in the area has been our investigation of mycobacterial lectins[21] as an extension of our long-term programme of structural and related studies on plant lectins.[22] In particular, a search of mycobacterial genomes with known sequences resulted in the identification of 94 lectins belonging to five structural families, which have homologues among the well established lectins of known three-dimensional structure.[23] Here we present a similar genomic search for archeal lectins. So far hardly anything is known on archeal lectins and they constitute a virgin area of research. Here again, our approach has been rigorous and only those which are homologous to lectins with known three-dimensional structure, have been

Additional Supporting Information may be found in the online version of this article.

*Correspondence to: M. Vijayan; Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560 012, India. E-mail: mv@mbu.iisc.ernet.in

Received 20 August 2015; Revised 13 October 2015; Accepted 19 October 2015
Published online 22 October 2015 in Wiley Online Library (wileyonlinelibrary.com). DOI: 10.1002/prot.24949

considered as candidate archeal lectins. Furthermore, the presence of appropriate binding site residues was ensured before final identification. Thus, it is not claimed that all archeal lectins have been identified, but those identified have a very high probability of being lectins. It was felt that while exploring a virgin area, a cautious approach is perhaps the best one to follow. Using our approach which predominantly involves data mining, fold assignment, sequence search and binding site analysis, we report the identification of a total of 46 archeal lectins distributed among six structural families. These families involve the 7-bladed β-propeller, the β-trefoil, the legume lectin, β-prism I, the C-type lectin, and the 5-bladed β-propeller lectin folds.

METHODS

Retrieval of whole genome sequences of archea

Genome sequence information on a total of 165 different completed genome projects spanning 78 different archeal species (as given in Supporting Information Table SI) were retrieved from the NCBI ftp server (http://www.ncbi. nlm.nih.gov/Ftp/genomes/). The information included the whole genome sequence file and the translated ORF protein sequences.

Domain identification

To start with, the Conserved Domain Database (CDD)[24] webserver at NCBI was used in the domain identification procedure. This involved first mapping all the lectin sequences from Lectin3D database[1] on to their respective domains. It was followed by generation of the domain definitions for translated ORFs of all the archeal genomes using the same server with identical parameters. Archeal proteins belonging to families overlapping with those of proteins from the lectin structural database mentioned above were selected for further consideration. Thus, the resulting ORFs included only those proteins which have homologues among the lectins of known three-dimensional structure compiled in the Lectin3D database. These sequences were further examined using the template based homology modelling PHYRE webserver[25] for their ability to form folded structures of the lectin domains. In parallel, a similar domain identification procedure was carried out using the SCOP server[26] as the starting point. The results of both these procedures converged. A consensus involving the results of PHYRE and SCOP was used for deciding domain boundaries. The resulting sequences, which formed well folded homology models wholly or substantially similar to those of lectins with known three-dimensional structure, were further examined in terms of their carbohydrate binding sites. The carbohydrate binding sites in terms of the amino acid residues in them and their locations in the sequence have been well characterized in lectins of known structure. Those archeal sequences which show substantial presence of these residues were finally accepted as archeal lectins. SCOP, CDD and PHYRE were also used to identify other, nonlectin domains, which occur in tandem with lectin domains in the chosen archeal sequences.

Sequence based search

A query database was constructed using the sequences of lectins listed in Lectin3D database. The data were made nonredundant by clustering the sequences with >90% sequence identity using a standalone version of CD-HIT.[27] PSI-BLAST was used for searching the genomic data using the query database. The programme was installed locally by implementing the BLAST 2.2.12[28,29] package downloaded from the NCBI ftp server (http://www.ncbi.nlm.nih.gov/ftp/genomes/). The translated ORF protein sequences of each genome were used as the database to search for sequences similar to each member of the query database using shell scripts compiled for running PSI-BLAST. A total of 10 iterations were performed. The output of the search was sparsed using a filter involving 50 amino acids overlap, 15% identity within the aligned region and an *e* value cut off of 0.01. The selected sequences were put through the domain identification procedures outlined earlier.

Annotated lectins

Proteins already annotated as lectins were also added to the list by including the sequences for which the header name consisted of keywords lectin, adhesin, agglutinin, carbohydrate-binding protein, sugar-binding protein, hyaluronic acid binding protein, chitin-binding protein, neuraminyllactose-binding protein, tenascin, agrin, lectin-like, and heparin-binding protein. Those sequences were also put through the fold recognition procedures.

Analysis

Sequence comparison of the protein sequences and foils was carried using MatGAT[30] and CLUSTALW.[31] Pymol[32] and Rasmol[33] were used for macromolecular visualization and binding site analysis.

RESULTS AND DISCUSSION

As mentioned in the introduction, the effort here was to identify archeal proteins which can be unambiguously assigned to a lectin family. The approach, therefore, was to bring to bear different search procedures on the problem. To start with, a total of 389,375 open reading frames (ORFs) belonging to 165 genomes spanning 78 distinct archeal species were searched using CDD. That resulted in the identification of 160 ORFs with homologues among well established lectin domains with known three-dimensional structure. Of these, only 145 sequences lent themselves to the formation of models with the expected lectin fold using PHYRE with good *e* value (<0.001) and high precision (>90%). Among the 145, only 45 had the expected binding site residues. A similar operation starting with SCOP resulted in the identification of 46 lectins which included the 45 identified using CDD. Thus, the two parallel domain identification procedures led to the designation of 46 archeal sequences as those containing lectins.

Sequence search through the genomes using the sequences of well established lectins with known three-dimensional structure, followed by domain identification procedures led to the designation of 44 sequences as those of lectins. These 44 formed a subset of the 46 picked up solely using domain identification procedures. The two proteins that were not picked up in the sequence search had *e* value higher than 0.01. A number of ORFs have been annotated as those of keywords related to lectins. All of them were put through the domain identification procedure. Only eight of them were unambiguously identified as lectins. They formed a small subset of the original 46. Thus, the critical criterion for identification was the ability of standard homology modelling programmes to construct an acceptable three-dimensional model using the given sequence on the basis of a lectin of known three-dimensional structure.

