

## Lectindb: a plant lectin database

Nagasuma R. Chandra<sup>1</sup>, Nirmal Kumar, Justin Jeyakani,  
Desh Deepak Singh, Sharan B. Gowda, and M. N. Prathima

Bioinformatics Centre and Supercomputer Education and Research  
Centre, Raman Building, Indian Institute of Science, Bangalore 560 012,  
Karnataka, India

Received on February 11, 2006; revised on May 17, 2006; accepted on  
June 6, 2006

**Lectins, a class of carbohydrate-binding proteins, are now widely recognized to play a range of crucial roles in many cell–cell recognition events triggering several important cellular processes. They encompass different members that are diverse in their sequences, structures, binding site architectures, quaternary structures, carbohydrate affinities, and specificities as well as their larger biological roles and potential applications. It is not surprising, therefore, that the vast amount of experimental data on lectins available in the literature is so diverse, that it becomes difficult and time consuming, if not impossible to comprehend the advances in various areas and obtain the maximum benefit. To achieve an effective use of all the data toward understanding the function and their possible applications, an organization of these seemingly independent data into a common framework is essential. An integrated knowledge base (Lectindb, <http://nscdb.bic.physics.iisc.ernet.in>) together with appropriate analytical tools has therefore been developed initially for plant lectins by collating and integrating diverse data. The database has been implemented using MySQL on a Linux platform and web-enabled using PERL-CGI and Java tools. Data for each lectin pertain to taxonomic, biochemical, domain architecture, molecular sequence, and structural details as well as carbohydrate and hence blood group specificities. Extensive links have also been provided for relevant bioinformatics resources and analytical tools. Availability of diverse data integrated into a common framework is expected to be of high value not only for basic studies in lectin biology but also for basic studies in pursuing several applications in biotechnology, immunology, and clinical practice, using these molecules.**

**Key words:** carbohydrate specificity/integrated database/  
lectin folds/sequence analysis/structural annotation

### Introduction

Lectins are carbohydrate-binding proteins that specifically recognize diverse sugar structures and mediate a variety of biological processes such as cell–cell and host–pathogen interactions, serum–glycoprotein turnover, and innate

immune responses (Lis and Sharon, 1998; Vijayan and Chandra, 1999). Although originally isolated from plant seeds, they are now known to be ubiquitously distributed in nature (Sharon and Lis, 1989). As their role in several cellular processes has begun to increasingly become evident, there has been simultaneous progress in various areas of lectin biology and chemistry (e.g., Gabius *et al.*, 2002; Goldstein, 2002; Hirabayashi, 2004; Brinda *et al.*, 2005; Jeyaprakash *et al.*, 2005; Liu and Rabinovich, 2005). The successful completion of several genome projects has made amino acid sequences of several lectins available. The vast amount of data on lectins, as in many other classes of proteins, makes it difficult and time consuming to simultaneously comprehend all the various advances, especially because the sources of these various pieces of data are highly diverse as also the areas of specialization they emerge from. To achieve effective use of all the data toward understanding of the function and for any possible application, an organization of these seemingly independent data into a common framework is essential. This can be made possible by collating and integrating these various bits of data using an appropriate framework.

Amino acid sequences of lectins and their tertiary structures where available provide a good framework upon which all other data can be integrated, also enabling the pursuit of the ultimate goal of understanding these molecules at the atomic level. They also provide a basis for a unique classification of this class of proteins (Peumans *et al.*, 2001; Bettler, Loris and Imberty, <http://www.cermav.cnrs.fr/lectines/>). Furthermore, chemical and biological data on lectins, as in the case of any family of proteins, when interpreted through their tertiary structures provide the greatest insight into their function and their role in biological systems (Sharma and Surolia, 1997; Vijayan and Chandra, 1999; Chandra *et al.*, 2001; Raval *et al.*, 2004). The vast number of sequences, a significant amount of biochemical data as well as several crystal structures reported in literature, in fact, necessitate a simultaneous analysis of all known members of the family to develop a broader perspective of the functionalities as well as potential uses of these lectins.

Plant lectins have been for decades, model systems of choice to study the molecular basis of these recognition events, because they are not only easy to purify but also easy to exhibit a wide range of carbohydrate specificities, despite strong sequence conservation (Loris *et al.*, 1998; Rudiger, 1998). An integrated knowledge base together with appropriate analytical tools has therefore been developed initially for plant lectins by collating, integrating, classifying primary data, and generating different derived data from several perspectives.

<sup>1</sup>To whom correspondence should be addressed; e-mail:  
[nchandra@serc.iisc.ernet.in](mailto:nchandra@serc.iisc.ernet.in)

## Materials and methods

### Database schema

A relationally organized database schema was designed to serve as a repository of lectins. The database was implemented using MySQL (<http://www.mysql.com>), a free-to-use RDBMS, whereas the scripts were written in PERL. CGI (<http://www.perl.org>) and Java applets for some modules were used for client–server programming, and a Linux web server was used to deliver the interface.

The schema has been designed to accommodate basic information about a lectin, its corresponding sequence and structural details, fold, family classification, and carbohydrate specificity and also enables easy addition of new information in the future. The database schema can also support addition of information on lectins from other sources such as animal, fungi, bacteria, and viruses that are planned to be integrated in the future. Derived data features such as domain boundaries, active site residues, structure prediction, fold classification, and phylogenetic results are stored in various file formats and are processed and accessed through PERL scripts.

### Database construction

A flowchart depicting the methodology used in constructing Lectindb is illustrated in Figure 1. The pipeline to construct the database was automated in parts but was also

manually checked at specific stages, to ensure minimizing of errors in the database.

