PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

Identification of mycobacterial lectins from genomic data

K. V. Abhinav, Alok Sharma and M. Vijayan*

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

ABSTRACT

Sixty-four sequences containing lectin domains with homologs of known three-dimensional structure were identified through a search of mycobacterial genomes. They appear to belong to the β-prism II, the C-type, the *Microcystis virdis* (MV), and the β-trefoil lectin folds. The first three always occur in conjunction with the LysM, the PI-PLC, and the β-grasp domains, respectively while mycobacterial β-trefoil lectins are unaccompanied by any other domain. Thirty heparin binding hemagglutinins (HBHA), already annotated, have also been included in the study although they have no homologs of known three-dimensional structure. The biological role of HBHA has been well characterized. A comparison between the sequences of the lectin from pathogenic and nonpathogenic mycobacteria provides insights into the carbohydrate binding region of the molecule, but the structure of the molecule is yet to be determined. A reasonable picture of the structural features of other mycobacterial proteins containing one or the other of the four lectin domains can be gleaned through the examination of homologs proteins, although the structure of none of them is available. Their biological role is also yet to be elucidated. The work presented here is among the first steps towards exploring the almost unexplored area of the structural biology of mycobacterial lectins.

Proteins 2012; 00:000–000.

Key words: host-pathogen interactions; homologs lectins; microbial lectins; hemagglutinin; domain identification.

INTRODUCTION

Lectins are carbohydrate-binding proteins, which exert their biological influence through the ability to recognize specific sugar structures.[1–6] They were originally isolated from plants and their best known property was the ability to agglutinate red blood cells. They used to be therefore referred to as phytohemagglutinins. Subsequently, they were found in all forms of life[6] with involvement in a variety of biological processes such as cell–cell interactions, innate immunity, mitogenesis, serum glycoprotein turnover.[7–9] With the increasing realization of the importance of protein-carbohydrate interactions in biological recognitive processes, investigations on lectins gathered momentum. Until very recently, our own effort in this area has been concerned with the structural biology of plant lectins.[2,10–19] We are now in the process of initiating work on microbial, particularly mycobacterial, lectins.[20,21]

Microbial lectins have so far received less attention than plant and animal lectins, although there have been very well known studies on some of them including influenza virus agglutinin[22,23] and enterotoxins.[24–27] Their importance in interactions with hosts, particularly host-pathogen interactions[28–30] is being increasingly recognized. Consequently, structural studies on them have also picked up momentum. However, studies of lectins from mycobacteria have been few and far between. In this context, we have extended our long range program on the structural biology of mycobacterial proteins[31–39] to include lectins as well. One of the lectins, identified on the basis of a bioinformatics search of *M. tuberculosis* H37Rv genome,[40] has been cloned, expressed and crystallized.[20] Also cloned, expressed and crystallized is another lectin from *M. smegmatis*.[21] Here we present a comprehensive thorough search of lectins of all the mycobacterial genomes of fully or partially known sequence. This search resulted in the identification of 94 lectins including 30 heparin binding hemagglutinins (HBHA),[41] which have no homologs with known three-dimensional structure.

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

Grant sponsors: the Department of Science and Technology, Government of India, Homi Bhabha Professor, CSIR Junior Research Fellow

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

Received 25 August 2012; Revised 16 November 2012; Accepted 17 November 2012
Published online 24 November 2012 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.24219

Table I
Lectin Domains Identified from Fully Sequenced Mycobacterial Genomes

Organism	Number of ORFs	RefSeq	Date of last modification	Gene-ids of ORFS identified with lectin domains				
				β-prism II lectins	C-type lectins	β-trefoil lectins	MVL lectin	HBHA
Mycobacterium avium 104	5120	NC_008595.1	02/11/11	0	0	0	118462906, 118462755	118463469
Mycobacterium avium paratuberculosis K-10	4350	NC_002944.2	10/22/10	0	0	0	41406213, 41407956	0
Mycobacterium bovis AF2122/97	3918	NC_002945.3	10/22/10	0	0	31792613	0	31791655
Mycobacterium bovis BCG str. Pasteur 1173P2	3949	NC_008769.1	12/14/10	0	0	121637349	0	121636391
Mycobacterium bovis BCG str. Tokyo 172	3944	NC_012207.1	10/22/10	0	0	224989824	0	224988863
Mycobacterium bovis BCG str. Mexico 86889	3952	NC_016804	29/2/12	0	0	378771183	0	378770225
Mycobacterium abscessus ATCC 19977	4941	NC_010397.1	01/06/11	169629459	0	0	169628098	169631161
Mycobacterium gilvum PYR-GCK	5579	NC_009338.1	10/22/10	0	0	0	145220609	0
Mycobacterium leprae Br4923	1604	NC_011896.1	10/22/10	0	0	0	0	221230806
Mycobacterium leprae TN	1605	NC_002677.1	02/13/11	0	0	0	0	15828329
Mycobacterium marinum M	5452	NC_010612.1	10/22/10	183983753	183983058	183983958	183980194, 183983846	183980824
Mycobacterium smegmatis str. MC2 155	6716	NC_008596.1	03/09/11	118468679	0	0	0	118468465
Mycobacterium sp. JLS	5739	NC_009077.1	10/22/10	0	0	0	0	0
Mycobacterium sp. KMS	5975	NC_008705.1	10/22/10	0	0	0	0	0
Mycobacterium sp. MCS	5615	NC_008146.1	11/22/10	0	0	0	0	0
*Mycobacterium_vanbaalenii*_PYR	**5979**	NC_008726.1	11/08/11	0	0	0	120401981	0
Mycobacterium ulcerans Agy99	4241	NC_008611.1	02/14/11	118618981	0	118619073	118619042, 118619962	118619638
Mycobacterium tuberculosis F11	3941	NC_009565.1	10/22/10	0	148823290	148822639	0	148821674
Mycobacterium tuberculosis H37Ra	4034	NC_009525.1	10/22/10	0	148661889	148661210	0	148660242
Mycobacterium tuberculosis H37Rv	3988	NC_000962.2	02/13/11	0	15609212	15608557	0	15607616
Mycobacterium tuberculosis KZN 1435	4059	NC_012943.1	10/22/10	0	253798868	253799531	0	253797403
Mycobacterium tuberculosis KZN 4207	3995	NC_016768	14/02/12	0	297634652	297633975	0	297632960
Mycobacterium tuberculosis CDC1551	4189	NC_002755.2	02/13/11	0	15841564	15840876	0	0
Mycobacterium africanum GM041182	3830	NC_015758.1	09/16/11	0	0	339631486	0	339630545
Mycobacterium canettii	3861	NC_015848.1	07/20/11	0	340627086	340626433	0	340625501
Mycobacterium sp. JDM601	4346	NC_015576.1	10/12/11	333990551	0	0	333988773	333989103
Mycobacterium sp. Spyr1	5349	NC_014814.1	05/12/11	0	0	0	315442443	0
Mycobacterium intracellulerae MOTT-02	5149	NC_016947	10/03/12	0	0	0	379752126, 379754394	379756580
Mycobacterium intracellulerae MOTT-64	5249	NC_016948	10/03/12	0	0	0	379759549, 379761671	379764109
Mycobacterium rhodesiae	6147	NC_016604	10/02/12	0	0	0	0	354587708

The remaining 64, of which 58 have not been annotated so far, have homologs with β-prism-II fold, C-type lectin fold, β-trefoil fold, and MVL fold. The identification of mycobacterial lectins, especially from pathogenic mycobacteria, presented here would hopefully help in future studies on host pathogen interactions.

METHODS

Retrieval of whole genome sequences of mycobacteria

Complete genome sequence information including the whole genome and translated ORF protein sequences from a total of 30 different completed genome projects spanning 20 distinct mycobacterial species were retrieved from NCBI ftp server (http://www.ncbi.nlm.nih.gov/Ftp/genomes/) on 18 March 2012 (Table I).

Domain identification in mycobacterial sequences

Domain definition of the sequences corresponding to all ORFs in the mycobacterial genomes was generated using the Conserved Domain Database (CDD) web server at NCBI[42] (http:// www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml). Only those sequences in which domains were designated as 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 by CDD were shortlisted. Among these, sequences, which have well established lectin homologs with known three-dimensional structure were chosen for further analysis. The sequences shortlisted after CDD analysis were also examined using the template based homology modeling PHYRE web server (http://www.sbg.bio.ic.ac.uk/phyre2), which searches for the profiles of the SCOP library of proteins in the query to assign them specific fold.[43] As a further check, models from the sequences of the accepted domains were also constructed using the method involving secondary structure enhanced profile–profile threading alignment (PPA) and iterative implementation of threading assembly (TASSER)[44] (http://zhanglab.ccmb.med. umich.edu/I-TASSER/). The results obtained by employing CDD, PHYRE, and I-TASSER were critically manually examined for the identification of folds.