Distribution of lectin families in the archeal genomes

Among the 165 archeal genomes belonging to 78 different species examined here, a total of 46 lectin domains could be identified. These proteins, spanning 29 distinct archeal strains, belong to 12 families (Table I). Almost half of them (21 in number) have the 7-bladed β-propeller lectin fold which occur in 9 out of the 12 families. The next largest group (16) is made up of domains with the β-trefoil fold present only in haloarchea. Three ORFs from methanoarchea have two legume lectin domains each while one such domain occurs in a sequence in *Nanoarchaeum equitans* Kin4. Two proteins belonging to different families assume the C-type lectin fold. Two ORFs from the same family assume the β-prism I fold while another from a different family has the 5-bladed β-propeller lectin domain. The alignment of a typical archeal lectin sequence from each structural family and that of a typical well characterized lectin with the corresponding fold, is given in Supporting Information Figure S1. In a majority of cases, nonlectin domains have also been identified in ORFs containing lectin domains (Fig. 1). Many ORFs also contain long nonannotated stretches.

7-bladed β-propeller lectin fold

The single largest group of lectin domains identified in archea shows homology to the integrin α-*N*-terminal domain observed in the mushroom lectin from *Psathyrella velutina* (PVL).[34] PVL is the only lectin of known structure in this family. The 401-amino acid long lectin assumes a 7-bladed β-propeller fold [Fig. 2(a)]. Sugar binding sites in it are located in the upper part of the β-propeller, in spaces between consecutive blades. A hydrophobic patch, mainly made up of aromatic residues, interacts with each bound sugar. The sugar is also hydrogen bonded to two main chain NH groups, an asparaginyl side-chain and a tryptophan side chain. The structure of PVL contains two to three calcium ions. The calcium binding site involves a loop with consensus sequence D × T × D × [L/C] × D.

An archeal sequence was considered to be a putative 7-bladed β-propeller fold lectin if PHYRE yielded a model with at least six blades with appropriate topology. It was further required that the folded structures should posses residues corresponding to at least one sugar binding site and one calcium binding site of the type observed in PVL. The 21 archeal sequences with a minimum length of 360 residues identified using these criteria are listed in Table I. They belong to 16 different archeal species. Some genomes contain more than one 7-bladed β-propeller fold lectin domain. Identity among these 21 sequences ranges between 93.5 and 15.0%. They have sequence identities of 21.6–16.9% with respect to PVL.

Among the 21 ORFs involving the PVL fold, four contain a little <400 residues each and can accommodate only one PVL like domain [Fig. 1(a)]. Then there are eight ORFs, each of which is long enough to accommodate one or more domains in addition to a PVL-like domain. However, PHYRE did not model any other domain in the additional polypeptide stretches that are present. One or more other domains have been identified in the remaining 9 ORFs. The most notable among them are the PKD[35] domains and the YVTN-repeat[36] domains. These domains are known to be involved in facilitating survival by adhesion to surfaces.[36] PVL, the only 7-bladed β-propeller lectin of known structure, exhibits near 7-fold symmetry. However, sequence identity among pairs of blades varies between 60.2 and 23.8%. In the corresponding archeal lectins with well-identified seven blades, the sequence identity between pairs of blades in each domain is the highest in *Methanomethylovorans hollandica* DSM 15978, the maximum and minimum values being 98.1 and 21.2%, respectively. In *Archaeoglobus sulfaticallidus* PM70, in which the identity is the minimum, the maximum and minimum values are 30.8 and 4.3%. When the sequence of every blade in archeal domains is compared with those of the blades in PVL, the pairwise identity ranges between 33.3 and

Table I
Details of ORFs Identified as Lectin-Related from Archeal Genomes

Sno.	Genome name	Total ORFs	7-bladed β-propeller	β-trefoil fold	Legume lectin	β-prism I fold	C-type lectin	5-bladed β-propeller
1	*Archaeoglobus fulgidus* DSM 4304 uid57717	2420	11499528					
2	*Archaeoglobus sulfaticallidus* PM70 1 uid201033	2216	488601359					
3	*Candidatus korarchaeum* cryptofilum OPF8 uid58601	1603	170290194					
4	*Candidatus nitrosopumilus* AR2 uid176130	1974					407465564	
5	*Haloarcula hispanica* ATCC 33960 uid72475	3859	344212090 344210096					
6	*Haloarcula marismortui* ATCC 43049 uid57719	4243	55378026	55377121 55380166				
7	*Halogeometricum borinquense* DSM 11551 uid54919	3898	313122374 313126621					
8	*Halomicrobium mukohataei* DSM 12286 uid59107	3349	257387795 257386210	257387442				
9	*Halopiger xanaduensis* SH 6 uid68105	4221		336251901 336252094 336255156				
10	*Haloquadratum walsbyi* DSM 16790 uid58673.ls	2643					110667057	
11	*Halorhabdus tiamatea* SARL4B uid214082	3023		529078136				
12	*Halorhabdus utahensis* DSM 12940 uid59189	2998		257053030 257053515 257053516				
13	*Haloterrigena turkmenica* DSM 5511 uid43501	5113		284167153 284167158 284167313 284172542 284172598 284172606				
14	*Methanocaldococcus* FS406 22 uid42499	1816			289191745[a]			
15	*Methanocaldococcus jannaschii* DSM 2661 uid57713	1771			499173326[a]			
16	*Methanocaldococcus vulcanius* M7 uid41131	1742			261402542[a]			
17	*Methanococcus maripaludis* C7 uid58847	1788				150401960		
18	*Methanococcus voltae* A3 uid49529	1717				297619264		
19	*Methanohalobium evestigatum* Z 7303 uid49857	2254	298674578					
20	*Methanomethylovorans hollandica* DSM 15978	2556	435850334 435850887 435850888					
21	*Methanosaeta harundinacea* 6Ac uid81199	2371	386002078					
22	*Methanosarcina barkeri Fusaro* uid57715	3625	73670249					
23	*Methanosphaerula palustris* E1 9c uid59193	2655						219852796
24	*Nanoarchaeum equitans* Kin4 M uid58009	540			41615226			
25	*Salinarchaeum laminariae* Harcht Bsk1 uid207001	3013	510881252					
26	*Staphylothermus hellenicus* DSM 12710 uid45893	1599	297526457					
27	*Thermococcus* 4557 uid70841	2133	341581909					
28	*Thermococcus barophilus* MP uid54733	2265	315231213					
29	*Thermococcus litoralis* DSM 5473 uid82997	2516	530548018					

[a]Sequences with two legume lectin domains each.

6.4%. Pairwise sequence identity among all the blades belonging to the 21 lectins varies between 98.1 and 1.9%. However, those belonging to the same family tend to cluster together. Despite the comparatively low sequence identity of the archeal sequences with that of PVL, good models of the former could be constructed using the three-dimensional structure of PVL as the template (Fig. 2). However, the number of binding sites identified in the archeal lectins is invariably less than the six in PVL.

