W452–W458 *Nucleic Acids Research, 2009, Vol. 37, Web Server issue* *Published online 27 May 2009*
doi:10.1093/nar/gkp409

PLecDom: a program for identification and analysis of plant lectin domains

Smriti Shridhar, Debasis Chattopadhyay and Gitanjali Yadav*

Computational Biology Laboratory, National Institute of Plant Genome Research, Aruna Asaf Ali Marg, New Delhi 110067, India

Received February 16, 2009; Revised April 25, 2009; Accepted May 2, 2009

ABSTRACT

PLecDom is a program for detection of Plant Lectin Domains in a polypeptide or EST sequence, followed by a classification of the identified domains into known families. The web server is a collection of plant lectin domain families represented by alignments and profile Hidden Markov Models. PLecDom was developed after a rigorous analysis of evolutionary relationships between available sequences of lectin domains with known specificities. Users can test their sequences for potential lectin domains, catalog the identified domains into broad substrate classes, estimate the extent of divergence of new domains with existing homologs, extract domain boundaries and examine flanking sequences for further analysis. The high prediction accuracy of PLecDom combined with the ease with which it handles large scale input, enabled us to apply the program to protein and EST data from 48 plant genome-sequencing projects in various stages of completion. Our results represent a significant enrichment of the currently annotated plant lectins, and highlight potential targets for biochemical characterization. The search algorithm requires input in fasta format and is designed to process simultaneous connection requests from multiple users, such that huge sets of input sequences can be scanned in a matter of seconds. PLecDom is available at http://www.nipgr.res.in/plecdom.html.

INTRODUCTION

Modern glycobiology revolves, to a large extent, around the potential biological information stored in cell surface carbohydrates, whose roles in cell growth, differentiation and surface recognition are increasingly being investigated using lectins, a large family of ubiquitous proteins with the ability to bind and agglutinate these sugars. Lectins display an enormous diversity in their sequence, biological activity and mono-/oligosaccharide specificity in addition to an unsurpassed structural versatility (1,2). In plants, lectins play crucial physiological roles in stress responses, defense, symbiotic communication, and are considered one of the most important biological recognition molecules (1,3,4).

Plant lectins have been variously classified into distinct families based on their structure, ligands and evolutionary relationships. Structurally, some of the major folds reported for plant lectins are beta prism-I, beta-prism-II, beta-trefoil, seven-bladed beta-propeller, knottins, jelly-roll (also called the lectin fold) and the P-domain fold of calnexin/calreticulin (5). These structural groups show varying degrees of overlap with the sequence-based classes, and new folds are constantly being identified. Until a few years back, a consensus of seven distinct lectin families were known in plants, based upon their carbohydrate binding domains (6). These were the amaranthins, cucurbitaceae phloem lectins (now called the Nictaba lectins), lectins with hevein domains, jacalin-related lectins, legume lectins, monocot mannose-binding lectins [now called the GNA-related lectins (7)] and type-II ribosome-inactivating proteins (RIPs) also known as the Ricin-B family. However, this classification has become inadequate with the very recent addition of new families or by regrouping of the existing families (8). Likewise, the identification of functional homologs of various animal lectin families in plants (9), as in the case of the galectins (previously called the S-type lectins, having a strong affinity for β-galactosides) and the calnexin/calreticulin lectin families, has further amplified the complexity in plant lectin classification (10,11). Among the newly identified and/or regrouped families, are the agaricus bisporus agglutinins (or ABA domains), class V chitinase homologs with lectin activity (12), the EEA (13), LysM family (14) and cyanovirins (15), all of which emphasize the need to re-address the question of lectin classification in light of new data.