The residues involved in carbohydrate-binding in the known homologs of the chosen mycobacterial lectins have already been identified by several authors using crystal structures of the sugar complexes of the relevant lectins. Only those sequences with substantial presence of these residues were accepted for further analysis. CDD and PHYRE were also used to identify other, nonlectin domains, which occur in tandem with lectin domains in the chosen mycobacterial sequences.

Sequence based search

In addition to the approach utilizing fold recognition explained above, a sequence based search for mycobacterial lectins was also conducted to ensure the completion of the effort.

Construction of the query database

Sequences of an initial list of query proteins were retrieved by searching for the same keywords used in domain identification, in Swiss Prot[45] in the ExPASy proteomics server. In addition, protein sequences of well annotated lectins with known crystal structures in PDB were also retrieved from the comprehensive lectin structural database at http://www.cermav.cnrs.fr/lectines/. The sequences thus retrieved were merged into a combined database, which was then made nonredundant by clustering the sequences with more that 90% sequence identity using a standalone version of CD-HIT.[46]

Sequence based identification followed by fold recognition

The identification of homologs of each of the sequence query in the mycobacterial genomes was carried out by using PSI-BLAST[47,48] installed locally employing the BLAST 2.2.12 package downloaded from the NCBI ftp server (http://www.ncbi.nlm.nih.gov/Ftp/genomes/). The translated ORF protein sequence file of each genome was used as the database against which each protein sequence in the query database was searched for similar protein sequences by shell scripts compiled for running PSI-BLAST. A total of 10 iterations were performed. The tabular output generated by PSI-BLAST was parsed by implementing filters of 110 amino acids overlap, 25% identity within the alignment and an e-value cut-off of 0.0001. The sequences thus obtained were then put through the fold recognition procedures described earlier.

Annotated genes in mycobacterial genomes

Genes in mycobacterial genomes annotated as 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 were also carefully examined. This examination resulted in the addition of more lectins for further consideration.

Identification of lectins from partial sequences of mycobacterial genomes

A search of partial genome sequences of mycobacteria for lectins was conducted in the NR database using the sequences of lectins already identified from the completely sequenced genomes employing NCBI-Protein BLAST. This search, followed by the fold recognition procedures yielded more putative lectins.

Sequence comparison and visualization

Sequence comparison was carried using MatGAT,[49] CLUSTALW.[50] Along with the results from CDD and PHYRE, sequence comparison was used to define the boundaries of appropriate domains or parts of them. Transmembrane segments were searched using SOSUI.[51] Coot, Pymol (http://www.pymol.org) and Rasmol (http://rasmol.org/) were used for macromolecular visualization and binding site analysis.

RESULTS

As detailed below, the identification of 94 mycobacterial lectins presented in this article has relied primarily on fold recognition, information on the annotation of mycobacterial genomes and the presence of appropriate sugar binding residues. A majority of the lectins were recognized through domain identification procedures using CDD, PHYRE, and I-TASSER. The results were confirmed by a sequence based approach using PSI-BLAST. Furthermore, proteins clearly identified as lectins through annotation of mycobacterial genomes were also included in the analysis.

The multipronged approach outlined in the methods section has resulted in a reasonably robust identification of a set of lectins or lectin domains in mycobacteria with known or partially known genome sequences. The results also bear testimony to the approach based on fold recognition. A search through 136,817 open reading frames in 30 fully known genome sequences using CDD resulted in 60 sequences, which have homologs with known lectin structure. PHYRE also predicted well annotated lectin folds with good e-value (<0.001) and high precision (>90%) for 43 out of the 60 sequences. Seventeen of the omitted sequences were very similar to the sequence of *Microcystis virdis* lectin (MVL).[52] Forty-four of the 60, which exhibited substantial presence of the appropriate sugar binding residues, were identified as putative lectins in mycobacteria from completely sequenced genomes.

The sequence based search involving lectin queries of the same ORF dataset carried out in parallel yielded 39 putative lectins. Interestingly, all of them had already been picked up by the fold recognition procedures as well. In fact, fold recognition had resulted in five more lectins. These five have an identity ranging from 20 to 25% with one or more of the query sequences. It is the criterion of a minimum of 25% identity used in the sequence search that resulted in them not being picked up when sequence alone was used for the search. In any case, the good convergence of the results obtained through two different approaches with a rational explanation for disagreements, adds to the confidence on the acceptability of these results. Subsequently, the sequences of the lectins identified from whole mycobacterial genomes were used to identify 20 more lectins from mycobacteria with partially sequenced genomes.

It is interesting to compare the approach adopted here with that used by Someya *et al.*[53] for prediction of carbohydrate-binding proteins. The latter is based exclusively on sequences while the approach here is primarily based on comparison with three-dimensional structures. Thus, the results obtained by employing the method of Someya *et al.* would encompass all possibilities whereas only ORFs with homologs, which have been established as lectins on the basis of three-dimensional structures, have been accepted for the present analysis except in the case of HBHAs (see below).

There are many genes annotated as lectins or lectin domains in mycobacterial genomes. A couple of them have homologs with known three-dimensional structure and they form part of the set identified on the basis of fold recognition. The annotated genes, in addition, contain 22 HBHAs from whole genomes and eight from partially sequenced genomes. The three-dimensional structure of no HBHA has been determined. However, their importance as lectins has been demonstrated.[40] Therefore, these proteins were also included in the list of mycobacterial lectins as a special group. Thus, the mycobacterial lectins or lectin domains identified in the present study consists of 64 sequences with homologs of known three-dimensional structure and 30 HBHAs.

Distribution of the lectin families in mycobacteria with completely sequenced genomes

The lectins or lectin domains identified by analyzing completely sequenced mycobacterial genomes (Table I) belong to five lectin families.

β-prism II lectins

Among the identified sequences, five involve a β-prism II fold domain along with a LysM domain, which is always found at the C-terminal region [Fig. 1(a)]. In none of the sequences, β-prism II fold appears alone or with any domain other than LysM. β-prism II fold lectins occur extensively in monocots and are invariably specific to mannose and higher oligomers.[12,54,55] The fold is threefold symmetric and roughly involves three Greek keys, which occur in a polypeptide chain of ~110 residues. The fold has therefore been suggested to have evolved through gene duplication and fusion of a primitive motif. Each Greek key carries a sugar binding site with a consensus sequence of Q-X-D-X-N-X-V-X-Y. These plant lectins mostly occur as tetramers or dimers. Some of these lectins possess antiretroviral activity[56] whereas some others do not. This difference has been explained in terms of the nature of oligomerization of the lectin.[57] Unlike in mycobacteria, β-prism II fold lectins occur independently in plants. The 50 residue long LysM domain, which occurs in tandem with the lectin domain in mycobacteria, is known to be a peptidoglycan-binding module useful for bacterial cell wall degradation.[58] In other eubacteria, the β-prism II fold lectin domain has been reported to occur in conjunction with LysM, PI-PLC, cysteine protease, and metalloprotease domains.[59] However, no structures of such proteins are currently available.

Homologs of β-prism II fold lectins, along with a LysM domain in each case, occur in *M. smegmatis*, *M. marinum, M. abscessus, M. sp.* JDM601 and *M. ulcerans* (Supporting Information Fig. 1). The consensus sugar binding sequence occurs in all of them except for variations at the last position. Except for *M. ulcerans*, these microorganisms are either nonpathogenic or only mildly pathogenic. The combined length of the two domains and the linker between them range between 188 and 209 residues. On an average, the lectin domain, the LysM domain and the linker region contain 105, 50, and 40 amino acids, respectively. The sequences from *M. marinum* and *M. ulcerans* are almost identical. The domains from *M. smegmatis, M. sp.* JDM601, and *M. abscessus* have sequence identities of 72%, 71% and 58%, respectively with respect to that from *M. marinum*.