β-trefoil fold

The most extensively studied β-trefoil lectins are those that occur along with a glycosidase domain in type II RIPs of plant origin, typified by ricin.[37] However, this lectin domain occurs independently and along with other domains in a variety of organisms. PHYRE indicated homology of segments of archeal sequences with three fungal β-trefoil lectin domains, one each from the genomes of *Marasmius oreades* (MOA),[38] *Sclerotinia sclerotiorum* (SSA),[39] and *Polyporus squamosus* (PSA).[40] Homology is also indicated with one trefoil domain each from three bacteria, namely, *Clostridium thermocellum*,[41] *Streptomyces avermitlis*,[42] and *Streptomyces olivaceoviridis*.[43] Each of the fungal and bacterial domains exhibits near threefold symmetry in three-dimensional structure and is approximately 140 residues long (Fig. 3). The number of binding sites in them varies between one and

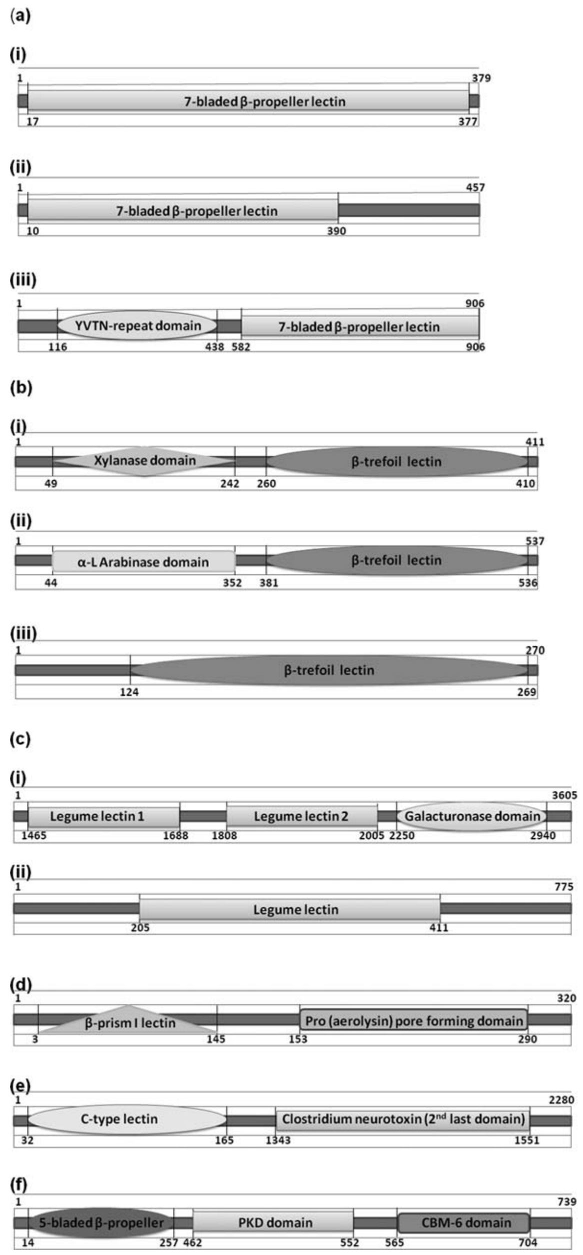

Figure 1

Domain organization of representative archeal proteins with lectin folds: (**a**) 7-bladed β-propeller lectin fold in (i) Hbor 36080 (313122374) from *Halogeometricum borinquense* DSM 11551, (ii) Hbor 18790 (313126621) from *Halogeometricum borinquense* DSM 11551, and (iii) Metho 0100 (435850334) from *Methanomethylovorans hollandica* DSM 15978; (**b**) β-trefoil lectin domain in (i) Endo-1,4-beta-xylanase (257053030) from *Halorhabdus utahensis* DSM 12940, (ii) Glycoside hydrolase family 43 protein (257053515) from *Halorhabdus utahensis* DSM 12940, and (iii) Hmuk 1387 (257387442) from *Halomicrobium mukohataei* DSM 12286; (**c**) Legume lectin domain in (i) Hypothetical protein (499173326) from *Methanocaldococcus jannaschii* and (ii) Hypothetical protein NEQ442 (41615226) from *Nanoarchaeum equitans* Kin4-M; (**d**) β-prism I fold lectin domain in MmarC7 0032 (150401960) from *Methanococcus maripaludis* C7; (**e**) C-type lectin domain in NSED_08560 (407465564) from *Candidatus Nitrosopumilus* sp. AR2, and (**f**) 5-bladed β-propeller fold in CBM-6 protein (219852796) from *Methanosphaerula palustris* E1-9c.

three. The binding sites in the lectins involve a D/N—X_1X_2Ar—N/H motif. D and N occur at the first and the last position in a majority of cases.

Archeal sequences with β-trefoil lectin fold containing at least one carbohydrate binding site were accepted as corresponding to lectin domains. Replacement of the aspartyl residue with a glutamyl residue at the binding site was also accepted. Threonine occurs at the last