Due to the huge variation in the sequence, structure and specificity of plant lectins and the complexity associated with their classification, methods that seek to improve detection, annotation or assignation of carbohydrate

*To whom correspondence should be addressed. Tel: +91 11 26735103; Fax: +91 11 26741658; Email: gy@nipgr.res.in specificity to these proteins would be of immediate interest to researchers. Currently, the determination of fine specificities of lectins remains largely experimental although attempts to understand the sugar recognition mechanisms within families have demonstrated a huge potential for development of bioinformatics-based predictive/automated tools. Some of the significant computational efforts have involved flexible docking between lectin and sugar molecules, structural mapping and pattern recognition in glycan branches via probabilistic sibling-dependent tree markov models (16–18). Efficient algorithms have been developed for training and improvement of the probabilistic models, but these have been tested on binding affinity data for a limited number of families (19). A comparative structural and specificity analysis led to the identification of three crucial residues for carbohydrate recognition in legume lectins (20), providing insights into the molecular interactions of lectins with simple sugars, involving a network of hydrogen bonds and an aromatic residue in the vicinity of the binding site. Despite these breakthroughs and pioneering work being carried out in lectin biology, there is no dedicated program or tool for identification of these domains. This inspired us to develop PLecDom, an online, predictive and interactive web server that can assist in the identification and analysis of these proteins using sequence information alone. The program has a very simple user friendly interface, and help pages, allowing users to submit their own queries or browse available data.

METHODOLOGY

Data collection

Published reports of characterized sequences assigned to distinct plant lectin families were searched from literature followed by keyword search of the major protein databanks as well as the plant lectin database (21). Homologs of animals origin, wherever available, were also compiled for these families. In all, 845 and 487 sequences, from plant and animal origins, respectively, were compiled for various families. Families for which a minimum of 30 representative sequences were not available were discarded from further analysis. These included the cyanovirins, ABA lectins, LysM, EEA, Amaranthins and class V chitinase homologs with lectin activity. The filtered starting dataset was composed of the eight remaining families, namely, the (i) GNA-related lectin domains, (ii) lectins with Hevein domains, (iii) Jacalin related lectins, (iv) Legume lectins, (v) Ricin–B lectins, (vi) Galectins, (vii) Calreticulin/Calnexins and (viii) Nictaba lectins. This dataset was called the curated sequence (CS) dataset (available as online Supplementary Data). Apart from the CS dataset, protein sequence prediction data was downloaded for 10 completed plant genome projects from NCBI, TIGR (22) and JGI web sites. These include four dicots (Arabidopsis, Poplar, Grape and Soybean), four chlorophytes (Chlamydomonas, Volvox and two species of Ostreococcus), a bryophyte (Physcomitrella) and Rice—a monocot. This data comprised of a total of 357 139 protein sequences and was called the 'Protein Complete Genome' (PCG) dataset. In addition, EST sequence data for 38 incomplete plant genomes, was downloaded from the TIGR database (22). These include eight monocots (Allium, Festuca, Hordeum, Wheat, Maize, Rye, Sorghum and Sugarcane), two gymnosperms (Pinus and Spruce), and 28 dicots including Apple, Aquilegia, Beetroot, Brassica, Capsicum, Cocoa, Coffee, Euphorbia, Sunflower, Ice plant, Ipomoea, Lotus, Medicago, Petunia, Phaseolus, Potato, Prunus, Tomato and two species each of Citrus, Gossypium, Nicotiana, Lactuca and Triphysaria. Full species details and scientific names can be found on the web server. This dataset includes 42% clustered ESTs and 57% individual ESTs and a fraction of ETs, adding to a total of 1 873 460 sequences in all. This was called the 'EST Incomplete Genome' (EIG) Dataset.

Analysis and program development

Multiple alignments and phylogenetic reconstruction of the sequences in the CS dataset were carried out using CLUSTALW (23). Profile HMMs were built using HMMER version 2.3.2 (24). The first set of profile HMMs were built using plant sequences and were trained on the animal data to strengthen the predictability and to enable identification of distant homologs. Different build-and-search parameters were tested, allowing only one parameter change at a time. The GNA-related lectins and calreticulin sequences gave optimal results when weighed using the Krogh/Mitchinson maximum entropy algorithm, a slightly more robust form of the Eddy/Mitchinson/Durbin maximum discrimination algorithm, giving an increase in sensitivity. For all other families, the default weighing method, i.e. the Gerstein/Sonhammer/Chothia tree-weighting algorithm gave best results. The BLOSUM62 scoring matrix was used for all families except the jacalins and legume lectins, both of which performed better with a heuristic PAM60 matrix. In the second step, the plant and animal sequences from the CS dataset were combined and profiles were rebuilt using parameters optimized in the first stage. This strategy offered us a larger dataset and greater representation of each family, and was thenceforth treated as the training dataset, having a total of 1198 sequences from eight families described in the previous section. For testing EST sequences, which often contain partial domains rather than full domains, the fragmentary search option was used in addition to the optimized parameters described earlier, to enable short fragments of target domains to be captured. An *E*-value filter of 0.01 was applied when using these HMMfs profiles. The EIG dataset was subjected to a six-frame translation using EMBOSS version 6.0.1 (25). For protein sequence search, default *E*-value was used with the optimized profile HMMs. Domain boundaries were identified using the alignments of sequences with profiles so that individual domains could be extracted for further analysis. Database match with query was done using a local version of BLAST program (26). In-house Fortran programs were used to streamline and automate the entire process and shell scripts were added for the testing of input sequences

