AnimalLectinDb: An integrated animal lectin database

Dharmendra Kumar[1]* & Yashoda Mittal[2]

[1]Department of Zoology, Government Post Graduate College, Rishikesh, Uttrakhand, India; [2]KLDAV, Post Graduate College, Roorkee, Uttrakhand, India; Dharmendra Kumar - Email: research.bioinformatics@gmail.com; *Corresponding author

Abstract:
Lectins, a class of carbohydrate-binding proteins and widely recognized to play a range of crucial roles in many cell–cell recognition events triggering several important cellular processes encompass different members that are diverse in their protein structures, carbohydrate affinities and specificities, their larger biological roles and potential applications. To attain an effective use of all the diverse data initially an animal lectin database 'AnimalLectinDb' with information pertaining to taxonomic, structural, domain architecture, molecular sequence, carbohydrate structure and blood group specificity has been developed. It is expected to be of high value not only for basic study in lectin biology but also for advanced research in pursuing several applications in biotechnology, immunology, and clinical practice.

Availability: http://www.research-bioinformatics.in

Background:
Lectins are glycoproteins 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 [1]. Although originally isolated from plant seeds, they are now known to be ubiquitously distributed in nature [2]. As their role in several cellular processes has begun to become increasingly evident, there has been simultaneous progress in various areas of lectin biology and chemistry [3].
The successful completion of several genome projects has made amino acid sequences of several lectins available. Amino acid sequences of lectins and their tertiary structures (where available) provide a good framework upon which all other data can be integrated enabling the pursuit of the ultimate goal of understanding these molecules at the atomic level also [4]. They also provide a basis for a unique classification of this class of proteins (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 [5]. The vast numbers of sequences, a significant amount of biochemical data as well as several crystal structures reported in the literature, in fact, demand 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. The number of animal lectins and their biological roles continue to increase with the new researches going on in the field of lectin biology and allied fields. It has also been shown that some of the newly described lectins are similar to previously investigated lectins, whereas others represent new structural groups. Progress has been made in understanding structure-function relationships for several lectins in both the old

and the new categories. In the present work the efforts have been made to develop an integrated knowledge based animal lectins database together with appropriate analytical tools which we named as "AnimalLectinDb".

Materials and Methodology:
Data collection:
Initially, a search with keywords "animal lectin" or "agglutinin" as source limits was carried out on all protein databases available through National Center for Biotechnology Information (NCBI). Repeated searches were made with the same keywords but with an additional source keyword pertaining to each of the animal lectin individually, obtained in the first list through the NCBI search engines, using GenBank, Swiss-Prot [6], and, Protein Data Bank (PDB) [7]. These searches were carried out through the Internet and results were downloaded onto local machines, along with other available associated details. This data was filtered through RDBMS (Relational Database Management System) technique by assigning primary key and foreign key to ensure there is no duplication of data. Unless otherwise explicitly stated, all further processing was carried out locally using default parameters. Any previously unidentified redundancies were removed. For each entry, basic information pertaining to the lectin name, source, protein, amino acid sequence length, molecular weight, carbohydrate specificity, and PDB identifiers was parsed from the respective entries in various databases.

Database schema:
An organized database schema was designed to serve as a repository for animal lectins. The database was implemented using HTML, XML. For some modules JavaScript and Java applets were used. The schema has been designed to

accommodate basic information about lectins viz. the structural details: fold, family classification, primary structure of their proteins, the corresponding nucleotide sequences and also the details of the carbohydrate specificities. Derived data features such as domain boundaries, active site residues, structure prediction, fold classification, and phylogenetic results are stored in various file formats. Many of the links are connected to different databases to acquire more recent data. The database schema also enables easy addition of new information about animal lectins in the future. Moreover, it can also support addition of information on lectins from other sources such as plants, fungi,

bacteria, and viruses that are planned to be integrated in the future as a "Comprehensive Lectin Database".