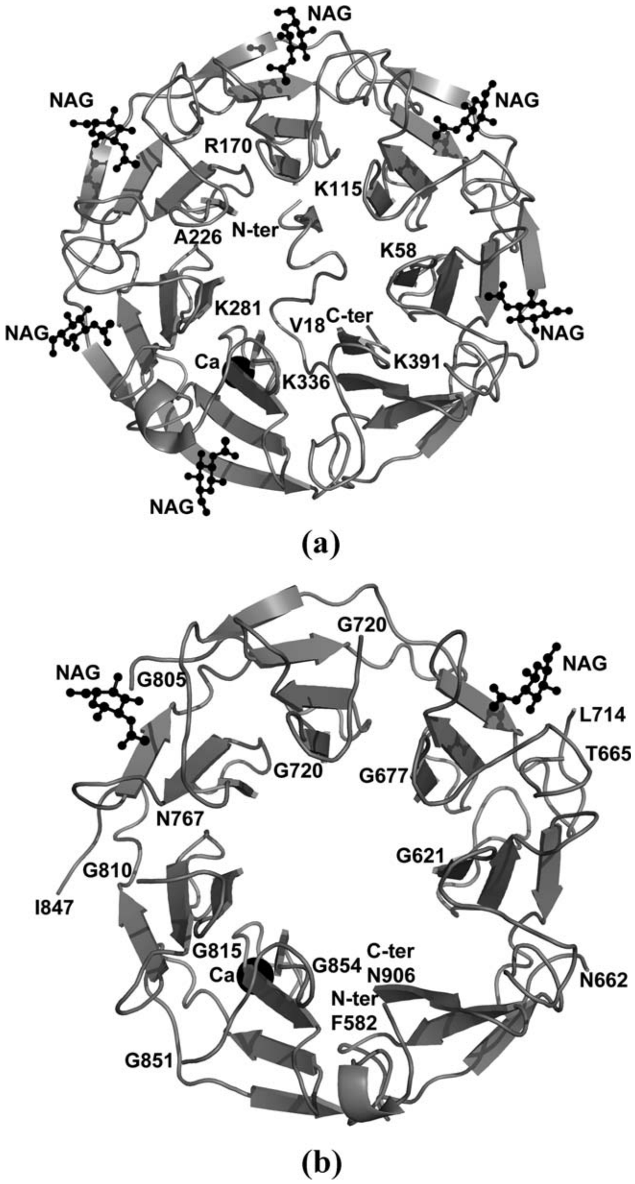

Figure 2

7-bladed β-propeller lectin fold: (**a**) three-dimensional structure of the 7-bladed β-propeller *Psathyrella velutina* lectin from *Lacrymaria velutina* in complex with N-acetyl glucosamine (PDB ID: 2C4D). (**b**) Homology model of the putative 7-bladed β-propeller lectin from *Methanomethylovorans hollandica* DSM 15978. The N-, C-terminal stretches and five strands at the periphery are missing in the model. The Ca^{2+} bound to the structure is shown as a van der Waals sphere in both the cases.

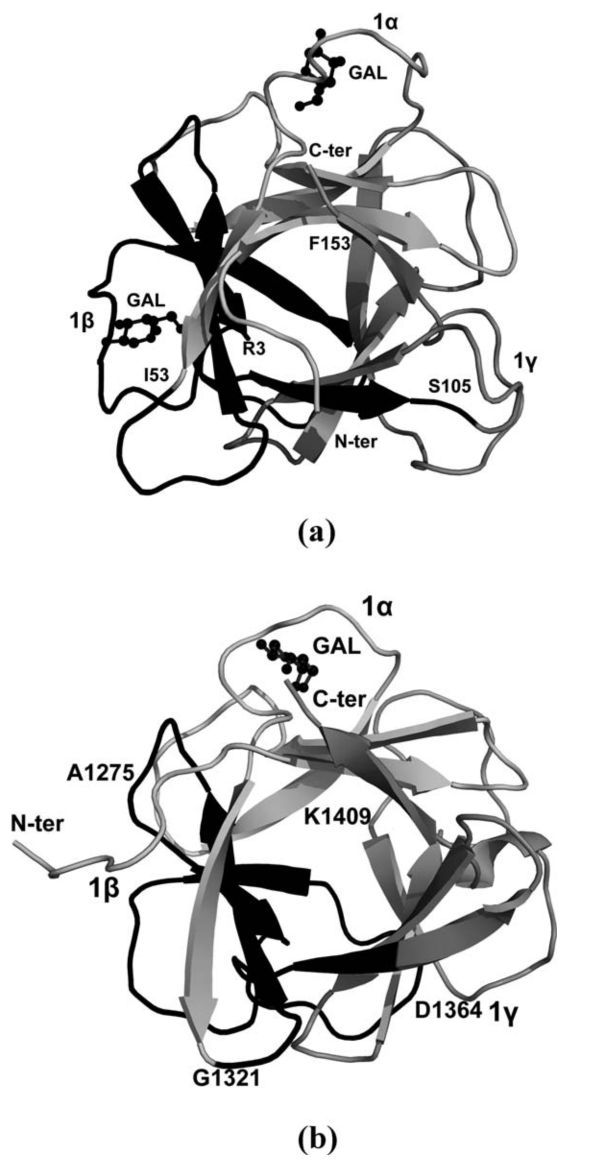

Figure 3

β-trefoil fold: (**a**) Three-dimensional structure of the β-trefoil domain of MOA lectin from *Marasmius oreades* bound to galactose (PDB ID: 2IHO). The three foils are labeled as 1α, 1β, and 1γ. (**b**) Homology model of the putative β-trefoil lectin from *Haloterrigena turkmenica* DSM 5511.

position of the motif in a couple of instances. The presence of T instead of N or H is compatible with the lectin–sugar interactions in the reference structures. Therefore, sites with T at this position were also accepted. On the basis of these criteria, 16 sequences containing the β-trefoil lectin was identified in six different archeal species belonging to two haloarchea, namely *Halobacteriaceae* and *Natrialbaceae* (Table I). The number of domains in the six genomes varies between 1 and 6. In 15 of the 16 sequences, the accompanying domains are of enzymes involved in modification of carbohydrates [Fig. 1(b)]. In all cases, the β-trefoil lectin domain is found in the C-terminal region. Even in the sequence in which only a lectin domain has been identified, that domain is preceded by a stretch of more than a hundred amino acid residues which could well correspond to a yet to be identified nonlectin domain. The sequence identity among the 16 lectin domains is between 86.0 and 20.1%. The sequence identity of archeal trefoils with the fungal trefoils ranges between 25.7 and 16.6%. That with the bacterial trefoils is between 37.4 and 16.7%. On an average, the sequence identity of the archeal trefoils with the bacterial trefoils is higher than that with the fungal trefoils. That with those in ricin is still lower, the maximum and the minimum values being 22.9 and 10.9%.

It has been suggested that the β-trefoil fold lectin domains originated through successive gene duplication, fusion, and divergent evolution of a primitive sugar binding motif.[44,45] In this context, it is interesting to examine the sequence identity among the three foils in each class of β-trefoil lectin domains. The average identity among pairs of three foils in the trefoil domain-1 in ricin, a representative of type II RIPs, is 15.1%. The corresponding value in domain-2 of ricin is 19.8%. The average identity between the three foils in the three fungal proteins ranges from 36.6 to 16.8%. The corresponding range in the three bacterial β-trefoils are 43.4 and 19.2%. The value in archeal trefoil lectin domains exhibit wide variation, the maximum being 63.1% in *Haloarcula marismortui* ATCC 43049 and minimum being 12.3% in *Halomicrobium mukohataei* DSM 12286. Each of the three foils carries a sugar binding site in five of the archeal domains. The average identity between pairs of foils in them ranges between 62.9 and 48.3% with an average of 55.4%. The range and average in the three domains that carry two binding sites each are 48.0–27.2 and 39.2%, respectively, while it is 53.1–12.3 and 28.8%, respectively, in trefoils which carry only one binding site. Thus, there appears to be a weak correlation between the number of binding sites in the archeal trefoil domain and the sequence identity among the foils.