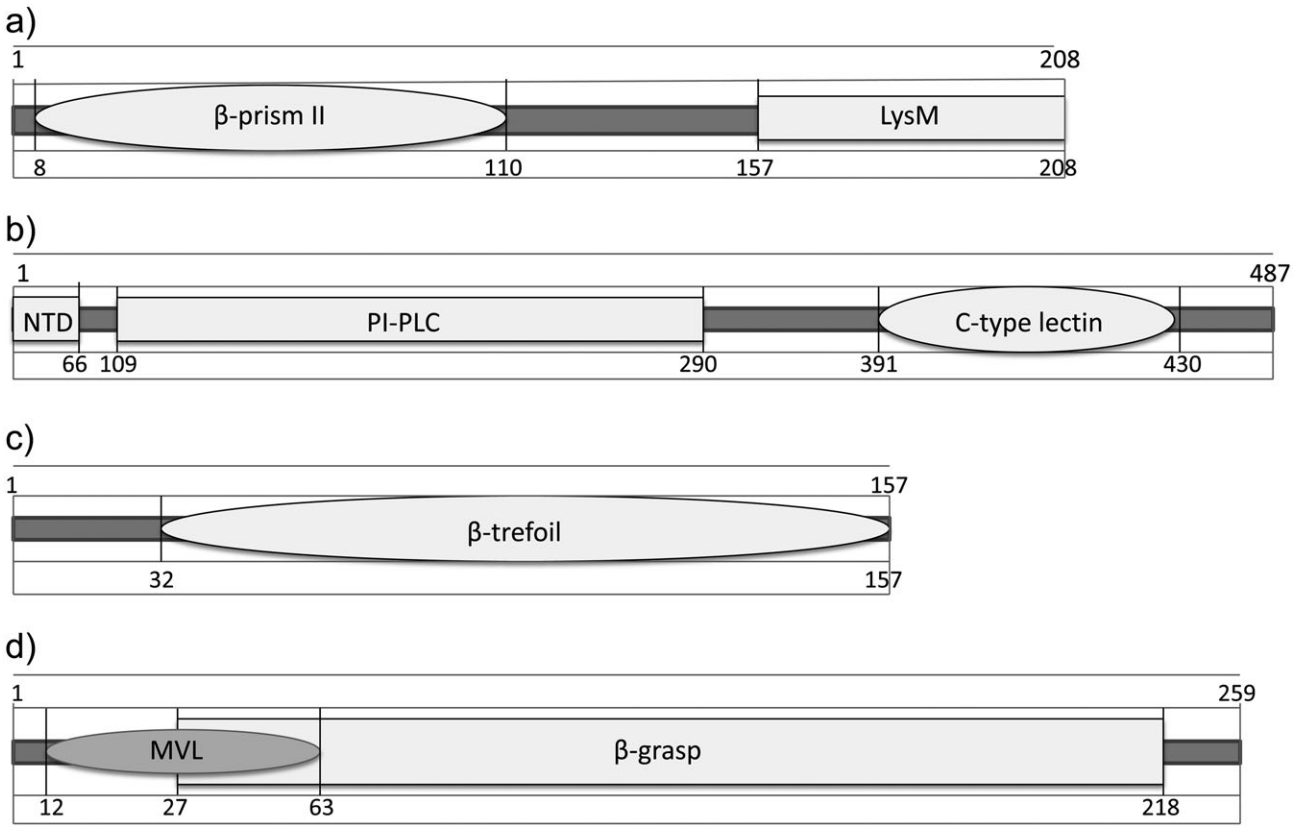


Figure 1

Examples of domain organization in mycobacterial sequences containing whole or part of (**a**) β-prism II, (**b**) C-type lectin, (**c**) β-trefoil, and (**d**) MVL domains. The accompanying domains are also indicated. The examples chosen are mentioned in the text under Discussion and in the legends for Figures 3–6.

C-type lectins

C-type lectins have been thoroughly characterized in animals. This carbohydrate-binding module most often occurs in tandem with other domains.[5,60] A segment of this module appears in eight mycobacterial sequences, in each case in tandem with a part of phosphoinositide-specific phospholipase C (PI-PLC) domain, which is always found at the N-terminal region [Fig. 1(b)]. The C-type lectin domain in animals is ~135 amino acids long and contains a recognizable consensus sequence rich in cysteine residues, which contribute to the stability of the domain through disulfide bridges. C-type lectins constitute a large family of proteins. They could be mannose specific or galactose specific. The binding sites of both are similar, but not identical. The distribution of amino acids at the site in the mycobacterial lectins identified here is closer to that in mannose specific lectins. The interactions at the binding site of mannose specific C-type lectins involve an E-X-N/R motif followed by a glutamyl residue after about five amino acid and then by an asparaginyl residue after about a dozen residues. A similar pattern is observed in the mycobacterial lectin segments. The middle E is however replaced by a Q, which can support most of the interactions involving E. However, N does not occur at the expected location. A close examination of known structures shows that part of the interactions involving N can be achieved by a T, which occurs towards the C-terminus. The functions of C-type lectin domains in animals have been well characterized. They include those in host pathogen interactions and innate immune response.[61,62] The roughly 286 amino acid long PI-PLC domain is an ubiquitous enzyme catalyzing the cleavage of the sn3-phosphodiester bond in the membrane phosphoinositide.[63]

The segment of C-type lectins, which occurs in tandem with part of PI-PLC domain, is found in all strains of *M. tuberculosis*, *M. marinum*, and *M. canetti* (Supporting Information Fig. 2). Sequences of the proteins from different *M. tuberculosis* strains are identical. The sequence of the protein from *M. canetti* is the same as that of the protein from *M. tuberculosis* except for one substitution. The protein from *M. marinum* exhibits a sequence identity of 72% with respect to that of *M. tuberculosis* and *M. canetti*. In the domain organization, the PI-PLC

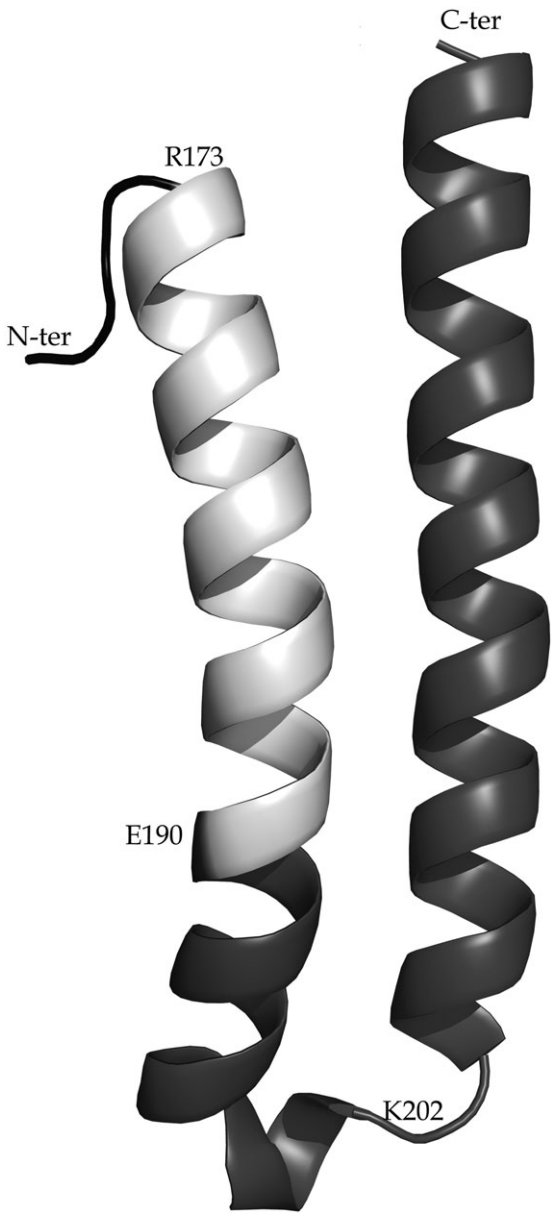


Figure 2

The C-domain (dark) and the insertion involving acidic residues (light) in the best model of the *M. smegmatis* HBHA predicted by I-TASSER.

domain is preceded by a transmembrane domain and followed by the lectin domain.

β-trefoil lectins

The β-trefoil was first characterized as a lectin domain in ricin from *Ricinus communis*. Ricin is a type II ribosome inactivating protein (RIP) containing a lectin chain and a toxin chain held together by a disulfide bond.[64] The lectin chain contains two β-trefoil domains, each with an approximate threefold symmetry. This galactose specific domain, roughly 140 amino acids long, has been suggested to have evolved through gene duplication and fusion of a carbohydrate binding motif. Among the three foils in each domain, only one carries the carbohydrate binding site. Most of the protein-sugar interactions at the site are through an individual aspartic acid residue. A glutamyl residue, an asparaginyl residue, and an aromatic residue are also present at the site.

The mycobacterial lectins identified from whole genomes in the present analysis include 14 with the β-trefoil fold (Supporting Information Fig. 3). These proteins are not associated with any other domain [Fig. 1(c)]. They are single domain β-trefoil lectins. They occur in all pathogenic mycobacteria except *M. leprae* and *M. avium*. The length of the lectin ranges from 156 to 158 amino acids. The lectins in different strains of *M. tuberculosis, M. africanum, M. canetti*, and *M. bovis* have nearly identical sequences. *M. ulcerans* and *M. marinum* cluster together with more than 98% sequence identity between them. The two clusters exhibit a sequence identity of nearly 70% between them. All the identified mycobacterial β-trefoil lectins have the aspartyl residue at the binding site referred to earlier. The asparaginyl and the aromatic residues found in lectins of the type found in RIPs, are present in the mycobacterial lectins as well, though at transposed but structurally close locations.

MVL lectins

The lectin from the cyanobacterium *Microcystis virdis*-MVL, which is inhibitory to HIV-1, was identified to have a new lectin fold through X-ray analysis in 2005.[52] Each subunit of this dimeric lectin is made up of two domains connected by a five residue linker. The two 54 amino acid long domains are homologs with a sequence identity of 50% between them. Each domain binds a complex carbohydrate. The primary carbohydrate binding site of both the domains is characterized by a consensus sequence element GQW.