Database assembly:
A flowchart depicting the methodology used in constructing **AnimalLectinDb** is illustrated in **Figure 1**. The pipeline to construct the database has been automated in parts and also manually checked at specific stages, to ensure of minimizing errors in the database.
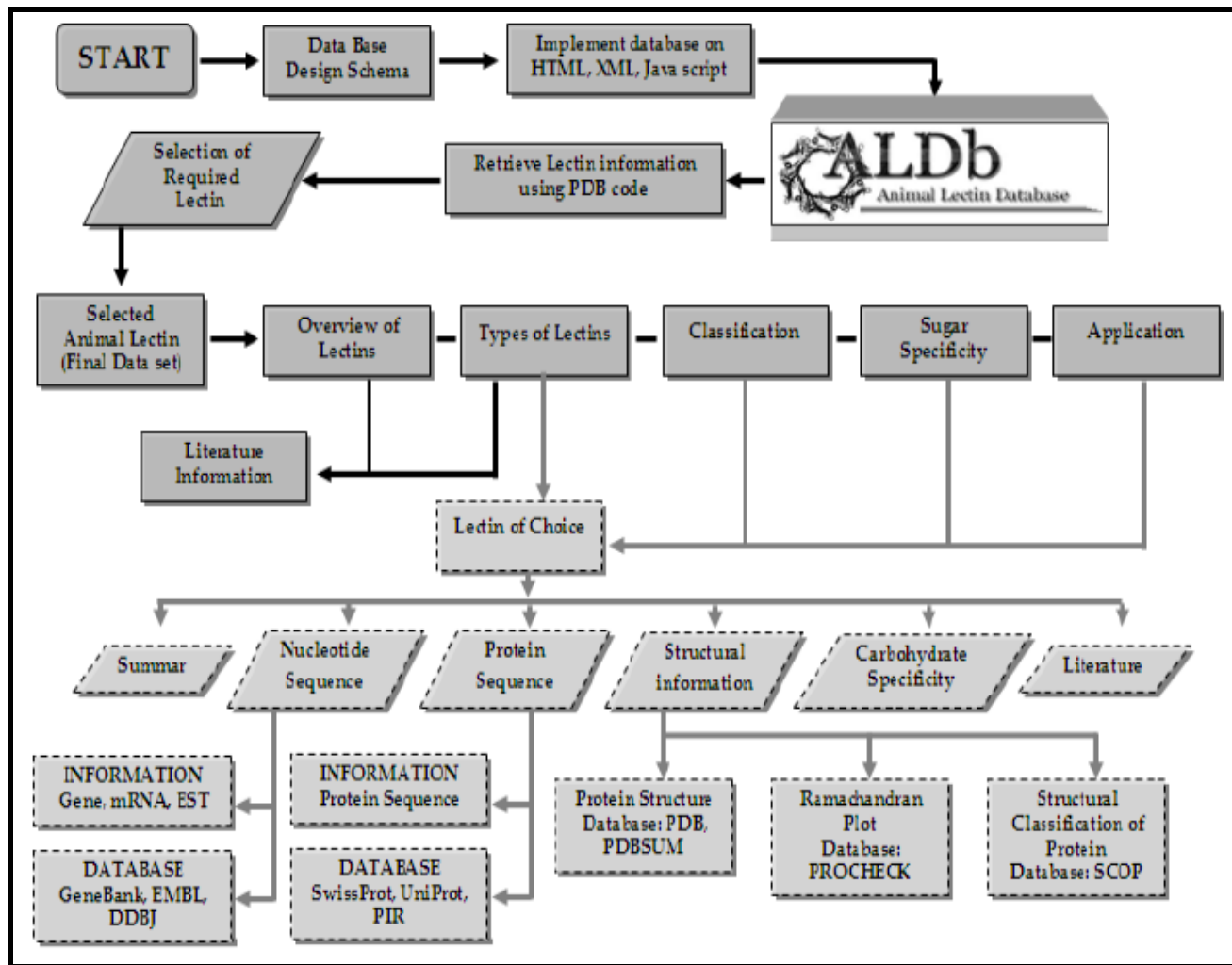

Figure 1: A schematic representation of the pipeline used in the development of AnimalLectinDb.

Results:
Using keywords "animal lectin"/ "agglutinin" search through NCBI resulted in an initial list of animal lectins = 1372 and agglutinin = 887 (**Figure 2**). Further information was collected by repeated searches for each animal lectin individually using different databases. The database assembly was designed as discussed in "Material and Methods". It is then run and contents were analyzed.