Legume lectin fold

The third most populous archeal lectin domains are those which exhibit homology with legume lectins. [Fig. 1(c)]. Those from leguminous plants undoubtedly constitute the most thoroughly studied structural family among lectins.[1,22] They occur as dimers or dimers of dimers. Each subunit is made up of at least a six stranded β- and a seven stranded β-sheet [Fig. 4(a)]. The carbohydrate binding site, made up of specific loops, in all cases contain an aspartyl residue and an asparaginyl residue separated by about 40 residues along the sequence. Legume lectins could be Gal/GalNAc specific or Man/Glc specific. In Gal/GalNAc specific lectins, an aromatic residue located close to the asparaginyl residue

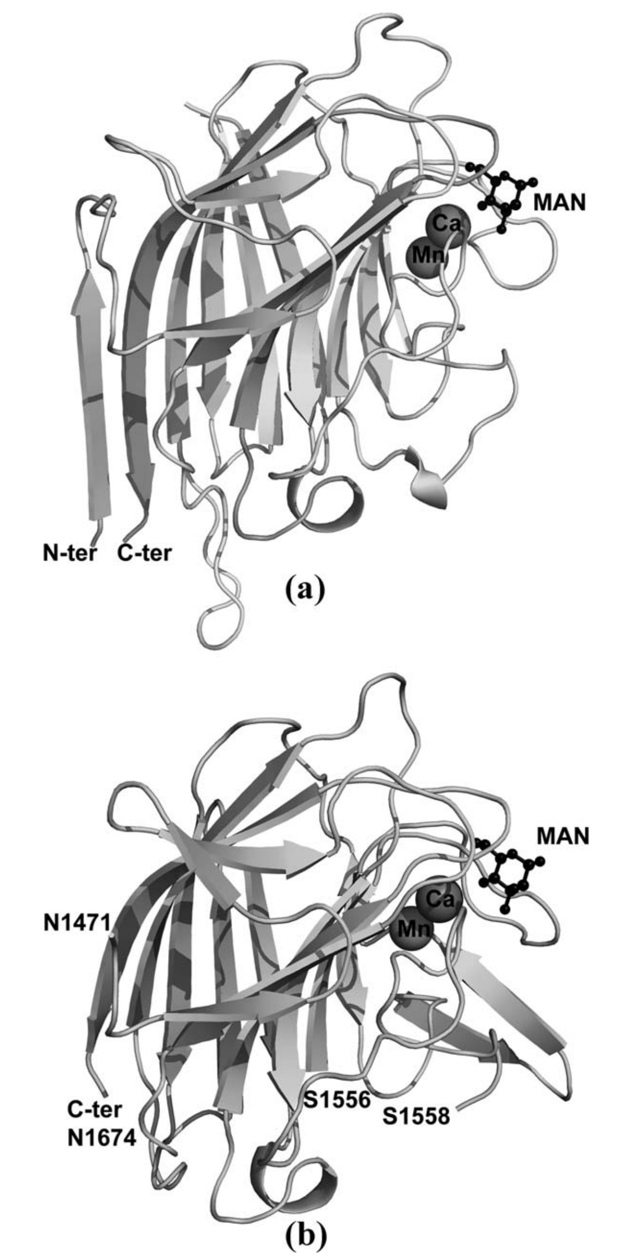

Figure 4

Legume lectin fold: (**a**) Three-dimensional structure of pea lectin (PSL) bound to mannose (PDB ID: 1RIN). (b) Homology model of the putative legume lectin domain 1 from *Methanocaldococcus vulcanius* M7. One strand and one loop are missing in the model. The manganese and calcium ions are shown as a van der Waals spheres in both the cases.

stacks against the bound galactose. This stacking interaction is not obligatory in Man/Glc legume lectins. Legume lectins invariably contain a calcium and a manganese ion. In addition to the asparaginyl residue which interacts with sugar, two aspartyl, and one glutamyl residues are involved in binding the metal ions.

Seven legume lectin domains belonging to four archeal species could be unambiguously identified (Table I). All but one contain the aromatic residue appropriate for stacking on the galactose ring. Other sugar and metal binding residues were present in most of the domains. Exceptions occur in two cases, where an aspartyl residue is replaced by a glutamyl residue in one case while one metal binding residue is missing in the other. The sequence identity ranges between 91.0 and 21.3% among the lectin regions of the archeal proteins. However, when the two domains occur in the same sequence, the sequence identity between them ranges between 91.0 and 79.2%. The identity of the sequences with respect to that of pea lectin[46] is between 25.7 and 17.5%.

Genomes of three organisms from *Methanocaldococcus* genus have one ORF containing two legume lectin domains each. Each ORF is about 2200 amino acids long and have two legume lectin-like domains accounting for only <500 residues. A galacturonase domain, downstream to the lectin domains, could also be identified in each case. A legume lectin like domain occurs in a 775 residue long ORF in *Nanoarchaeum equitans* Kin4 as well. Homology models for all the legume–lectin domains based on the structure of pea lectin could be constructed using PHYRE [Fig. 4(b)].

β-prism I lectin fold

β-prism I fold was first characterized as a lectin fold through the structure analysis of the galactose specific tetrameric jacalin,[47] one of the two lectins present in the seeds of jackfruit (*Artocarpus integrifolia*). The structure of the other lectin, artocarpin,[48] was also subsequently determined. Artocarpin is again tetrameric, but mannose specific, and is structurally very similar to jacalin. The β-prism I fold involves three Greek keys arranged around an approximate three-fold axis [Fig. 5(a)]. Jacalin and artocarpin, between them, define the characteristics of the primary binding sites of galactose and mannose specific β-prism I fold lectins. The structures of a number of β-prism I fold lectins are currently available.[49] All of them are dimeric or tetrameric. The carbohydrate binding site(s) at one end of the prism is characterized by a G. . .GXXXD motif. Thus the only side-chain involved in interactions at the primary binding site is that of an aspartyl residue. In addition, an aromatic residue stacks against the sugar ring in the galactose specific lectins.