One MVL lectin domain with an approximate length of 50 residues occurs in 17 of the identified sequences (Supporting Information Fig. 4). Invariably, the sequence also contains a nearly 200 residue long immunogenic protein with β-grasp domain from *Mycobacterium tuberculosis* with unknown function (DUF3298) [Fig. 1(d)].[65] As discussed later, the two domains overlap substantially. Interestingly, two such sequences occur in the whole genomes of *M. marinum, M. ulcerans, M. avium* 104, two strains of *M. intracellulerae* (MOTT-02 and MOTT-64) and *M. avium paratuberculosis* while one occurs in those of *M. abscessus, M. sp.* JDM601, *M. vanbalenii, M. gilvum,* and *M. sp.* Spyr1. The 17 sequences from the 11 whole genomes could not be clustered in any sensible manner. The sequence identities between pairs of them vary from 50 to 100%. The consensus sequence element at the carbohydrate site of MVL-lectins is substantially conserved in the mycobacterial lectin domains. In this

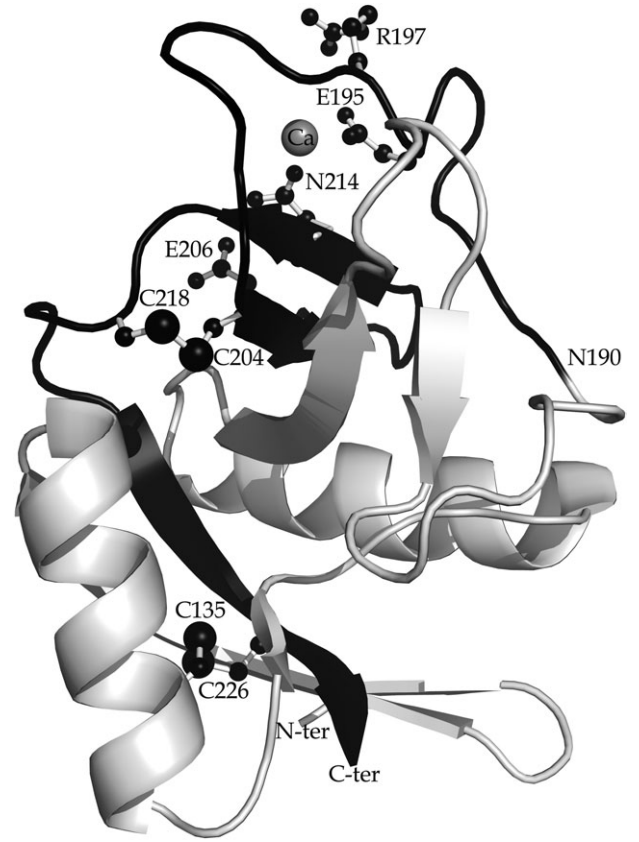


Figure 3

The C-type lectin domain of lung surfactant protein, SP-A (PDB code: 1R13). The segment identified in Rv2075c from *Mycobacterium tuberculosis* H37Rv is highlighted (in dark). The calcium ion is indicated. The large filled circles represent sulfur atoms in disulphide bridges. In this and the subsequent figures, sugar binding residues are shown in ball and stick representation.

three residue element, the first is always G and the third is always an aromatic residue. There is some variability at the second position, which is most often a glutamine.

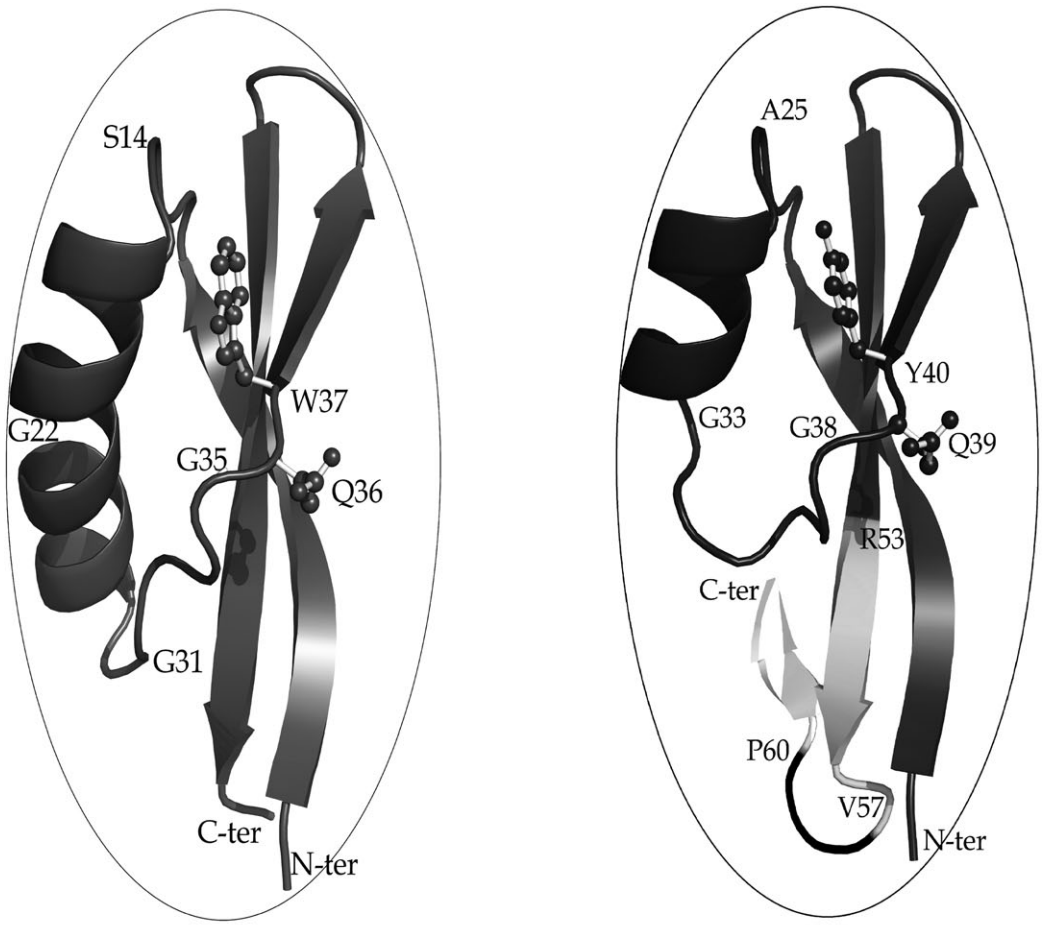
Heparin binding hemagglutinin

HBHA was first characterized in mycobacteria in the 1990s.[66] The 199 amino acid long glycosylated protein from *M. tuberculosis* has a molecular weight of 22 kDa. It agglutinates rabbit erythrocytes and promotes mycobacterial aggregation *in vitro*.[41,66–69] These and further studies indicated that the protein consists of a N-terminal 18 residue long trigger sequence, an 81 residue long α-helical coiled coil domain and a C-terminal Lys-Pro-Ala rich domain that mediates sugar binding. The molecule exists as a dimer in solution.[68,70,71] The N-terminal and the coiled-coil region are involved in dimerization.[72] It was also shown that HBHA is required for the extrapulmonary dissemination of *M. tuberculosis*.[41] The lectin binds to the sulfated glyconjugates present on the surface of the human respiratory epithelial cells including heparin, dextran sulfate, fucodian, and chondroitin sulfate.[68,69] Sugar binding is mediated by the C-terminal domain, which contains lysine rich repeats (KKAAPA) and (KKAAAKK).[68] The lectin binds heparin sulfate only when specific lysines are methylated.[73–75] Binding of HBHA to cell-surface sugar leads to a variety of biological effects. It was also shown that HBHA induces receptor mediated endocytosis through the recognition of heparin sulfate-containing proteoglycans by the heparin binding domain of the adhesin. It was also suggested that HBHA induces epithelial trancytosis,[74] which may represent a macrophage independent extrapulmonary dissemination mechanism leading to systemic infection by *M. tuberculosis*.[76] *M. tuberculosis* HBHA is also known to bind the human complement component C3 and mediate attachment and phagocytosis of the pathogen by mononuclear phagocytosis.[77] Thus HBHA mediates two critical functions: adherence to epithelial cells and establishment of pathogenesis. Although first thoroughly characterized in *M. tuberculosis*, HBHA from *M. leprae, M. bovis,* and *M. smegmatis* have also been extensively studied in recent years.[73,78,79] These results provide evidence that interactions of adhesins such as HBHA with nonphagocytic cells have an important role in the pathogenesis of tuberculosis. In addition to mycobacteria, this protein has been reported to occur in other gram positive aerobic nonsporulating bacteria like *Rhodococcus erythropolis*, *Clavibacter michiganesis, and Nocardia farcinica,* which are taxonomically very close to mycobacteria.[59]

The present analysis indicates that HBHA occurs in all but one strain of *M. tuberculosis*, *M. leprae*, *M. bovis*, *M. avium*, *M. ulcerans*, *M. africanum*, *M. canetti*, *M. sp* JDM 601, *M. intracellulerae*, *M. rhodesiae*, *M. abscessus*, *M. marinum*, *M. smegmatis* (Supporting Information Fig. 5). Of these, the first 11 species are pathogenic to different extents while the last two are nonpathogenic or mildly pathogenic. HBHA from *M. tuberculosis*, *M. bovis*, *M. africanum,* and *M. canetti* have identical sequences and forms a major cluster. The proteins from *M. leprae*, *M. avium*, *M. marinum*, *M. intracellulerae,* and *M. ulcerans* cluster together with a sequence identity between pairs of species varying between 73 and 96%. Each of them has sequence identity in the range of 81–86% with that of the HBHA from *M. tuberculosis* H37Rv and its companions. The sequences of the protein from *M. smegmatis*, *M. sp.* JDM601, *M. abscessus*, and *M. rhodesiae* exhibit identities of 59–74% with respect to that of the *M. tuberculosis* H37Rv. Sequence identities of pairs among the four vary between 53 and 70%.