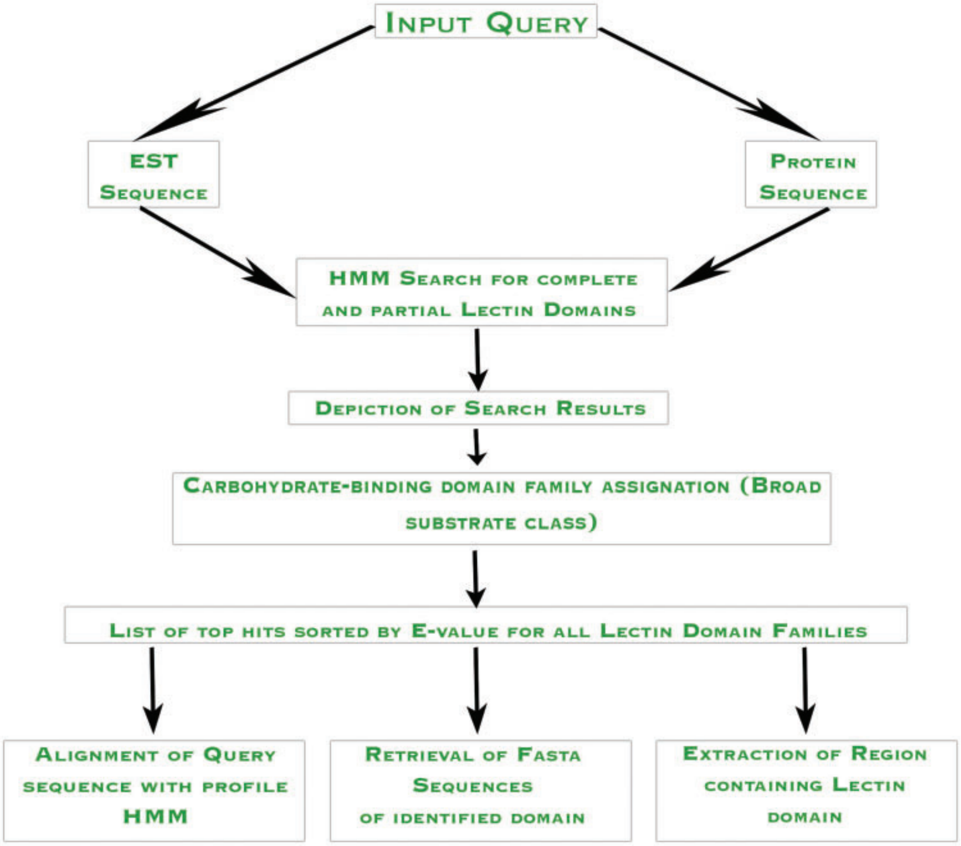

Input Query
EST Sequence
Protein Sequence
HMM Search for complete and partial Lectin Domains
Depiction of Search Results
Carbohydrate-binding domain family assignation (Broad substrate class)
List of top hits sorted by E-value for all Lectin Domain Families
Alignment of Query sequence with profile HMM
Retrieval of Fasta Sequences of identified domain
Extraction of Region containing Lectin domain

Figure 1. Flowchart depicting the query submission protocol in PLecDom.

and to allow multiple users to use the program simultaneously. The program thus developed was converted into a web server using html and back-end CGI coding. Figure 1 depicts a schematic overview of the PLecDom query submission protocol.