Homologues of β-prism I fold lectins could be found in *Methanococcus voltae* A3 and *Methanococcus maripaludis* C7. In the former, a single 145 residue long gene product constitutes the lectin. In the later, the sequence contains a lectin domain and a 137 amino acid segment of a very well characterized 380 amino acid long pore forming lobe of cytolytic pore-forming toxin (Pro)aerolysin exported by *Aeromonas hydrophila*, a Gram-negative bacterium associated with diarrhoeal diseases and deep wound infections [Fig. 1(d)]. The sequence identity

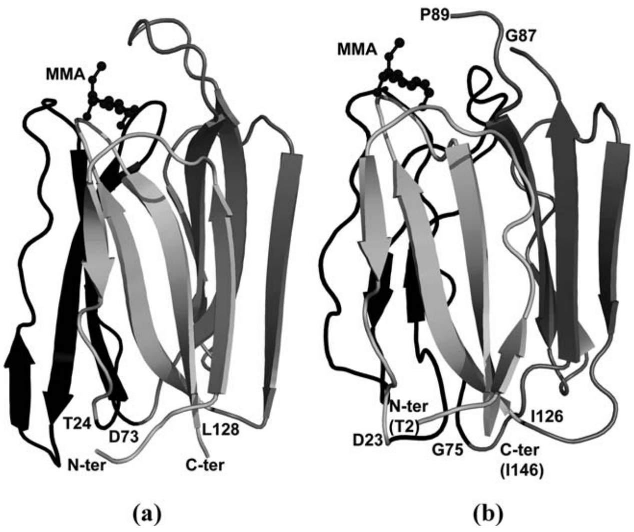

Figure 5

β-prism I fold lectin. (**a**) Three-dimensional structure of the monomeric unit of artocarpin from *Artocarpus integer* bound to Me-α-mannose (PDB ID: 1J4U). (**b**) Homology model of the putative β-prism I lectin domain from *Methanococcus maripaludis* C7.

between the two lectin domains is 29.3%. They have sequence identities of 18.3 and 21.9% with artocarpin.

Homology models of the two archeal domains could be readily constructed using banana lectin[50] as the template. The β-prism I fold consists of three Greek keys which form three sides of a nearly threefold symmetric prism. Two of the three keys carry primary sugar binding sites in banana lectin. However, only one of the keys has a binding site in the archeal lectin domains, as in the case of artocarpin. It has been suggested in the case of β-prism fold lectins also that the fold could have resulted from successive gene duplication, fusion and divergent evolution of a primitive Greek key based carbohydrate binding motif.[47,51] It has been further suggested that there is a correlation between the average sequence identity among the three Greek keys and the number of binding sites. For instance, griffithsin,[52] which carries three binding sites, has an average identity of 27.7%. The average identity in banana lectin which carries two binding sites is 20.6%. The corresponding values in artocarpin with one binding site is 16.7%. The value of 17.6% in the lectin domain in *Methanococcus voltae* A3 is close to that in artocarpin while that in the domain in *Methanococcus maripaludis* C7 is 11.7%, which is well below that in artocarpin. Thus, the identification of one binding site in the archeal lectin is in consonance with the general observation on the correlation involving sequence identity.

C-type lectin fold

C-type lectin domains, approximately 135 residues in length, which often occur in tandem with other domains, have been extensively characterized in animals. Two archeal domains exhibit structural homology to langerin,[53] a well-known C-type lectin. Like other mannose-binding C-type lectins, the sugar binding sites of C-type lectins involves an E-X-N/R motif followed by a glutamyl residue after about five residues and then by a ND doublet after about a dozen residues. However, unlike other C-type lectins, which contain two calcium ions, langerin binds only one calcium ion [Fig. 6(a)].

A C-type lectin domain has been identified in a long, 9159 residue long ORF in the genome of *Haloquadratum wbyi* DSM 16790. No other domain could be identified in the ORF. Another such domain was located in a 2280 residue ORF in *Haloarcula hispanica* ATCC 33960 along with a *Clostridium* neurotoxin domain. Both the lectin domains could be successfully modeled with the structure of mannose specific langerin as the template. The models are similar to the structure of langerin except in the loops. The sequence identities of the two archeal domains with respect to langerin are 23.6 and 26.9%, respectively, and that between them is 29.8%. The sugar binding motif in the domain from *Haloquadratum walsbyi* DSM 16790 is E-P-N—E-ND while that from *Candidatus nitrosopumilus* sp. AR2 is E-P-N—E-ID. Unlike in C-type lectins from animal sources, archeal domains do not contain disulphide bridges. The lectin domain in *Haloquadratum walsbyi* DSM 16790 occurs in a long sequence in which no other domains could be identified. The *Candidatus nitrosopumilus* sp. AR2 sequence is

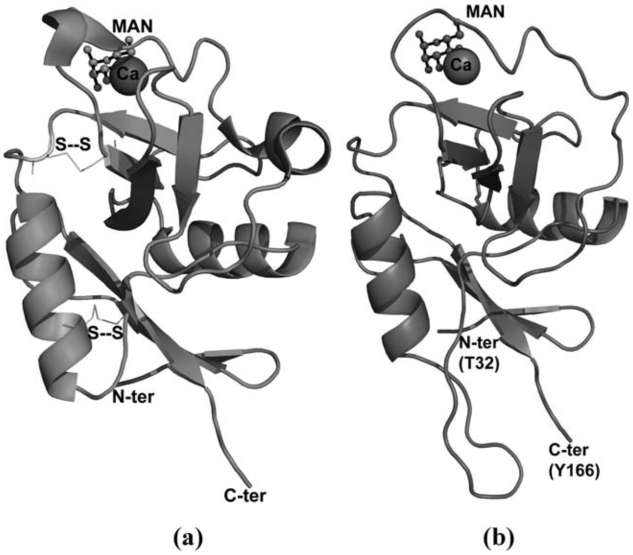

Figure 6

C-type lectin fold: (**a**) Three-dimensional structure of langerin from *Homo sapiens* bound to α-D-mannose (PDB ID: 3P5E). (**b**) Homology model of the putative C-type lectin domain from *C. nitrosopumilus* AR2. The calcium bound to the structure is shown as a van der Waals sphere in both the cases.

accompanied by the clostridium neurotoxin domain [Fig. 1(e)].

5-bladed β-propeller lectin domain

The GalNac/GlcNAc specific, 236 residue long tachylectin-2[54] isolated from the large granules of Japanese horseshoe crab has a 5-bladed β-propeller fold. Each blade has a four stranded antiparallel β-sheet with W-like topology. The lectin has five virtually identical sugar binding sites, one on each β-sheet. Each binding site contains two tryptophan residues. The side-chain of one is hydrogen bonded to the sugar while that of the other has a hydrophobic interaction with the sugar. In addition, five main chain NH or CO groups interact with the sugar.