Lectins identified from partially sequenced genomes

Lectins identified from partial genome sequences (Table II) are only mentioned in passing here as they are




```
MVL              1 -----------ASYKVNIPAGPLWSNAEAQQVGPKIAAAHQGNFTGQWT-     38
                             |:.....|.....:.|....:|        |::.||:.
Mvan_0966        1 MRRAFAAAAAAATLLATAPTASASAVAFCDGLG--------GSWNGQYCH     42

MVL             39 TVVESAMSVVEVELQVASYKVNIPAGPLWSNAEAQQVGPKIAAAHQGNFT     88
                   |.|:|..:.|.      ..||.|||..
Mvan_0966       43 TTVQSERNAVR------DIKVAIPAEL----------------------     63

MVL             89 GQWTTVVESAMSVVEVELQV    108

Mvan_0966       64 --------------------     63
```

(c)

Figure 4

(**a**) N-terminal domain of *Microcystis virdis* lectin (PDB code: 1ZHS). (**b**) A model of the homologs domain constructed on the basis of the alignment shown in (c). (**c**) Sequence alignment between two N-terminal domains of MVL in tandem and the first 63 residues of Mvan_0966 from *Mycobacterium vanbaalenii*.

likely to be reviewed after the sequencing is completed. Among the 20 putative lectin sequences obtained from partially sequenced mycobacterial genomes, 3 contain a β-prism II domain, 5 a C-type lectin domain, and 9 a MVL domain with the same domain organization found in sequences obtained from the completed genome proj-

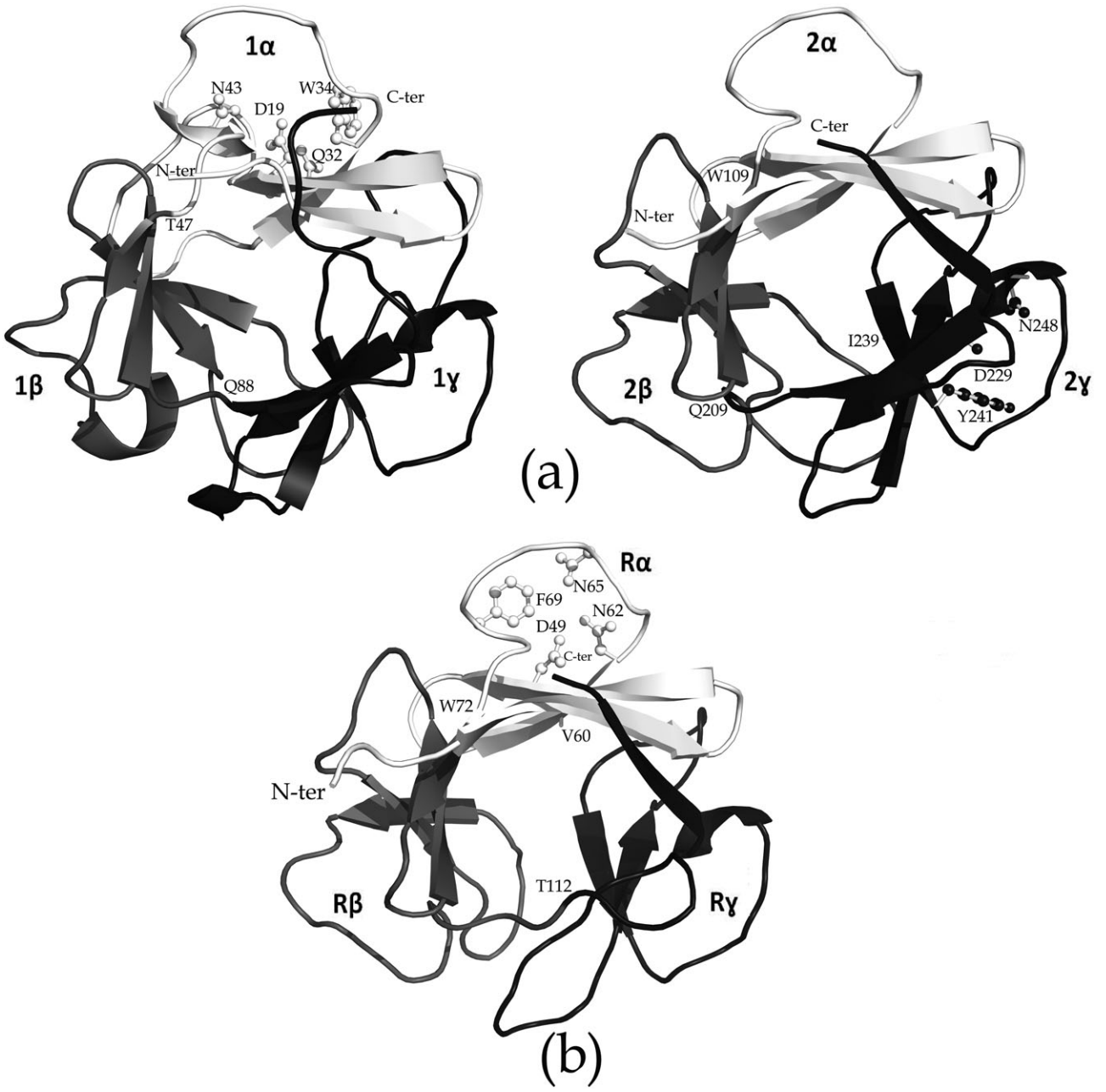


Figure 5

(**a**) The lectin domains of Himalayan mistletoe (PDB code: 1YF8, B-chain). Foils in the lectin domain 1 are designated 1α, 1β, and 1γ while those in domain 2 are designated as 2α, 2β, and 2γ. (**b**) Homology model of the lectin domain in Rv1419 from *Mycobacterium tuberculosis* H37Rv based on the domain 1 of the lectin chain of Himalayan mistletoe lectin.

ects, in addition to three containing a β-trefoil domain. Eight HBHAs have also been identified in sequences from partially sequenced genomes.

DISCUSSION

Among the lectin families identified in mycobacteria, the most important is perhaps the HBHA family. Of the 30 fully sequenced mycobacterium genomes, as many as 22 contain HBHA genes. More pertinently, available evidence suggests important roles for this lectin in site specific adherence to the host and establishment of pathogenesis. The site is known to involve sulfated glyconjugates. In this context, differences between pathogenic mycobacteria such as *M. tuberculosis* and *M. leprae* on the one hand and the nonpathogenic *M. smegamtis* on the other,

Table II
Lectin Domains Identified from Partially Sequenced Mycobacterial Genomes

	Names of the lectins identified from partially sequenced genomes				
Organism	β-prism II lectins	C-type lectins	β-trefoil lectins	MVL lectin	HBHA
Mycobacterium kansasii ATCC 12478	1	1	1	0	1
Mycobacterium tuberculosis SUMu012	0	1	0	0	0
Mycobacterium tuberculosis 94_M4241A	0	1	0	0	0
Mycobacterium tuberculosis GM 1503	0	1	0	0	0
Mycobacterium tuberculosis CPHL_A	0	0	1	0	0
Mycobacterium tuberculosis `98-R604 INH-RIF	0	0	0	0	1
Mycobacterium tuberculosis T46	0	1	0	0	0
Mycobacterium colombiense CECT 3035	1	0	0	2	1
Mycobacterium tuberculosis SUMu006	0	0	1	0	0
Mycobacterium rhodesiae JS60	0	0	0	0	1
Mycobacterium avium paratuberculosis S397	0	0	0	0	1
Mycobacterium abscessus 47J26	1	0	0	1	0
Mycobacterium intracellulare ATCC 13950	0	0	0	2	1
Mycobacterium avium subsp. avium ATCC 25291	0	0	0	2	1
Mycobacterium thermoresistible ATCC 19527	0	0	0	1	0
Mycobacterium parascrofulaceum ATCC BAA-614	0	0	0	1	1

in the 20 amino acid peptide stretch that precedes the C-domain is of considerable interest. The C-domain is characterized by the presence of a number of positively charged lysyl residues. In *M. smegmatis,* unlike in pathogenic bacteria, the preceding stretch contains an insertion involving negatively charged glutamyl residues.[78] This insertion possibly interferes with the interaction of the C-domain with the negatively charged complex oligosaccharides like heparin sulfate. Modeling studies on the proteins from *M. tuberculosis, M. leprae,* and *M. smegmatis* employing I-TASSER are consistent with the above inference. Admittedly, results of the modeling need to be treated with caution. However, it is interesting that the 35 amino acid long C-domain of *M. tuberculosis* HBHA and the 23 amino acid long C-domain of *M. leprae* HBHA always form a single helix each. However, the 41 amino acid long C-domain of *M. smegmatis* protein form two antiparallel helices of unequal lengths such that the C-terminal stretch (203–232) is in close proximity to the insertion containing the acidic residues (Fig. 2), thus impairing its ability to interact with the negatively charged stretch of the host binding site.[78]