Program testing and prediction accuracy

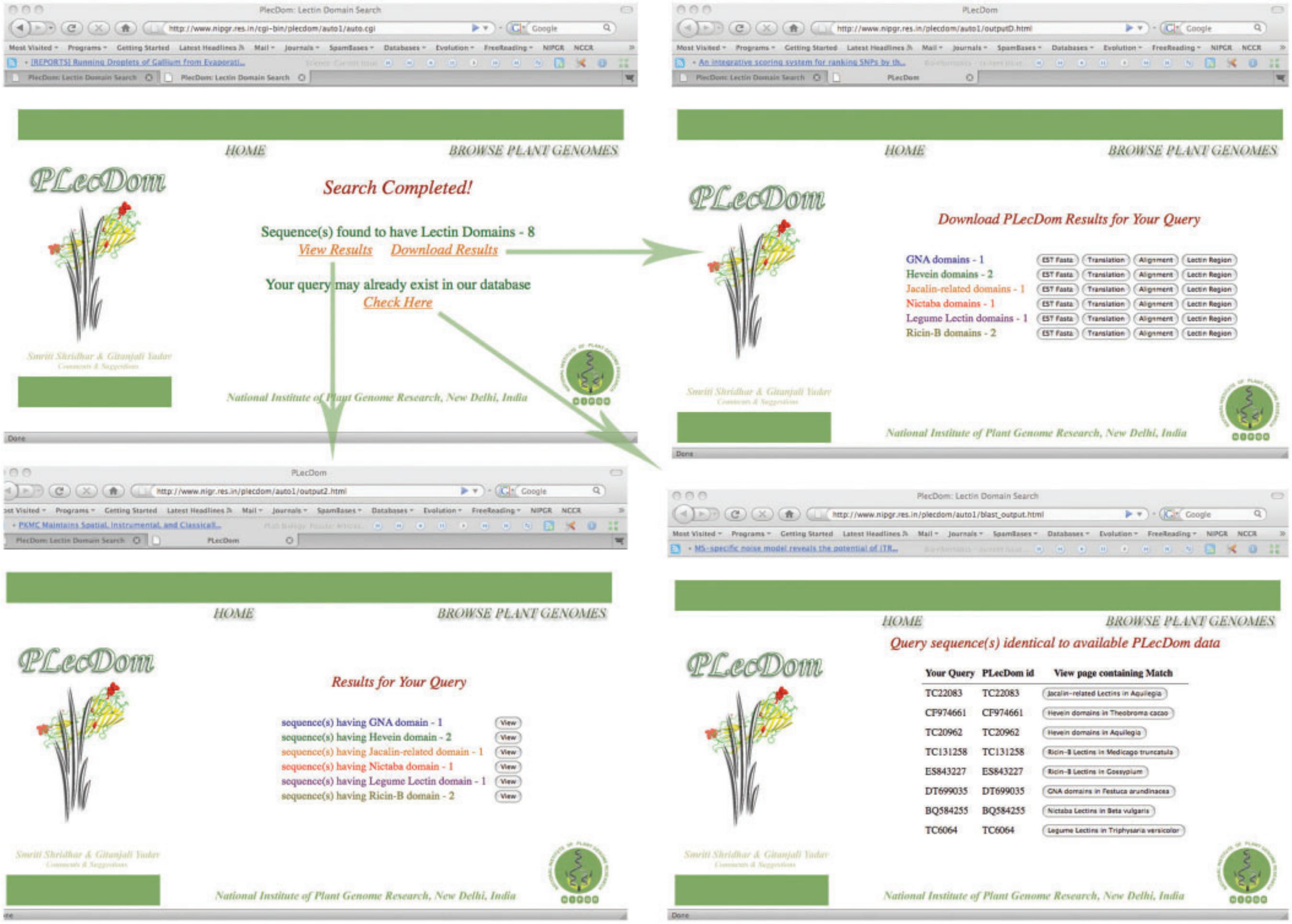
The PLecDom server has been tested on several browsers and platforms, including Safari, Firefox, Konqueror and IE on Macintosh, Linux as well as Windows workstations thereby making it cross-platform compatible. The validation dataset was separated at the time of data collection and comprised of 128 sequences having representatives from all families. The results of PLecDom on this dataset were compared with those from major annotation databases like Pfam (27), PANTHER (28) and SMART (29). In order to check the precision accuracy of the program, and more importantly, its negative prediction ability, a negative dataset comprising 1146 sequences was added to the 128 positives. The performance of the optimized profiles on this dataset was tested using the statistical concepts of sensitivity, specificity and precision. Sensitivity measures the proportion of actual positives, which are correctly identified as such, and was calculated for each family as the ratio of true positives to combined true positives and false negatives. Specificity is a measure of the proportion of negatives that are correctly identified, and was calculated for each domain family as the ratio of true negatives to combined true negatives and false positives. Precision, or positive predictive value, refers to the fraction of returned positives that are true positives, and is often considered more important than accuracy, which estimates the overall proportion of true positives in the population. Precision was calculated as the ratio of true positives to the combined true and false positives.