A sequence from *Methanosphaerula palustris* E1-9c contains a 232-amino acid residue long stretch homologous to tachylectin-2 with a sequence identity of 14.2%. The 739 amino acid sequence contains, in addition, a cellulose-binding β-sandwich domain, a carbohydrate binding module 6[55] and a PKD domain [Fig. 1(f)]. The PHYRE model of the lectin domain is consistent with five main chain interactions with the sugar. The Trp residue, whose side-chain hydrogen bonds sugar in tachylectin-2, is a valine in the archeal lectin domain. However, the Trp residue involved in the hydrophobic interaction with the sugar is present in the archeal domain as well. Modelling of the archeal lectin on the basis of tachylectin was substantially but not wholly successful. A part of the fifth blade could not be modeled using PHYRE.

The near 5-folded symmetry of tachylectin as well as the archeal lectin is reflected their sequences as well. The sequence identity among pairs of blades in tachylectin varies between 66 and 46.8%. The corresponding values in the case of the archeal lectin are 87.5 and 38.3%. The sequence identity between the blades in tachylectin on the one hand and the archeal lectin on the other, is much lower, the maximum and minimum values being 23.4 and 8.5%.

CONCLUSION

The work presented here represents the first comprehensive attempt to delineate lectin domains in archea. The results demonstrate the presence of homologues of six structural classes of eukaryotic and bacterial lectins in archea. This clearly indicates the ancient origin of lectins. Many of the lectin folds must have evolved into functional proteins before the three domains of life diverged. However, the number of archeal species in which lectins could be identified is disconcertingly low. This could well be because of the rigorous criteria employed in the present search. It is also possible that hitherto unidentified lectin folds exist in archea. This is an area which merits further exploration.

ACKNOWLEDGMENTS

The authors thank the Science and Engineering Research Board, India for financial support. Part of the computations was performed at the Interactive Graphics Facility supported by the Department of Biotechnology, Government of India. E. S. was a UGC Kothari Postdoctoral Fellow. M. V. is the Albert Einstein Professor of the Indian National Science Academy.

REFERENCES

1. Pérez S, Sarkar A, Rivet A, Breton C, Imberty A. Glyco3D: a portal for structural glycosciences. Methods Mol Biol 2015;1273:241–258.
2. Vijayan M, Chandra NR. Lectins. Curr Opin Struct Biol 1999;9: 707–714.
3. Feizi T. Carbohydrate-mediated recognition systems in innate immunity. Immunol Rev 2000;173:79–88.
4. Sharon N, Lis H. History of lectins: from hemagglutinins to biological recognition molecules. Glycobiology 2004;14:53R–62R.
5. Wong JH, Wong CC, Ng TB. Purification and characterization of a galactose-specific lectin with mitogenic activity from pinto beans. Biochim Biophys Acta 2006;1760:808–813.
6. Ngai PHK, Ng TB. A lectin with antifungal and mitogenic activities from red cluster pepper (*Capsicum frutescens*) seeds. Appl Microbiol Biotechnol 2007;74:366–371.
7. Zhang GQ, Sun J, Wang HX, Ng TB. A novel lectin with antiproliferative activity from the medicinal mushroom *Pholiota adiposa.* Acta Biochim Pol 2009;56:415–421.
8. Dam TK, Brewer CF. Lectins as pattern recognition molecules: the effects of epitope density in innate immunity. Glycobiology 2010;20: 270–279.
9. Weis WI, Drickamer K. Structural basis of lectin-carbohydrate recognition. Annu Rev Biochem 1996;65:441–473.
10. Loris R. Principles of structures of animal and plant lectins. Biochim Biophys Acta 2002;1572:198–208.
11. Prabu MM, Suguna K, Vijayan M. Variability in quaternary association of proteins with the same tertiary fold: a case study and rationalization involving legume lectins. Proteins 1999;35:58–69.
12. Sharma A, Vijayan M. Quaternary association in beta-prism I fold plant lectins: insights from X-ray crystallography, modelling and molecular dynamics. J Biosci 2011;36:793–808.
13. Drickamer K. Multiplicity of lectin–carbohydrate interactions. Nat Struct Biol 1995;2:437–439.
14. Jeyaprakash AA, Srivastav A, Surolia A, Vijayan M. 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 2004;338:757–770.
15. Sixma TK, Pronk SE, Kalk KH, Wartna ES, van Zanten BA, Witholt B, Hol WG. Crystal structure of a cholera toxin-related heat-labile enterotoxin from *E. coli.* Nature 1991;351:371–377.
16. Jones CH, Pinkner JS, Roth R, Heuser J, Nicholes AV, Abraham SN, Hultgren SJ. Fim H adhesin of type 1 pili is assembled into a fibrillar tip structure in the *Enterobacteriaceae.* Proc Natl Acad Sci USA 1995;14:2081–2085.
17. Zhang RG, Scott DL, Westbrook ML, Nance S, Spangler BD, Shipley GG, Westbrook EM. The three-dimensional crystal structure of cholera toxin. J Mol Biol 1995;25:563–573.
18. Swaminathan S, Eswaramoorthy S. Structural analysis of the catalytic and binding sites of *Clostridium botulinum* neurotoxin B. Nat Struct Biol 2000;7:693–699.