Database content:
There are some specific databases which may provide information on animal lectins, but all these databases need to be updated with latest information as per requirement of scientific community. The Animal lectin Database, 'AnimalLectinDb' provides all required information on animal lectins such as Express Sequence Tags and mRNA detail along with complete nucleotide sequence and Protein sequence from different databases. It also provides the microscopic details of protein such as visualization of dihedral angles ψ against φ of amino acid residues in protein structure, ligand information and domain etc, not available in any other lectin database. This database integrates information about lectins such as, Nucleic acid and protein sequence databases from Swiss-Prot, GenBank, EMBL and DDBJ; Taxonomy database; Protein

structure database from PDB; Structural classification databases from CATH, SCOP; Ramachandran Plot from PROCHECK and Interaction with other molecules from MSD. Besides these, it has functional information for all lectins fetched from literature, functional annotations derived from Swiss-Prot and GenBank function cards as well as from Protein Information Resource (PIR). The functional information pertaining to carbohydrate specificities, blood group specificities, and biological processes have been mapped. 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. 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. Carbohydrate specificities obtained from the literature have often pertained to specific animal lectin and a general functional annotation to lectins of a given animal has also been provided. In our database some of the functions of lectins and the broad potential applications they lead have been provided as a different section of the database.
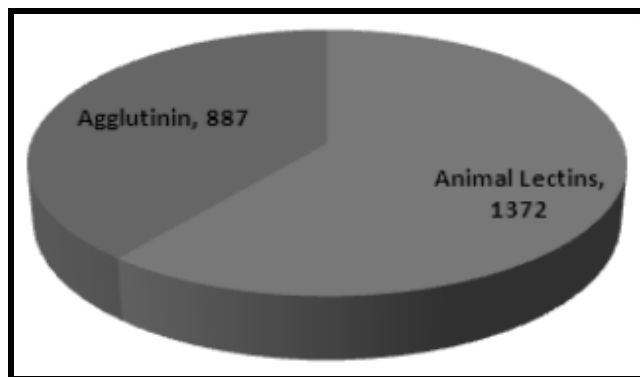
Figure 2: Statistcis of Animal Lectins in Database

Discussion and Utility:
Our studies with analysis of databases for cystic fibrosis and mutational analysis of H5N1 and H1N1 were found significant for ascertaining test systems [8, 9]. 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 them the ready recognition sites for various proteins interacting with them. Lectins form an important class of these carbohydrates containing molecules. 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 [10, 11]. 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 [3]. AnimalLectinDb 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 animal, sequence class etc. In the pursuit of all the applications, it is our belief that this database will serve as a useful repository of manually curetted information pertaining to sequence, structure, and function, all integrated into a single framework similar to our previously developed phyto chemical database for diabetes "Phyto-Mellitus" [12].

Conflict of interest:
The authors have no conflicts of interest to declare.

References:
[1] Sharon N & Lis H. *Science* 1972 **177**: 949 [PMID: 5055944]
[2] Vijayan M & Chandra N. *Curr Opin Struct Biol*. **9**: 707 [PMID: 10607664]
[3] Gabius HJ *et al*. *Biochim Biophys Acta*. 2002 **1572**: 165 [PMID: 12223267]
[4] Sharon N. *Trends Biochem Sci*. 1993 **18**: 221 [PMID: 8346557]
[5] Gourdine JP *et al*. *Fish Shellfish Immunol*. 2007 **23** (4): 831 [PMID: 17493832]
[6] Bairoch A & Boeckmann B. *Nucleic Acids Res*. 1992 **20** Suppl: 2019 [PMID: 1598233]
[7] Berman HM *et al*. *Nucleic Acids Res*. 2000 **28**: 235 [PMID: 10592235]
[8] Kumar D & Mittal Y. *Proc BIOCOMP'07 (CSREA Press)*. 2007 **2**: 727
[9] Mittal Y *et al*. *Journal of Computational Intelligence in Bioinformatics* 2010 **3**(1): 91
[10] Agrawal BB & Goldstein IJ. *Biochem J*. 1967 **96**: 23C–5C [PMID: 5862398]
[11] Hammarstrom S & Kabat EA. *Biochemistry* 1969 **8**: 2696 [PMID: 4241272]
[12] Middha SK *et al*. *Bioinformation* **4**(2): 78 [PMID: 20198174]