### Data collection

Initially, a search with keywords “lectin” or “agglutinin” with “viridiplantae” as source limits was carried out on all protein databases available through National Center for Biotechnology Information (NCBI), which resulted in an initial list of 3272 sequence entries from 259 plants. To remove cross-database (Xref) redundancy, we repeated searches with the same keywords but with an additional source keyword pertaining to each of the plants obtained in the first list, using GenBank (Benson *et al.*, 2005; <http://www.ncbi.nlm.nih.gov/>), Swiss-Prot (Bairoch and Boeckmann, 1992), and Protein Data Bank (PDB) (Berman *et al.*, 2000) individually as the database limits, available through the NCBI search engines. These searches were carried out through the Internet and results downloaded onto local machines, along with other available associated details (Fasta, GenPept, view graph format, PubMed entries as well as Swiss-Prot/Trembl “.dat” files from Swiss-Prot). Unless otherwise explicitly stated, all further processing was carried out locally using default parameters.

The resulting list containing 1064 entries was then verified for uniqueness of each entry by carrying out BLAST (Altschul *et al.*, 1990) sequence searches within the data set to find sequences that were 100% identical to any other in the data set (Table I). Any previously unidentified redundancies

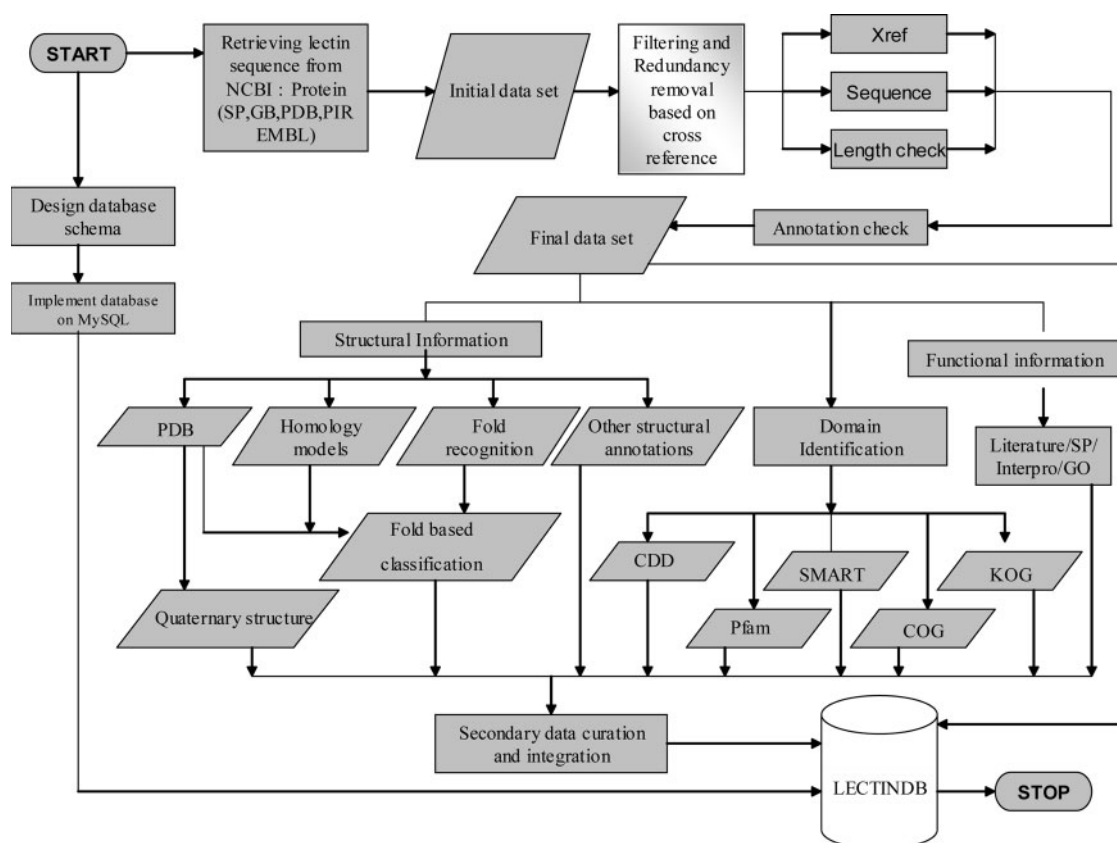


Fig. 1. A schematic representation of the pipeline used in the development of Lectindb.

**Table I.** Various statistics pertaining to Lectindb construction and content

|                                                                                                                                                 |  |             |          |                                    |                                 |               |
|-------------------------------------------------------------------------------------------------------------------------------------------------|--|-------------|----------|------------------------------------|---------------------------------|---------------|
| Data set construction                                                                                                                           |  |             |          |                                    |                                 |               |
| Primary set (lectin sequences from PIR/PRF/PDB/DDBJ/EMBL/GenBank/Swiss-Prot)                                                                    |  |             |          | 3272                               |                                 |               |
| Filtered set (number of entries after filtering)                                                                                                |  |             |          |                                    |                                 |               |
| Xref redundancy                                                                                                                                 |  |             |          | 1064                               |                                 |               |
| Length criteria                                                                                                                                 |  |             |          | 999                                |                                 |               |
| Sequence redundancy                                                                                                                             |  |             |          | 966                                |                                 |               |
| Manual inspection of annotation                                                                                                                 |  |             |          | 947                                |                                 |               |
| Final set (Lectindb)                                                                                                                            |  |             |          | 947                                |                                 |               |
| Number of plants in the final set                                                                                                               |  |             |          | 241                                |                                 |               |
| Domain annotation                                                                                                                               |  |             |          |                                    |                                 |               |
| Total number of domains in Lectindb                                                                                                             |  |             |          | 3583                               |                                 |               |
| Total/unique domains identified from                                                                                                            |  | Cdd: 842/12 | Cog: 1/1 | Kog: 318/17                        | Smart: 809/27                   | Pfam: 1613/35 |
| Total number of domains: sum of unique domains in all domain databases                                                                          |  |             |          |                                    | 92 (contains 21 lectin domains) |               |
| Unique domains from total number of domains                                                                                                     |  |             |          |                                    | 63 (contains 11 lectin domains) |               |
| Number of lectin sequences with multiple domains                                                                                                |  |             |          | 538                                |                                 |               |
| Structural annotation                                                                                                                           |  |             |          |                                    |                                 |               |
| Lectins whose structure is available through PDB                                                                                                |  |             |          | 71                                 |                                 |               |
| Lectins whose structure can be built by homology modeling                                                                                       |  |             |          | 433                                |                                 |               |
| Lectins or domains in lectin containing proteins, whose structure can be predicted (in part or full) by fold recognition                        |  |             |          | 399                                |                                 |               |
| Quality of prediction (as judged by mGenthrader)                                                                                                |  |             |          | Certain: 307, high: 49, medium: 43 |                                 |               |
| Lectin containing proteins with no structural annotation                                                                                        |  |             |          | 44                                 |                                 |               |
| Distribution of other domains co-occurring with lectins                                                                                         |  |             |          | 256                                |                                 |               |
| Functional annotation                                                                                                                           |  |             |          |                                    |                                 |               |
| Number of lectins with annotation from                                                                                                          |  |             |          |                                    |                                 |               |
| InterPro                                                                                                                                        |  |             |          | 820                                |                                 |               |
| GO                                                                                                                                              |  |             |          | 575                                |                                 |               |
| PRINTS                                                                                                                                          |  |             |          | 115                                |                                 |               |
| PRODOM                                                                                                                                          |  |             |          | 364                                |                                 |               |
| PROSITE                                                                                                                                         |  |             |          | 653                                |                                 |               |
| Number of plants with known carbohydrate specificity                                                                                            |  |             |          | 481                                |                                 |               |
| Number of plants with known carbohydrate specificity for one or more lectins in them (that are not resolved as yet to the individual sequences) |  |             |          | 56                                 |                                 |               |
| Number of lectins predicted to have a signal peptide                                                                                            |  |             |          | 492                                |                                 |               |