19. Mitchell E, Houles C, Sudakevitz D, Wimmerova M, Gautier C, Pérez S, Wu AM, Gilboa-Garber N, Imberty A. Structural basis for oligosaccharide-mediated adhesion of *Pseudomonas aeruginosa* in the lungs of cystic fibrosis patients. Nat Struct Biol 2002;9:918–921.
20. Wiley DC, Wilson IA, Skehel JJ. Structural identification of the antibody-binding sites of Hong Kong influenza haemagglutinin and their involvement in antigenic variation. Nature 1981;289:373–378.
21. Patra D, Mishra P, Surolia A, Vijayan M. Structure, interactions and evolutionary implications of a domain-swapped lectin dimer from *Mycobacterium smegmatis*. Glycobiology 2014;24:956–965.
22. Abhinav KV, Vijayan M. Structural diversity and ligand specificity of lectins. The Bangalore effort. Pure Appl Chem 2014;86:1335–1355.
23. Abhinav KV, Sharma A, Vijayan M. Identification of mycobacterial lectins from genomic data. Proteins 2013;81:644–657.
24. Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, Geer RC, He J, Gwadz M, Hurwitz DI, Lanczycki CJ, Lu F, Marchler GH, Song JS, Thanki N, Wang Z, Yamashita RA, Zhang D, Zheng C, Bryant SH. CDD: NCBI's conserved domain database. Nucleic Acids Res 2015;43(Database issue):D222–D226.
25. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc 2015;10:845–858.
26. Wilson D, Pethica R, Zhou Y, Talbot C, Vogel C, Madera M, Chothia C, Gough J. SUPERFAMILY—sophisticated comparative genomics, data mining, visualization and phylogeny. Nucleic Acids Res 2009;37(Database issue):D380–D386.
27. Huang Y, Niu B, Gao Y Fu L. W, Li. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics 2010;26:680–682.
28. Altschul SF. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl Acids Res 1997;25:3389–3402.
29. Schaffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, Koonin EV, Altschul SF. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucl Acids Res 2001;29:2994–3005.
30. Campanella JJ, Bitincka L Smalley J., MatGAT. An application that generates similarity/identity matrices using protein or DNA sequences. BMC Bioinform 2003;4:1471–2105.
31. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. ClustalW ClustalX version 2.0. Bioinformatics 2007;23:2947–2948.
32. The PyMOL Molecular Graphics System, Version 1.7.4 Schrödinger, LLC.
33. Sayle RA, Milner-White EJ. RASMOL: biomolecular graphics for all. Trends Biochem Sci 1995;20:374.
34. Cioci G, Mitchell EP, Chazalet V, Debray H, Oscarson S, Lahmann M, Gautier C, Breton C, Perez S, Imberty A. β propeller crystal structure of *Psathyrella velutina* lectin: an integrin-like fungal protein interacting with monosaccharides and calcium. J Mol Biol 2006;14:1575–1591.
35. Bycroft M, Bateman A, Clarke J, Hamill SJ, Sandford R, Thomas RL, Chothia C. The structure of a PKD domain from polycystin-1: implications for polycystic kidney disease. Embo J 1999;18:297–305.
36. Jing H, Takagi J, Liu JH, Lindgren S, Zhang RG, Joachimiak A, Wang JH, Springer TA. Archaeal surface layer proteins contain beta propeller, PKD, and beta helix domains and are related to metazoan cell surface proteins. Structure 2002;10:1453–1464.
37. Rutenber E, Katzin BJ, Ernst S, Collins EJ, Mlsna D, Ready MP, Robertus JD. Crystallographic refinement of ricin to 2.5 Å. Proteins 1991;10:240–250.
38. Grahn E, Askarieh G, Holmner A, Tateno H, Winter HC, Goldstein IJ, Krengel U. Crystal structure of the *Marasmius oreades* mushroom lectin in complex with a xenotransplantation epitope. J Mol Biol 2007;369:710–721.
39. Sulzenbacher G, Roig-Zamboni V, Peumans WJ, Rougé P, Van Damme EJ, Bourne Y. Crystal structure of the GalNAc/Gal-specific agglutinin from the phytopathogenic ascomycete *Sclerotinia sclerotiorum* reveals novel adaptation of a β-trefoil domain. J Mol Biol 2010;400:715–723.
40. Kadirvelraj R, Grant OC, Goldstein IJ, Winter HC, Tateno H, Fadda E, Woods RJ. Structure and binding analysis of *Polyporus squamosus* lectin in complex with the Neu5Ac α-(2-6) Gal β(1-4)GlcNAc human-type influenza receptor. Glycobiology 2011;21:973–984.
41. Jiang D, Fan J, Wang X, Zhao Y, Huang B, Liu J, Zhang XC. Crystal structure of 1,3Gal43A, an exo-β-1,3-galactanase from *Clostridium thermocellum*. J Struct Biol 2012;180:447–457.
42. Ichinose H, Fujimoto Z, Honda M, Harazono K, Nishimoto Y, Uzura A, Kaneko SA. βl-Arabinopyranosidase from *Streptomyces avermitilis* is a novel member of glycoside hydrolase family 27. J Biol Chem 2009;284:25097–25106.
43. 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;8:65–78.
44. Ready MP, Brown DT, Robertus JD. Extracellular localization of pokeweed antiviral protein. Proc Natl Acad Sci USA 1986;83:5053–5056.
45. Sharma A, Pohlentz G, Bobbili KB, Jeyaprakash AA, Chandran T, Mormann M, Swamy MJ, Vijayan M. The sequence and structure of snake gourd (*Trichosanthes anguina*) seed lectin, a three-chain nontoxic homologue of type II RIPs. Acta Crystallogr D Biol Crystallogr 2013;69:1493–1503.
46. Rini JM, Hardman KD, Einspahr H, Suddath FL, Carver JP. X-ray crystal structure of a pea lectin-trimannoside complex at 2.6 Å resolution. J Biol Chem 1993;268:10126−10132.
47. 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 β-prism fold. Nat Struct Biol 1996;3:596–603.
48. Pratap JV, Jeyaprakash AA, Rani PG, Sekar K, Surolia A, Vijayan M. 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 2002;317:237–247.
49. Chandran T, Sharma A, Vijayan M. Generation of ligand specificity and modes of oligomerization in β-prism I fold lectins. Adv Protein Chem Struct Biol 2013;92:135–178.
50. Meagher JL, Winter HC, Ezell P, Goldstein IJ, Stuckey JA. Crystal structure of banana lectin reveals a novel second sugar binding site. Glycobiology 2005;15:1033–1042.
51. Sharma A, Chandran D, Singh DD, Vijayan M. Multiplicity of carbohydrate-binding sites in beta-prism fold lectins: occurrence and possible evolutionary implications. J Biosci 2007;32:1089–1110.
52. Ziółkowska NE, O'Keefe BR, Mori T, Zhu C, Giomarelli B, Vojdani F, Palmer KE, McMahon JB, Wlodawer A. Domain-swapped structure of the potent antiviral protein griffithsin and its mode of carbohydrate binding. Structure 2006;14:1127–1135.
53. Feinberg H, Taylor ME, Razi N, McBride R, Knirel YA, Graham SA, Drickamer K, Weis WI. Structural basis for langerin recognition of diverse pathogen and mammalian glycans through a single binding site. J Mol Biol 2011;405:1027–1039.
54. Beisel HG, Kawabata S, Iwanaga S, Huber R, Bode W. Tachylectin-2: crystal structure of a specific GlcNAc/GalNAc-binding lectin involved in the innate immunity host defense of the Japanese horseshoe crab *Tachypleus tridentatus*. Embo J 1999;18:2313–2322.
55. Boraston AB, Bolam DN, Gilbert HJ, Davies GJ. Carbohydrate-binding modules: fine-tuning polysaccharide recognition. Biochem J 2004;382:769–781.