The set of proteins containing a segment of the C-type lectin domain along with a part of PI-PLC domain presents an interesting case. This segment overlaps with a nearly 40 amino acid long stretch of classical C-type lectins (Fig. 3). The segment forms part of the sequence motif suggested earlier by Drickamer and contains most of the carbohydrate binding residues.[60] The location of cysteine residues in the segment is compatible with the location of the conserved disulfide bridge within the corresponding stretch in the C-type lectin. Presumably, the conformation of the stretch in C-type lectins is dependent of the rest of molecule as well. In the mycobacterial sequences, the segment is preceded and followed by unannotated polypeptide stretches. Perhaps one or the other or both the stretches help the segment to assume the required three-dimensional structure.

Although CDD and PHYRE identified a PIPLC domain in the mycobacterial sequences containing the C-type lectin segment, a good sequence match for the whole domain could not be obtained. Taking all the results together, the match between 109 and 290 stretch in Rv2075c from *Mycobacterium tuberculosis* H37Rv and the 21–190 stretch of the 339 residue calcium dependent phosphatidyl inositol-specific phosphatase from *Streptomyces antibioticus* is reasonable. The sequence identity between the two stretches is still low, but significant at 22%. The sequence identity between the 40 residue segment of the same mycobacterial protein and SP-A is substantially higher at 30%.

The sequences containing homologs of MVL also present a situation involving domain identification with low sequence similarity. Both the domains of the two domains MVL overlap individually with the N-terminal segment of the relevant mycobacterial sequence with a sequence identity of about 25% in the case of the typical mycobacterial protein Mvan_0966 from *Mycobacterium vanbalenii* (Fig. 4). A still better overlap with the first domain involving an arrangement reminiscent of circular permutation exists. As the sequence similarity is low, it is difficult to choose from among the very similar possibilities. The balance of evidence indicates that the MVL-lectin domain spans the 12–63 stretch in the mycobacterial sequence. Unlike the other lectin domains identified in mycobacteria, the MVL-like domains have only one homolog of known structure. However, the homolog belongs to another bacterium. Furthermore, this domain could represent a very early carbohydrate binding motif as it occurs in ancient organisms like the cyanobacteria. The poor sequence similarity between the cyanobacterial MVL and the MVL-like domains in mycobacteria

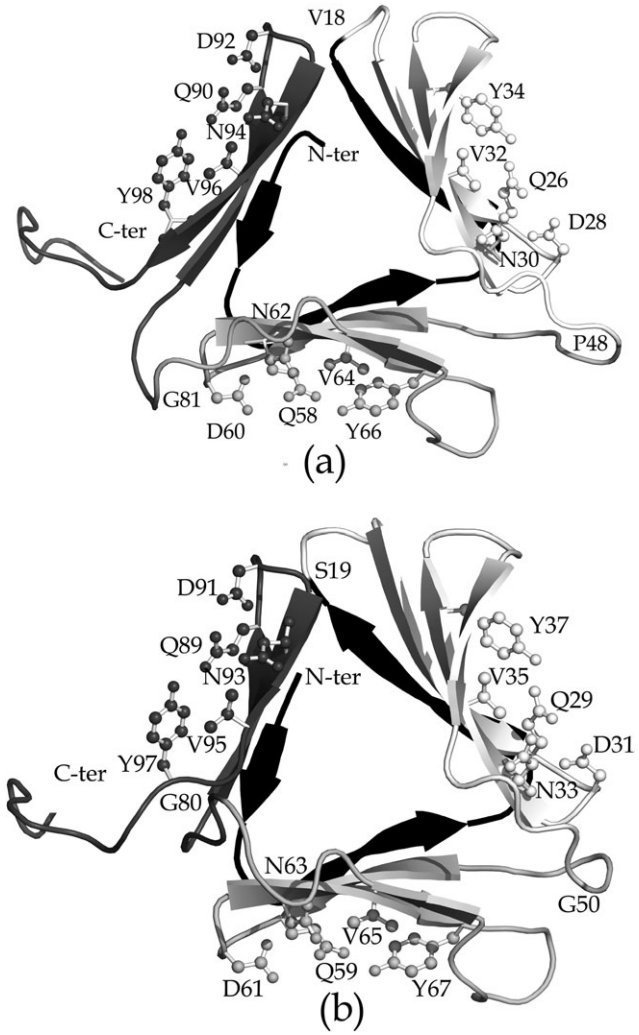


Figure 6

(**a**) A subunit of garlic lectin (PDB code: 1KJ1, D-chain). (**b**) Homology model of the lectin domain of protein from *Mycobacterium smegmatis* based on (a).

perhaps indicates divergent evolution. However, as indicated in Figure 4(a), the mycobacterial lectin Mvan_0966 forms a model similar in shape and size to the MVL domain, which can be constructed based on the sequence alignment shown in Figure 4(c).

An intriguing feature of the mycobacterial sequences containing the MVL domain is the substantial overlap between the MVL domain and the β-grasp domain. The MVL domain appears to involve the 12–63 polypeptide stretch in the *M. vanbalenii* sequence. At the same time, the 27–218 stretch of the same sequence aligns with the sequence of the 204 amino acid long *M. tuberculosis* β-grasp domain[65] with a sequence identity of 18%, thus indicating a 36 amino acid overlap between the MVL domain and the β-grasp domain. Comparison between the relevant three-dimensional structures indicates that the region in the β-grasp domain corresponding to this 36 residue stretch could constitute a MVL-domain sans the first long β-strand and part of the helix. Therefore, the possible ability of the β-grasp domain to bind sugar merits further exploration.

The sequence similarity between mycobacterial β-trefoil lectin domains and their plant homologs, although low, presents an interesting situation, which is best illustrated by comparing the two (Fig. 5). The bacterial protein contains one trefoil domain each while the lectin chain of the well established Type II RIPs accounts for two such domains. As indicated earlier, each domain is believed to have originated through successive gene duplication, fusion, and divergent evolution of a primitive carbohydrate-binding motif. In this regard, the sequence identity between pairs of foils in the typical mycobacterial lectin Rv1419 from *Mycobacterium tuberculosis* H37Rv varies between 16 and 24%. In the lectin domain 1 of Himalayan mistletoe,[80] the range is 10–21% in domain 1 while it is 14–23% in domain 2. The sequence identity between the foils in domain 1 and those in domain 2 varies between 14 and 20%. The sequence identity between the foils in domain 1 of Himalayan mistletoe lectin and those in Rv1419 ranges between 11 and 22%. The corresponding values when domain 2 is used are 14 and 22%. Thus, the foils in the selected plant lectin and the mycobacterial lectin exhibit nearly the same degree of relatedness irrespective of whether the comparison is between the foils of the same domain, two different domains or from a plant and a bacterium. This could perhaps mean that the primitive carbohydrate-binding motif referred to earlier is of very ancient origin.

The mycobacterial sequences containing the β-prism II fold presents the simplest case of domain identification. The lectin domain of the sequence from *M. smegmatis*, for instance, has a sequence identity of 44% with garlic lectin[57,81] a well known β-prism II fold plant lectin. The LysM domain of the same ORF exhibits a sequence identity of 40% with the MoCVNH LysM module of the rice blast fungus *Magnaporthe oryzae* protein.[82] The sequence identity among the three Greek keys in garlic lectin is around 30%. The corresponding value in the *M. smegmatis* lectin is around 40%. Indeed among the mycobacterial domains identified in the present study, those with the β-prism II fold exhibit the maximum similarity with the homologs from other sources with known three-dimensional structure (Fig. 6).