were removed at this stage. The list of sequences obtained at this stage was manually inspected to ensure that only those sequences belonging to lectins were retained in the database, based on their primary annotations in their respective databases. The filtering step also removed peptide fragments of lectins that were <50 residues in length, which resulted in a final set of 947 sequence entries. This list includes isolectins or even the two chains of heterodimers as separate entries. For each entry, basic information pertaining to the lectin name, amino acid sequence length, molecular weight, carbohydrate specificity, source, protein, and PDB identifiers was parsed from the respective entries in various databases.

#### *Curation of derived data*

Several additional features have been derived for each of the lectin sequences compiled as described above, using

well-established methods. Domains contained in each of the entries were obtained by scanning locally downloaded databases of CDD (Marchler-Bauer *et al.*, 2002), SMART (Schultz *et al.*, 1998), Pfam (Bateman *et al.*, 2004), and KOG/COG databases (Tatusov *et al.*, 2003) using RPS-BLAST (cutoffs employed were 30% identity and 60% coverage in length of the known domain sequence). Alignments of isolectins within a given plant were obtained using Clustalw package (Thompson *et al.*, 1994), followed by a phylogenetic analysis using the MEGA toolkit (Kumar *et al.*, 2004). Information pertaining to sequence neighbors and active site residues identified by the sequence annotated by structure (SAS) server (Milburn *et al.*, 1998; <http://www.ebi.ac.uk/thornton-srv/databases/sas/>), based on related structures in PDB, was also obtained. In addition to the precompiled analyses results, dynamic links have been provided for obtaining sequence neighbors for

each of the lectin entries against NR, PDB as well as within Lectindb.

The experimentally derived structures, where available, were obtained from the PDB. The remaining entries were subjected to BLAST against PDB to find appropriate templates (using cutoffs of 30% identity and 60% coverage of the length of the hit) for building comparative models. For those cases where homologous proteins in PDB were identified, a fold was assigned to either the whole lectins or domains in them. Alignments from the highest scoring templates were obtained, and models were built for these proteins. Fold recognition based on Threader (Jones *et al.*, 1999; <http://bioinf.cs.ucl.ac.uk/psipred/>) and the obtained hits that have been classified as either certain or high were further inspected. A structure-based classification scheme (Lo Conte *et al.*, 2000) was then applied to the different lectins, searchable through the “fold class” query. For the rest of the lectin entries, a separate secondary structure prediction (<http://npsa-pbil.ibcp.fr/>) was attempted.

Quaternary structure information obtained from PDB or in a few cases from the PQS (Henrick and Thornton, 1998) server, wherever available, has been mapped onto the respective lectin entries. The sequences have been scanned for possible signal peptides using SignalP (Bendtsen *et al.*, 2004). Functional information pertaining to carbohydrate and blood group specificities obtained through literature searches for each of the entries has been integrated into the database. Function annotations from Swiss-Prot/InterPro/Gene Ontology (Apweiler *et al.*, 2000; Harris *et al.*, 2004; The GO Consortium), where such information was available, have also been included. Access to various other resources such as Blink, reference to original entries in the primary databases such as PubMed/NCBI-protein/PDB has been made available through the web interface.

## Results

### Database content

Lectindb provides a single point for accessing diverse information pertaining to plant lectins, because it integrates information from Swiss-Prot, GenBank nucleic acid and protein sequence databases, Blink-sequence distribution across species (<http://www.ncbi.nlm.nih.gov/sutils/static/blinkhelp.html>), taxonomy database (<http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html>), PDB-protein structure database, SCOP-structural classification database (Lo Conte *et al.*, 2000), CDD, Pfam, SMART, KOG-domain databases, SAS-structure annotated sequence database, and quaternary structure information from PDB. Furthermore, multiple alignments, structure predictions through homology modeling, and fold recognition, for those lectins whose structures are not available in PDB, have also been carried out and integrated into Lectindb. Besides these, functional information from literature has been fetched for several lectins. The functional information pertaining to carbohydrate specificities, blood group specificities, and biological processes have been mapped from ontology information available through the InterPro and GO resources in many cases. An easy-to-navigate web interface has been built to allow the user to search through the

database, which offers a variety of search options. For example, the user can retrieve information based on a source, domain, fold class, PDB identity, quaternary structure, or carbohydrate specificity. In addition, a sequence-based search option that scans a user-input query sequence for similarities within Lectindb has also been incorporated using a Blast interface.

In total, 947 entries were obtained as unique, nonredundant entries spread across 241 different plant sources (Table I). These plants were found to encompass a wide range of both monocot and dicotyledonous varieties (Figure 2). *Arabidopsis thaliana* and *Oryza sativa* contained the largest number of lectin sequences summing up to 196 and 87, respectively, which is not surprising because their entire (or nearly complete) genome sequences have been deciphered. Some other plants such as *Allium sativum* and *Viscum album* also contained several lectins, most of them closely related to each other, thus qualifying as isolectins. An example snapshot of the database querying is shown in Figure 3A. Multiple sequence alignments of all lectins within each plant help in identifying the similarities and the subtle but perhaps important differences among them. A portion of such an alignment in *A. sativum* is shown as an example (Figure 3B). While many lectins were made of single domains, several others contained lectin domains as part of their larger polypeptide sequences that contained several other domains as well [e.g., lectins from *O. sativa* and *Zea mays* contained protein kinase and epidermal growth factor (EGF) domains in combination with the bulb lectin (B-lectin) domains]. In yet other cases, multiple copies of the same lectin or two or three different lectin domains were also found (e.g., jacalin repeats were present in *A. thaliana* and *O. sativa*). The distribution of number of domains across the entries in Lectindb, illustrated in Figure 4A, reveals that majority of the sequences contained two or three domains, while there were about 23 sequences that contained >12 domains in a single polypeptide chain, of which one or more are lectin domains. A total of 3583 domains were identified in the data set of 947 sequences, which contained 63 unique domain types, of which there were 11 unique lectin domains (Table I). A snapshot of the webpage illustrating the domain architecture of one of the entries in the database is shown in Figure 4B. Identification of domains obtained through a comprehensive analysis, integrated into the database, will help greatly in understanding domain architectures and their functional implications. Understanding lectin domains that coexist either with other lectin domains or with any other proteins provides significant insight into their larger biological roles.

Precompiled lists of sequence neighbors for each of the lectin entries in the database, determined through BlastP searches, have been made available on the database, besides which dynamic links to repeat such analysis with updated databases have also been provided. Most sequences have several good homologs ( $E < 0.05$ ) from plants in the same phyla. Such high conservation within the respective phyla indicates the functional importance of the lectin molecules. Diversity in sequence and as a corollary in function, observed among different phyla, also suggests how specific lectin types may have evolved to suit individual phyletic requirements. Sequence neighbors within the lectin database,

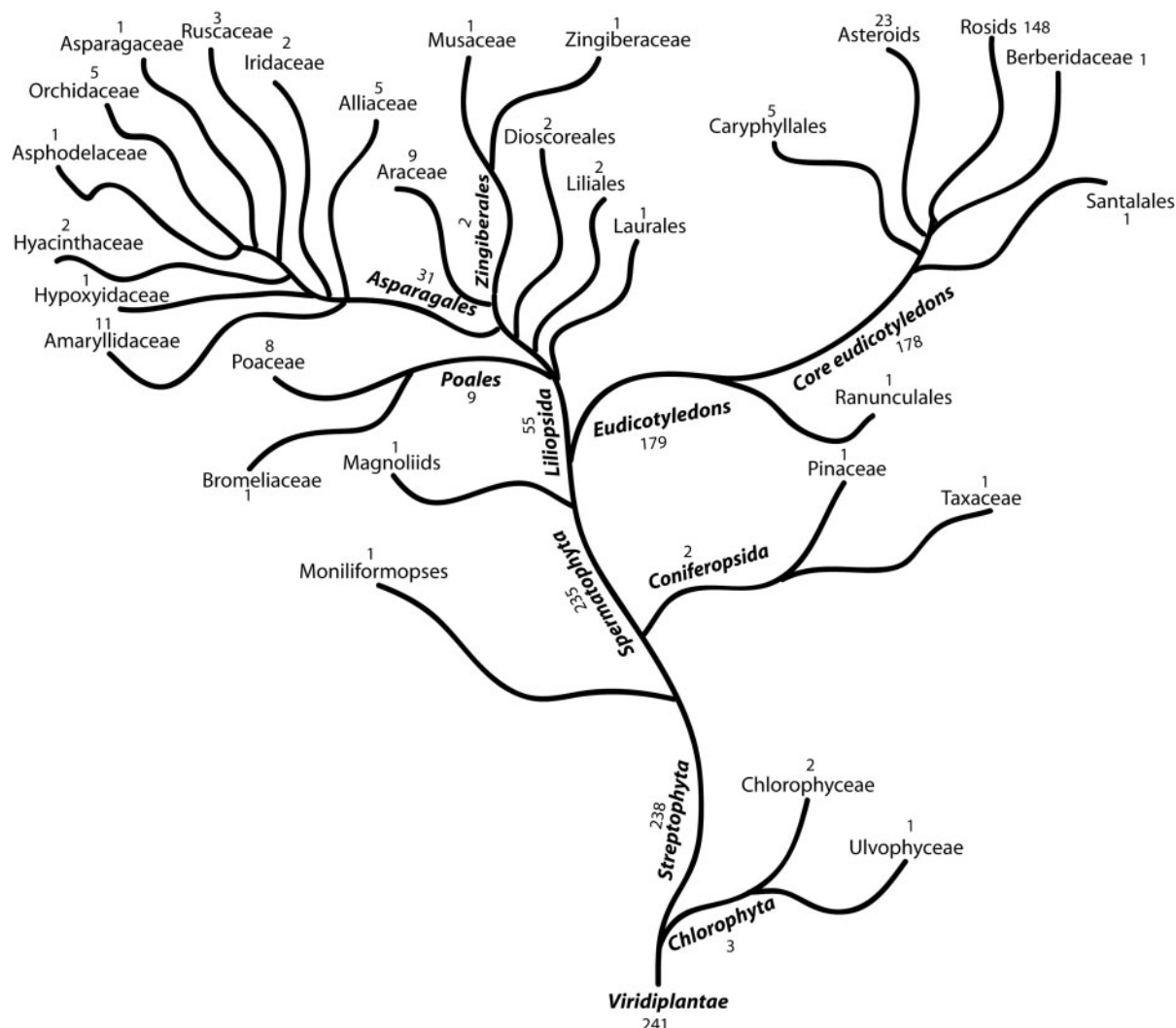


**Fig. 2.** A taxonomical view of the distribution of known lectins in various plant families. The tree has been prepared based on the lectins collated in the database along with the taxonomical information obtained from the NCBI databases. The numbers indicate the number of known plant species containing lectins that are in the database.

identified through BlastP, have also been made available in the database.