Dozens of lectin domains have been identified through structural, primary X-ray crystallographic, studies. Interestingly, only 5 of them appear to occur in mycobacteria. Of these 5, the functional role of only one, namely, HBHA, has been studied thoroughly. The molecular structure of this protein is, however, yet to be established. The only structural investigations reported on mycobacterial lectins so far are accounts of crystallization and

preliminary X-ray studies on a *M. tuberculosis* lectin[20] and a *M. smegmatis* lectin domain[21] that emanated from this laboratory. Biochemical characterization of the *M. tuberculosis* lectin Rv1419 has also been reported indicating its ability to agglutinate rabbit erythrocytes.[83] Thus, mycobacterial lectins constitute a largely unexplored area. Recognition of specific carbohydrates is important in cell–cell interactions including host-pathogen interactions. Therefore, mycobacterial lectins merit a thorough study. The effort presented here is a systematic attempt to enable such a study, which has already been initiated in this laboratory.

ACKNOWLEDGMENTS

The authors thank N. Srinivasan for suggestions. Facilities at the Interactive Graphics Facility, supported by the Department of Biotechnology, Government of India have been used in this work.

REFERENCES

1. Drickamer K, Taylor ME. Biology of animal lectins. Annu Rev Cell Biol 1993;9:237–264.
2. Vijayan M, Chandra NR. Lectins. Curr Opin Struct Biol 1999;9:707–714.
3. Loris R. Principles of structures of animal and plant lectins. Biochim Biophys Acta 2002;1572:198–208.
4. Sharon N. Lectins: carbohydrate-specific reagents and biological recognition molecules. J Biol Chem 2007;282:2753–2764.
5. Veldhuizen EJA, van Eijk M, Haagsman HP. The carbohydrate recognition domain of collectins. FEBS J 2011;278:3930–3941.
6. Chandra NR, Kumar N, Jeyakani J, Singh DD, Gowda SB, Prathima MN. Lectindb: a plant lectin database. Glycobiology 2006;16:938–946.
7. 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.
8. 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.
9. 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.
10. Banerjee R, Mande SC, Ganesh V, Das K, Dhanaraj V, Mohanta SK, Suguna K, Surolia A, Vijayan M. Crystal structure of peanut lectin, a protein with an unusual quaternary structure. Proc Natl Acad Sci U S A 1994;91:227–231.
11. Sankaranarayanan R, Sekar K, Banerjee R, Sharma V, Surolia A, Vijayan M. Nat Struct Biol 1996;3:596–603.
12. Ramachandraiah G, Chandra NR, Surolia A, Vijayan M. Computational analysis of multivalency in lectins: structures of garlic lectin-oligosaccharide complexes and their aggregates. Glycobiology 2003;13:765–775.
13. 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.
14. Singh DD, Saikrishnan K, Kumar P, Sekar K, Surolia A, Vijayan M. Unusual sugar specificity of banana lectin from *Musa paradisiaca* and its probable evolutionary origin. Crystallographic and modelling studies. Glycobiology 2005;15:1025–1032.
15. Kulkarni KA, Katiyar S, Surolia A, Vijayan M, Suguna K. Generation of blood group specificity: new insights from structural studies on the complexes of A- and B-reactive saccharides with basic winged bean agglutinin. Proteins 2007;68:762–769.
16. Sharma A, Sekar K, Vijayan M. Structure, dynamics, and interactions of jacalin. Insights from molecular dynamics simulations examined in conjunction with results of X-ray studies. Proteins 2009;77:760–777.
17. Chandran T, Sharma A, Vijayan M. Crystallization and preliminary X-ray studies of a galactose-specific lectin from the seeds of bitter gourd *Momordica charantia*. Acta Cryst 2010;F66:1037–1104.
18. Sharma A, Vijayan M. Influence of glycosidic linkage on the nature of carbohydrate binding in β-prism I fold lectins: an X-ray and molecular dynamics investigation on banana lectin–carbohydrate complexes. Glycobiology 2011;21:23–33.
19. Natchiar SK, Suguna K, Surolia A, Vijayan M. Peanut agglutinin, a lectin with an unusual quaternary structure and interesting ligand binding properties. Crystallogr Rev 2007;13:3–28.
20. Patra D, Srikalaivani R, Misra A, Singh DD, Selvaraj M, Vijayan M. Cloning, expression, purification, crystallization and preliminary X-ray studies of a secreted lectin (Rv1419) from *Mycobacterium tuberculosis*. Acta Cryst 2010;F66:1662–1665.
21. Patra D, Sharma A, Chandran D, Vijayan M. Cloning, expression, purification, crystallization and preliminary X-ray studies of the mannose-binding lectin domain of MSMEG_3662 from *Mycobacterium smegmatis*. Acta Cryst 2011;F67:596–599.
22. Wiley DC, Wilson IA, Skehel JJ. Structure of the haemagglutinin membrane glycoprotein of influenza virus at 3 Å resolution. Nature 1981;289:373–378.
23. Skehel JJ, Wiley DC. Receptor binding and membrane fusion in virus entry: the influenza hemagglutinin. Annu Rev Biochem 2000;69:531–569.
24. Zhang RG, Westbrook ML, Westbrook EM, Scott DL, Otwinowski Z, Maulik PR, Reed RA, Shipley GG. The 2.4Å crystal structure of cholera toxin B subunit pentamer: choleragenoid. J Mol Biol 1995;251:550–562.
25. Merritt EA, Kuhn P, Sarfaty S, Erbe JL, Holmes RK, Hol WG. The 1.25 Å resolution refinement of the cholera toxin B-pentamer: evidence of peptide backbone strain at the receptor-binding site. J Mol Biol 1998;282:1043–1059.
26. Swaminathan S, Eswaramoorthy S. Structural analysis of the catalytic and binding sites of Clostridium botulinum neurotoxin B. Nat Struct Biol 2000;7:683–699.
27. Emsley P, Fotinou C, Black I, Fairweather NF, Charles IG, Watts C, Hewitt E, Isaacs NW. The structures of the H_C fragment of tetanus toxin with carbohydrate subunit complexes provide insight into ganglioside binding. J Biol Chem 2000;275:8889–8894.
28. Duguid JP, Old DC.Bacterial adherence, receptors and recognition, Series B, Vol.6. London: Chapman and Hall; 1980.
29. Imberty A, Wimmerova M, Mitchell EP, Gilboa-Garber N. Structures of the lectins from *Pseudomonas aeruginosa*: insights into the molecular basis for host glycan recognition. Microbes Infect 2004;6:221–228.
30. Bewley CA. Protein-carbohydrate interactions in infectious diseases. Cambridge: Royal Society of Chemistry; 2006.
31. Vijayan M. Structural biology of mycobacterial proteins: the Bangalore effort. Tuberculosis 2005;85:357–366.
32. Krishna R, Prabu JR, Manjunath GP, Datta S, Chandra NR, Muniyappa K, Vijayan M. Snapshots of RecA protein involving movement of the C-domain and different conformations of the DNA-binding loops: Crystallographic and comparative analysis of 11 structures of *Mycobacterium smegmatis* RecA. J Mol Biol 2007;367:1130–1144.
33. Selvaraj M, Roy S, Singh NS, Sangeetha R, Varshney U, Vijayan M. Structural plasticity and enzyme action: crystal structures of *Mycobacterium tuberculosis* peptidyl-tRNA hydrolase. J Mol Biol 2007;372:186–193.
34. Roy S, Saraswathi R, Chatterji D, Vijayan M. Structural studies on the second *Mycobacterium smegmatis* Dps: Invariant and variable