Furthermore, each lectin entry in the database has been tagged with structural annotation in a layered fashion, depending upon the extent of information available about them (Figure 3A). Of the 947 entries, 71 polypeptide chains, referring to lectins from 43 plants, had an experimentally derived three-dimensional structure, available from PDB (corresponding to 251 entries in PDB). The first layer encompasses such sequences for which experimentally determined structures are available through PDB, followed by those sequences which have homologous sequences in PDB, and hence can be modeled by homology modeling, followed by sequences which have lesser similarity to any sequence in PDB but which have structure profiles or folds that can be identified by fold recognition methods (Table I). This is followed finally by those sequences that have even more indirect structural information such as through

the identification of motifs and patterns. Annotations also contain information about secondary structure predictions, where appropriate. For about 44 lectin sequences, no structural annotation has been possible with the existing databases, which have been indicated appropriately. A fold-based classification, utilizing the fold information of either the known lectin structures or the high-confidence structural templates of lectin sequences, from SCOP database, has been carried out (Figure 5A and B). This classification scheme groups known lectin types into 7-folds, which are Con-A like  $\beta$ -sandwich,  $\beta$ -Prism-I,  $\beta$ -Prism-II,  $\beta$ -trefoil, Knottins, double-psi- $\beta$ -barrel, and P-domain of calnexin/calreticulin folds. In addition, fold recognition studies are also suggestive of some plant lectins adopting a 6/7-bladed  $\beta$ -propeller fold and a double-stranded  $\beta$ -helix fold. Furthermore, multiple alignments of lectins within each fold class have been carried out, followed by phylogenetic analyses, which are useful to understand the extent of divergence in detail and

A



| Source information       |                            | Image will be displayed, if available                                               |
|--------------------------|----------------------------|-------------------------------------------------------------------------------------|
| Source                   | Allium cepa                |  |
| Common name              | Onion                      |                                                                                     |
| Taxonomy                 | <a href="#">CLICK HERE</a> |                                                                                     |
| Number of lectin entries | 3                          |                                                                                     |

## Sequence details

| Uniprot Entry          | Secondary accession      | Name                                     | Length | Molecular weight | BLINK                    | BLAST against NR       | Protein Sequence         | DNA Sequence             |
|------------------------|--------------------------|------------------------------------------|--------|------------------|--------------------------|------------------------|--------------------------|--------------------------|
| <a href="#">Q38687</a> | <a href="#">AAC37359</a> | mannose specific lectin                  | 164    | 17537            | <a href="#">404658</a>   | <a href="#">BLAST!</a> | <a href="#">Sequence</a> | <a href="#">L12171</a>   |
| <a href="#">Q679P2</a> | <a href="#">AAR23522</a> | mannose-binding lectin precursor         | 110    | 12418            | <a href="#">38532555</a> | <a href="#">BLAST!</a> | <a href="#">Sequence</a> | <a href="#">AY376826</a> |
| <a href="#">Q2XT42</a> | <a href="#">ABB54692</a> | mannose-binding insecticidal leaf lectin | 110    | 12478            | <a href="#">80978888</a> | <a href="#">BLAST!</a> | <a href="#">Sequence</a> | <a href="#">DQ255944</a> |

[MULTIPLE ALIGNMENT](#) [PHYLOGENETIC TREE](#)

To download all sequences in FASTA format [CLICK HERE](#)

## Structural information

| Uniprot entry          | Secondary accession      | PDB | Homolog                                           | THREADER | Fold Class (from SCOP) | Secondary structure                 | BLAST against PDB      |
|------------------------|--------------------------|-----|---------------------------------------------------|----------|------------------------|-------------------------------------|------------------------|
| <a href="#">Q38687</a> | <a href="#">AAC37359</a> | -   | <a href="#">1K71</a><br><a href="#">Alignment</a> | -        | BETA-PRISM II          | <a href="#">Secondary structure</a> | <a href="#">BLAST!</a> |
| <a href="#">Q679P2</a> | <a href="#">AAR23522</a> | -   | <a href="#">1K71</a><br><a href="#">Alignment</a> | -        | BETA-PRISM II          | <a href="#">Secondary structure</a> | <a href="#">BLAST!</a> |
| <a href="#">Q2XT42</a> | <a href="#">ABB54692</a> | -   | <a href="#">1K71</a><br><a href="#">Alignment</a> | -        | BETA-PRISM II          | <a href="#">Secondary structure</a> | <a href="#">BLAST!</a> |

## Functional information

| Uniprot entry          | Secondary accession      | Domain annotation                                                                                                                                                                                                                                                      | SAS                | Carbohydrate specificity (obtained from literature) | SignalP |
|------------------------|--------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------|-----------------------------------------------------|---------|
| <a href="#">Q38687</a> | <a href="#">AAC37359</a> | <a href="#">smart00108</a> Bulb-type mannose-specific lectin<br><a href="#">pfam01453</a> Bulb-type lectin D-mannose binding lectin<br><a href="#">cd00028</a> Bulb-type mannose-specific lectin<br><a href="#">Graphical representation</a> <a href="#">Alignment</a> | <a href="#">SS</a> | Mannose                                             | Yes     |
| <a href="#">Q679P2</a> | <a href="#">AAR23522</a> | <a href="#">smart00108</a> Bulb-type mannose-specific lectin<br><a href="#">pfam01453</a> Bulb-type lectin D-mannose binding lectin<br><a href="#">cd00028</a> Bulb-type mannose-specific lectin<br><a href="#">Graphical representation</a> <a href="#">Alignment</a> | <a href="#">SS</a> | Mannose                                             | No      |

B

|    |            |     |                                                                    |
|----|------------|-----|--------------------------------------------------------------------|
| qb | AAB64237.1 | 41  | YAGQSLDVEQVKIMQDCNVLVLYEYSTPIWASNTGVTGNGCRVLMCRDGNFVYDVNGCPVWASN   |
| gb | AAB64238.1 | 41  | YAGQSLDVEQVKIMQDCNVLVLYEYSTPIWASNTGVTGNGCRVLMCRDGNFVYDVNGCPVWASN   |
| qb | AAR23523.1 | 11  | YAGQSLDVEQVKIMQDCNVLVLYEYSTPIWASNTGVTGNGCRVLMCRDGNFVYDVNGCPVWASN   |
| gb | AAW48531.1 | 13  | YAGQSLDVEQVKIMQDCNVLVLYEYSTPIWASNTGVTGNGCRVLMCRDGNFVYDVNGCPVWASN   |
| qb | AAA32643.1 | 39  | YAGQSLDVEFYHIMQDCNVLVLYDHSSTAVWTSNTDIPGKKGCKAVLQSDGNFVYDAEGRSLWASH |
| gb | AAA32645.1 | 35  | YAGQSLDVEFYHIMQDCNVLVLYDHSSTAVWTSNTDIPGKKGCKAVLQSDGNFVYDAEGRSLWASH |
| qb | AAA32644.1 | 28  | YAGQSLDVEFYHIMQDCNVLVLYDHSSTAVWTSNTDIPGKKGCKAVLQSDGNFVYDAEGRSLWASH |
| gb | AAA32647.1 | 193 | YAGQSLDVEFYHIMQDCNVLVLYDHSSTAVWTSNTDIPGKKGCKAVLQSDGNFVYDAEGRSLWASH |
| qb | AAA32649.1 | 187 | YAGQSLDVEFYHIMQDCNVLVLYDHSSTAVWTSNTDIPGKKGCKAVLQSDGNFVYDAEGRSLWASH |
| gb | AAA32646.1 | 197 | YAGQSLDVEFYHIMQDCNVLVLYDHSSTAVWTSNTDIPGKKGCKAVLQSDGNFVYDAEGRSLWASH |
| qb | AAA32648.1 | 187 | YAGQSLDVEFYHIMQDCNVLVLYDHSSTAVWTSNTDIPGKKGCKAVLQSDGNFVYDAEGRSLWASH |
| gb | AAB64239.1 | 165 | SEEPFNTQQFKLAAQDCNVLVLYEYNIVIWASNTVVGAAHAGCRAMQVDCNFEVTEFNLHATSSH  |

**Fig. 3. (A)** A snapshot illustrating an example of a source-based search of Lectindb. The information is grouped into three categories: (1) sequence-related, (2) structure-related, and (3) function-related. The information retrieved for *Allium cepa* indicates that there are three lectins from this source in the database, whose details are listed. It also shows that the structures of these proteins have not been determined experimentally and hence not available from PDB, but its sequence has close homology to a protein with known structure (which in this case turns out to be garlic lectin), and hence, the structure can be built using homology modeling techniques. The retrieved functional information indicating that these lectins are mannose specific and bind to rabbit erythrocytes is also shown. **(B)** A snapshot of a portion of the multiple alignment of isolectins from a given plant species, which in the example shown, is from *Allium sativum*. Identical and highly conserved stretches are highlighted in dark and light grey, respectively.

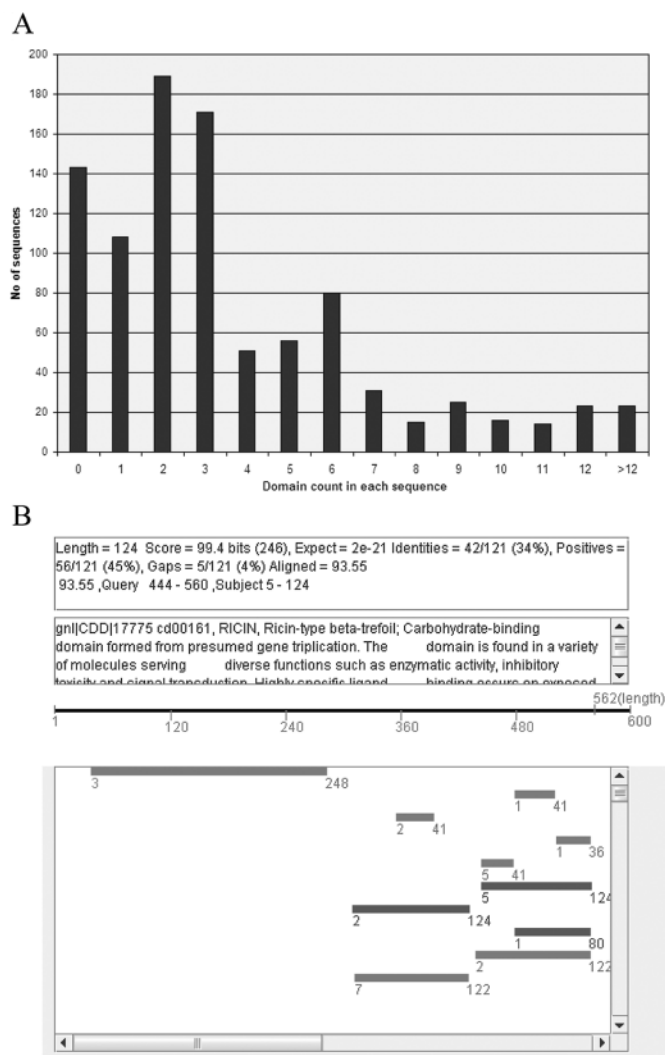


**Fig. 4.** (A) The distribution of number of domains across the database. The chart shows the numbers of lectin entries in the database with 1, 2, 3,...12, and >12 nonoverlapping domains as identified by one or the other domain databases. The column marked 0 refers to those sequences, which do not contain a domain recognized by any of the domain databases used here. It however does not rule out the presence of one or more domains in these lectins that may be detected in the future through further updated databases or through other bioinformatics approaches. (B) A sample snapshot indicating the domain architecture of a typical entry in the database. Different shades displayed by the applet indicate that they have been derived from different sources. Respective alignments and comments can also be displayed through the top panels.

hence subtle but definite functional differences/adaptations, within each fold. Beyond the overall fold information, the structural annotations also comprise information regarding quaternary structure of the different lectins in the database.