features of structure, assembly and function. J Mol Biol 2008;375:948–959.
35. Kaushal PS, Talawar RK, Krishna PDV, Varshney U, Vijayan M. Unique features of the structure and interactions of mycobacterial uracil-DNA glycosylase: structure of a complex of the *Mycobacterium tuberculosis* enzyme in comparison with those from other sources. Acta Cryst 2008;D64:551–560.
36. Prabu R, Manjunath GP, Chandra NR, Muniyappa K, Vijayan M. Functionally important movements in RecA molecules and filaments: studies involving mutational and conformational changes. Acta Cryst 2008;D64:1146–1157.
37. Prabu R, Thamotharan S, Khanduja JS, Chandra NR, Muniyappa K, Vijayan M. Crystallographic and modelling studies on *Mycobacterium tuberculosis* RuvA: additional role of RuvB-binding domain and inter species variability. Biochim Biophys Acta 2009;1794:1001–1009.
38. Chetnani B, Kumar P, Surolia A, Vijayan M. *M. tuberculosis* pantothenate kinase: dual substrate specificity and unusual changes in ligand locations. J Mol Biol 2010;400:171–185.
39. Chetnani B, Kumar P, Abhinav KV, Chhibber M, Surolia A, Vijayan M. Location and conformation of pantothenate and its derivatives in *Mycobacterium tuberculosis* pantothenate kinase: Insights into enzyme action. Acta Cryst 2011;D67:774–783.
40. Singh DD, Chandran D, Jeyakani J, Chandra NR. Scanning the genome of *Mycobac-terium tuberculosis* to identify potential lectins. Prot Pept Lett 2007;14:683–691.
41. Pethe K, Alonso S, Biet F, Delogu G, Brennan MJ, Locht C, Menozzi FD. The heparin-binding haemagglutinin of *M. tuberculosis* is required for extrapulmonary dissemination. Nature 2001;412: 190–194.
42. Marchler-Bauer A, Panchenko AR, Shoemaker BA, Thiessen PA, Geer LY, Bryant SH. CDD: a database of conserved domain alignments with links to domain three-dimensional structure. Nucleic Acids Res 2002;30:281–283.
43. Kelley LA, Sternberg MJE. Protein structure prediction on the web: a case study using the Phyre server. Na Protocols 2009;4:363–371.
44. Zhang Y. I-TASSER server for protein 3D structure prediction. BMC Bioinform 2008;9:40.
45. Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A. ExPASy: the proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res 2003;31:3784–3788.
46. Huang Y, Niu B, Gao Y, Fu L, Weizhong Li. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics 2010;26:680–682.
47. Altschul SF. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl Acids Res 1997;25:3389–3402.
48. 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.
49. Campanella JJ, Bitincka L, and Smalley J. MatGAT: An application that generates similarity/identity matrices using protein or DNA sequences. BMC Bioinformatics 2003;4:1471–2105.
50. 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 and ClustalX version 2.0. Bioinformatics 2007;23:2947–2948.
51. Hirokawa T, Boon-Chieng S, Mitaku S. SOSUI: classification and secondary structure prediction system for membrane proteins. Bioinformatics 1998;14:378–379.
52. William DC, Lee JY, Cai M, Bewley CA, Clore GM. Crystal structures of the HIV-1 inhibitory cyanobacterial protein MVL free and bound to $Man_3GlcNAc_2$. J Biol Chem 2005;280:29269–29276.
53. Someya S, Kakuta M, Morita M, Sumikoshi K, Cao W, Ge Z, Hirose O, Nakamura S, Terada T, Shimizu K. Prediction of carbohydrate-binding proteins from sequences using Support vector machines. Adv. Bioinform 2010;2010:289301.
54. Hester G, Kaku H, Goldstein IJ, Wright CS. Structure of mannose-specific snowdrop (*Galanthus nivalis*) lectin is representative of a new plant lectin family. Nat Struct Biol 1995;2:472–479.
55. Sharma A, Chandran D, Singh DD, Vijayan M. Multiplicity of carbohydrate-binding sites in β-prism fold lectins: occurrence and possible evolutionary implications. J Biosci 2007;32:1089–1110.
56. Balzarini J, Schols D, Neyts J, Van Damme EJM, Peumans WJ, Clercq ED. α-(1–3)- and α-(1–6)-D-mannose-specific plant lectins are markedly inhibitory to human immuno-deficiency virus and cytomegalovirus infections *in vitro*. Antimicrob Agents Chemother 1991;35:410–416.
57. Chandra NR, Ramachandraiah G, Bachhawat K, Dam TK, Surolia A, Vijayan M. Crystal structure of a dimeric mannose-specific agglutinin from garlic: quaternary association and carbohydrate specificity. J Mol Biol 1999;285:1157–1168.
58. Bateman A, Bycroft M. The structure of a LysM domain from *E. coli* membrane-bound lytic murein transglycosylase D (MltD). J Mol Biol 2000;299:1113–1119.
59. Punta M, Coggill PC, Eberhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer ELL, Eddy SR, Bateman A, Finn RD. The Pfam protein families' database. Nucl Acids Res 2012;40:D290–D301.
60. Drickamer K. Ca^{2+}-dependent carbohydrate-recognition domains in animal proteins. Curr Opin Struct Biol 1993;3:393–400.
61. Taylor ME, Conary JT, Lennartz MR, Stahl PD, Drickamer K. Primary structure of the mannose receptor contains multiple motifs resembling carbohydrate-recognition domains. J Biol Chem 1990;265:12156–12162.
62. Weis WI, Taylor ME, Drickamer K. The C-type lectin superfamily ìn the immune system. Immunol Rev 1998;163:19–34.
63. Moser J, Gerstel B, Meyer JEW, Chakraborty T, Wehland J, Heinz DW. Crystal structure of the phosphatidylinositol-specific phospholipase C from the human pathogen *Listeria monocytogenes*. J Mol Biol 1997;273:269–282.
64. 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.
65. Wang Z, Potter BM, Gray AM, Sacksteder KA, Geisbrecht BV, Laity JH. The solution structure of antigen MPT64 from *Mycobacterium tuberculosis* defines a new family of β-grasp proteins. J Mol Biol 2007;366:375–381.
66. Menozzi FD, Bischoff R, Fort E, Brennan MJ, Locht C. Molecular characterization of the mycobacterial heparin-binding hemagglutinin, a mycobacterial adhesin. Proc Natl Acad Sci U S A 1998;95: 12625–12630.
67. Menozzi FD. Identification of a heparin-binding hemagglutinin present in mycobacteria. J Exp Med 1996;184:993–1001.
68. Delogu G, Brennan MJ. Functional domains present in the mycobacterial hemagglutinin, HBHA. J Bacteriol 1999;181:7464–746.
69. Pethe K, Aumercier M, Fort E, Gatot C, Locht C, Menozzi FD. Characterization of the heparin binding site of the mycobacterial heparin-binding hemagglutinin adhesin. J Biol Chem 2000;275: 14273–1428.
70. Esposito C. Evidence for elongated dimeric structure of heparin-binding hemagglutinin from *Mycobacterium tuberculosis*. J Bacteriol 2008;190:4749–4753.
71. Esposito C. Dimerisation and structural integrity of heparin binding hemagglutinin A from *Mycobacterium tuberculosis*: implications for bacterial agglutination. FEBS Lett 2010;584:1091–1096.
72. Lomino JV, Tripathy A, Redinbo MR. Triggered *Mycobacterium tuberculosis* heparin-binding hemagglutinin adhesin folding and dimerization. J Bacteriol 2011;193:2089–2096.
73. Pethe K, Bifani P, Drobecq J, Sergheraert C, Debrie AS, Locht C, Menozzi FD. Mycobacterial heparin-binding hemagglutinin and laminin binding protein share antigenic methyllysines that confer resistance to proteolysis. Proc Natl Acad Sci U S A 2002;99: 10759–10764.

74. Temmerman S, Pethe K, Parra M, Alonso S, Rouanet C, Menozzi FD, Sergheraert S, Brennan MJ, Mascart F, Locht C. Methylation-dependent T cell immunity to *Mycobacterium tuberculosis* heparin-binding hemagglutinin. Nat Med 2004;10:935–941.
75. Delogu G, Bua A, Pusceddu C, Parra M, Fadda G, Brennan MJ, Zanetti S. Expression and purification of recombinant methylated HBHA in *Mycobacterium smegmatis*. FEMS Microbiol Lett 2004;239:33–39.
76. Krishnan N, Robertson BD, Thwaites G. The mechanisms and consequences of the extra-pulmonary dissemination of *Mycobacterium tuberculosis*. Tuberculosis 2010;90:361–366.
77. Mueller-Ortiz SL, Wanger AR, Norris SJ. Mycobacterial protein HBHA binds human complement component C3. Infect Immun 2001;69:7501–7511.
78. Biet F, Marques MA, Grayon M, Xavier da Silveira EK, Brennan PJ, Pessolani MC, Locht C, Menozzi FD. *Mycobacterium smegmatis* produces an HBHA homologue which is not involved in epithelial adherence. Microbes Infect 2007;9:175–182.
79. De-lima CS, Marques MA, Debrie AS, Almeida EC, Silva CA, Brennan PJ, Pessolani MV. HBHA from *Mycobacterium leprae* is expressed during infection and enhances bacterial adherence to epithelial cells. FEMS Microbiol Lett 2009;292:162–169.
80. Mishra V, Bilgrami S, Sharma RS, Kaur P, Yadav S, Krauspenhaar R, Betzel C, Voelter W, Babu CR, Singh TP. Crystal structure of Himalayan mistletoe ribosome-inactivating protein reveals the presence of a natural inhibitor and a new functionally active sugar-binding site. J Biol Chem 2005;280:20712–20721.
81. Ramachandraiah G, Chandra NR, Surolia A, Vijayan M. Re-refinement using reprocessed data to improve the quality of the structure: a case study involving garlic lectin. Acta Cryst 2002;D58: 414–420.
82. Koharudin MI, Viscomi AR, Montanini B, Kershaw MJ, Talbot NJ, Ottonello S, Gronenborn A. Structure-function analysis of a CVNH-LysM lectin expressed during plant infection by the rice blast fungus *Magnaporthe oryzae*. Structure 2011;19:662–674.
83. Nogueira N, Cardoso FC, Mattos AM, Bordignon J, Bafica A. *M. tuberculosis* Rv1419 encodes a secreted 13 kDa lectin with immunological reactivity during human tuberculosis. Eur J Immunol 2010;40:744–753.