The next level of information in the database pertains to the known function(s) of the lectins. Here again, the information spans a wide hierarchical range, starting from individual monosaccharide specificities to larger roles in various cellular events. Putative signal peptides have been identified using SignalP. Of the 947 lectin entries, 492 of them were strongly predicted to contain a signal peptide. Analysis of the lectins in the databases using SAS bioinformatics tool

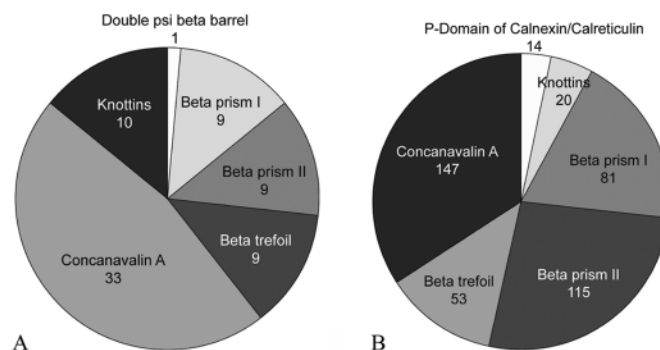


**Fig. 5.** A glimpse of the structural annotation of entries in Lectindb. The fold profile of structures in the first two categories in Lectindb indicating the distribution of the lectin entries across different folds: (A) fold profile for lectins with known structure and (B) major fold profile for lectins built by homology modeling. Numbers in each sector indicate the total number of lectin entries belonging to that fold.

helped in identifying the conservation of active residues in each class of lectins. Functional annotations of the lectins have also been derived from Swiss-Prot and GenBank function cards as well as from InterPro and GO. For example, a 109 amino acid putative secretory protein from Aloe species (P49329) has been annotated as a mannose-specific lectin, containing a B-type lectin domain, which shows agglutinating activity toward rabbit erythrocytes and mitogenic activity toward mouse lymphocytes. It is also annotated to possibly be either a homotrimer or a homotetramer. Its similarity to curculin-like mannose-binding proteins as well as sequence motifs indicating selective binding to carbohydrates is also included in the annotation. Carbohydrate specificities obtained from the literature have often pertained to lectins in general of a given plant rather than to a particular lectin of that plant. For example, a collection reported by Wu and others (2001) indicates that lectins from *Allium ursinum* are mannose specific. In these situations, where a detailed mapping to every sequence has not been possible, a general functional annotation to lectins of a given plant has been provided.

## Utility and discussion

Lectindb provides an easy-to-use web interface with flexibility to select for an entry or a collective set of entries matching users' criteria such as name of the plant, sequence class, domain type, structural class, quaternary type, fold class or carbohydrate specificity, and keyword searches. An example analysis in a set of 167 jacalin-related lectin domains carried out by us recently showed that hypervariability in the binding site loops generates carbohydrate recognition diversity, a strategy analogous to that in legume lectins (Raval *et al.*, 2004). Analyses of the size, conformation, and sequence variability in key regions in this data set revealed the existence of a common theme, encoded as a set of structural features over a common scaffold, in defining specificity.

Lectins bind to various sugars in a highly selective manner. This selectivity enables lectins to display many significant biological activities such as blood group-specific agglutination,

preferential agglutination of tumor cells, and a variety of other functions. Molecular recognition is indeed a key event in many biological processes. It is also known that almost all cells carry carbohydrates on their surfaces, providing ready recognition sites for various proteins interacting with them, of which lectins form an important class. Hence, studies with lectins have proved invaluable in the understanding of molecular mechanisms of various cellular processes and deciphering the code contained within the sugar molecules. In the laboratory, lectins are also attractive biotechnological tools because they are highly stable, exquisitely specific for carbohydrate determinants and amenable to chemical modification and conjugation.

Some of the functions of lectins and the broad potential applications they lead to are given below: (1) Agricultural biotechnology: in plant defense as insecticides; (2) Clinical diagnosis I: diagnosis of diseases such as cancer (e.g., peanut lectin and jacalin can specifically recognize a surface carbohydrate expressed on cancer cells); as a useful surrogate marker for qualitative and quantitative deficiency of CD4+ T cells in human immunodeficiency virus-1 (HIV-1) infection (e.g., jacalin); (3) Clinical diagnosis II: in blood typing (different lectins agglutinate different blood group carbohydrates); (4) Laboratory use: molecular biology and biotechnology (tools for protein purification, e.g., in affinity chromatography), analysis of oligosaccharide structure; (5) Clinical use: treatment, mitogenic stimulants, use in transplantation; (6) Immunology studies: immunosuppressive and immunomodulatory properties, inhibition of stimulated T cells, immobilization of Fc and epsilon receptors and as allergens; (7) Understanding/modification of various physiological processes, for example, cell-signaling events (i) understanding physiological processes, (ii) correcting abnormal physiology, and (iii) exploiting modified physiology to advantage in diseased processes; and (8) Drug delivery and targeting (lectin-mediated drug targeting/use of lectins in drug delivery to oral mucosa).

In the pursuit of the above applications, it is our belief that this database will serve as a useful repository of manually curated information pertaining to sequence, structure, and function, all integrated into a single framework (availability: Lectindb is available freely for academic use from <http://nscdb.bic.physics.iisc.ernet.in>, contact [nchandra@serc.iisc.ernet.in](mailto:nchandra@serc.iisc.ernet.in) for further information).

## Acknowledgments

We thank Prof. M. Vijayan and Prof. A. Surolia for constant encouragement and useful discussions. We also thank Prof. C. Kameswara Rao (formerly with Bangalore University) for providing us experimentally derived functional data for some lectins. Financial support from Department of Biotechnology (DBT), Govt. of India is gratefully acknowledged. Use of facilities at the Bioinformatics Centre and Supercomputer Education and Research Centre and interactive graphics facility supported by DBT is also acknowledged.

## Conflict of interest statement

None declared.

## Abbreviations

NCBI, National Center for Biotechnology Information; PDB, Protein Data Bank; SAS, sequence annotated by structure.

## References

- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
- Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M.D., and others (2000) InterPro – an integrated documentation resource for protein families, domains and functional sites. *Bioinformatics*, **16**, 1145–1150.
- Bairoch, A. and Boeckmann, B. (1992) The SWISS-PROT protein sequence data bank. *Nucleic Acids Res.*, **20**, 2019–2022.
- Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L., and others (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138–D141.
- Bendtsen, J.D., Nielsen, H., von Heijne, G., and Brunak, S. (2004) Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.*, **340**, 783–795.
- Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Wheeler, D.L. (2005) GenBank. *Nucleic Acids Res.*, **33**, D34–D38.
- Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
- Brinda, K.V., Surolia, A., and Vishveshwara, S. (2005) Insights into the quaternary association of proteins through structure graphs: a case study of lectins. *Biochem. J.*, **391**, 1–15.
- Chandra, N.R., Prabu, M.M., Suguna, K., and Vijayan, M. (2001) Structural similarity and functional diversity in proteins containing the legume lectin fold. *Protein Eng.*, **14**, 857–866.
- Gabius, H.J., Andre, S., Kaltner, H., and Siebert, H.C. (2002) The sugar code: functional lectinomics. *Biochim. Biophys. Acta*, **1572**, 165–177.
- Goldstein, I.J. (2002) Lectin-structure-activity: the story is never over. *J. Agric. Food Chem.*, **50**, 6583–6585.
- Harris, M.A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., and others (2004) Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.
- Henrick, K. and Thornton, J.M. (1998) PQS: a protein quaternary structure file server. *Trends Biochem. Sci.*, **23**, 358–361.
- Hirabayashi, J. (2004) Lectin-based structural glycomics: glycoproteomics and glycan profiling. *Glycoconj. J.*, **21**, 35–40.
- Jeyaprakash, A.A., Jayashree, G., Mahanta, S.K., Swaminathan, C.P., Sekar, K., Surolia, A., and Vijayan, M. (2005) Structural basis for the energetics of jacalin–sugar interactions: Promiscuity versus specificity. *J. Mol. Biol.*, **347**, 181–188.
- Jones, D.T., Tress, M., Bryson, K., and Hadley, C. (1999) Successful recognition of protein folds using threading methods biased by sequence similarity, and predicted secondary structure. *Proteins*, **3**, 104–111.
- Kumar, S., Tamura, K., and Nei, M. (2004) MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. *Brief. Bioinform.*, **5**, 150–163.
- Lis, H. and Sharon, N. (1998) Lectins: carbohydrate-specific proteins that mediate cellular recognition. *Chem. Rev.*, **98**, 637–674.
- Liu, F.T. and Rabinovich, G.A. (2005) Galectins as modulators of tumour progression. *Nat. Rev. Cancer*, **5**, 29–41.
- Lo Conte, L., Ailey, B., Hubbard, T.J., Brenner, S.E., Murzin, A.G., and Chothia, C. (2000) SCOP: a structural classification of proteins database. *Nucleic Acids Res.*, **28**, 257–259.
- Loris, R., Hamelryck, T., Bouckaert, J., and Wyns, L. (1998) Legume lectin structure. *Biochim. Biophys. Acta*, **383**, 9–36.
- Marchler-Bauer, A., Panchenko, A.R., Shoemaker, B.A., Thiessen, P.A., Geer, L.Y., and Bryant, S.H. (2002) CDD: a database of conserved domain alignments with links to domain three-dimensional structure. *Nucleic Acids Res.*, **30**, 281–283.
- Milburn, D., Laskowski, R.A., and Thornton, J.M. (1998) Sequences annotated by structure: a tool to facilitate the use of structural information in sequence analysis. *Protein Eng.*, **11**, 855–859.

- Peumans, W.J., Van Damme, E.J., Barre, A., and Rouge, P. (2001) Classification of plant lectins in families of structurally and evolutionary related proteins. *Adv. Exp. Med. Biol.*, **491**, 27–54.
- Raval, S., Gowda, S.B., Singh, D.D., and Chandra, N.R. (2004) A database analysis of jacalin-like lectins: sequence-structure-function relationships. *Glycobiology*, **14**, 1247–1263.
- Rudiger, H. (1998) Plant lectins – more than just tools for glycoscientists: occurrence, structure, and possible functions of plant lectins. *Acta Anat (Basel)*, **161** (1–4), 130–152.
- Schultz, J., Milpetz, F., Bork, P., and Ponting, C.P. (1998) SMART, a simple modular architecture research tool: identification of signaling domains. *Proc. Natl. Acad. Sci. U. S. A.*, **95**, 5857–5864.
- Sharma, V. and Surolia, A. (1997) Analyses of carbohydrate recognition by legume lectins: size of the combining site loops and their primary specificity. *J. Mol. Biol.*, **267**, 433–445.
- Sharon, N. and Lis, H. (1989) Lectins as cell recognition molecules. *Science*, **246**, 227–234.
- Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., and others (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4** [Epub 11 September 2003].
- Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994) CLUSTALW: Improving the sensitivity of progressive multiple sequence alignment through sequence weighing, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.
- Vijayan, M. and Chandra, N. (1999) Lectins. *Curr. Opin. Struct. Biol.*, **9**, 707–714.
- Wu, A.M., Song, S., Tsai, M., and Herp, A. (2001) A guide to the carbohydrate specificities of applied lectins. In Wu, A.M. (ed.), *The Molecular Immunology of Complex Carbohydrates*. Kluwer Academic and Plenum Publishers, pp. 551–585.