RESULTS

PLecDom represents a collection of profile Hidden Markov Models (HMMs) based on a rigorous analysis of eight distinct sugar-binding domain families of plant lectins, namely, (i) GNA-related lectin domains, (ii) lectins with Hevein domains, (iii) Jacalin-related lectins, (iv) Legume lectins, (v) Ricin–B lectins, (vi) Galectins, (vii) Calreticulin/Calnexins and (viii) Nictaba lectins. PLecDom is an online, automated, interactive and predictive search tool, the first of its kind dedicated to lectin domains.

Input and output

The PLecDom search algorithm requires input in Fasta format. It can accept protein as well as EST data and automatically recognizes the type of input without user intervention. EST data is translated into six frames, and each frame is tested for presence of lectin domains using the optimized fragmentary search profiles. If multiple


PLecDom: Lectin Domain Search
http://www.nipgr.res.in/cgi-bin/plecdom/auto1/auto.cgi
HOME
BROWSE PLANT GENOMES
PLecDom
Smriti Shridhar & Gitanjali Yadav
Comments & Suggestions
Done

Search Completed!

Sequence(s) found to have Lectin Domains - 8

View Results Download Results

Your query may already exist in our database

Check Here

National Institute of Plant Genome Research, New Delhi, India

PLecDom
http://www.nipgr.res.in/plecdom/auto1/outputD.html
HOME
BROWSE PLANT GENOMES
PLecDom
Smriti Shridhar & Gitanjali Yadav
Comments & Suggestions
Done

Download PLecDom Results for Your Query

GNA domains - 1	EST Fasta	Translation	Alignment	Lectin Region
Hevein domains - 2	EST Fasta	Translation	Alignment	Lectin Region
Jacalin-related domains - 1	EST Fasta	Translation	Alignment	Lectin Region
Nictaba domains - 1	EST Fasta	Translation	Alignment	Lectin Region
Legume Lectin domains - 1	EST Fasta	Translation	Alignment	Lectin Region
Ricin-B domains - 2	EST Fasta	Translation	Alignment	Lectin Region

National Institute of Plant Genome Research, New Delhi, India

PLecDom
http://www.nigr.res.in/plecdom/auto1/output2.html
HOME
BROWSE PLANT GENOMES
PLecDom
Smriti Shridhar & Gitanjali Yadav
Comments & Suggestions

Results for Your Query

sequence(s) having GNA domain - 1	View
sequence(s) having Hevein domain - 2	View
sequence(s) having Jacalin-related domain - 1	View
sequence(s) having Nictaba domain - 1	View
sequence(s) having Legume Lectin domain - 1	View
sequence(s) having Ricin-B domain - 2	View

National Institute of Plant Genome Research, New Delhi, India

PLecDom: Lectin Domain Search
http://www.nipgr.res.in/plecdom/auto1/blast_output.html
HOME
BROWSE PLANT GENOMES
PLecDom
Smriti Shridhar & Gitanjali Yadav
Comments & Suggestions
Done

Query sequence(s) identical to available PLecDom data

Your Query	PLecDom id	View page containing Match
TC22083	TC22083	Jacalin-related Lectins in Aquilegia
CF974661	CF974661	Hevein domains in Theobroma cacao
TC20962	TC20962	Hevein domains in Aquilegia
TC131258	TC131258	Ricin-B Lectins in Medicago truncatula
ES843227	ES843227	Ricin-B Lectins in Gossypium
DT699035	DT699035	GNA domains in Festuca arundinacea
BQ584255	BQ584255	Nictaba Lectins in Beta vulgaris
TC6064	TC6064	Legume Lectins in Triphysaria versicolor

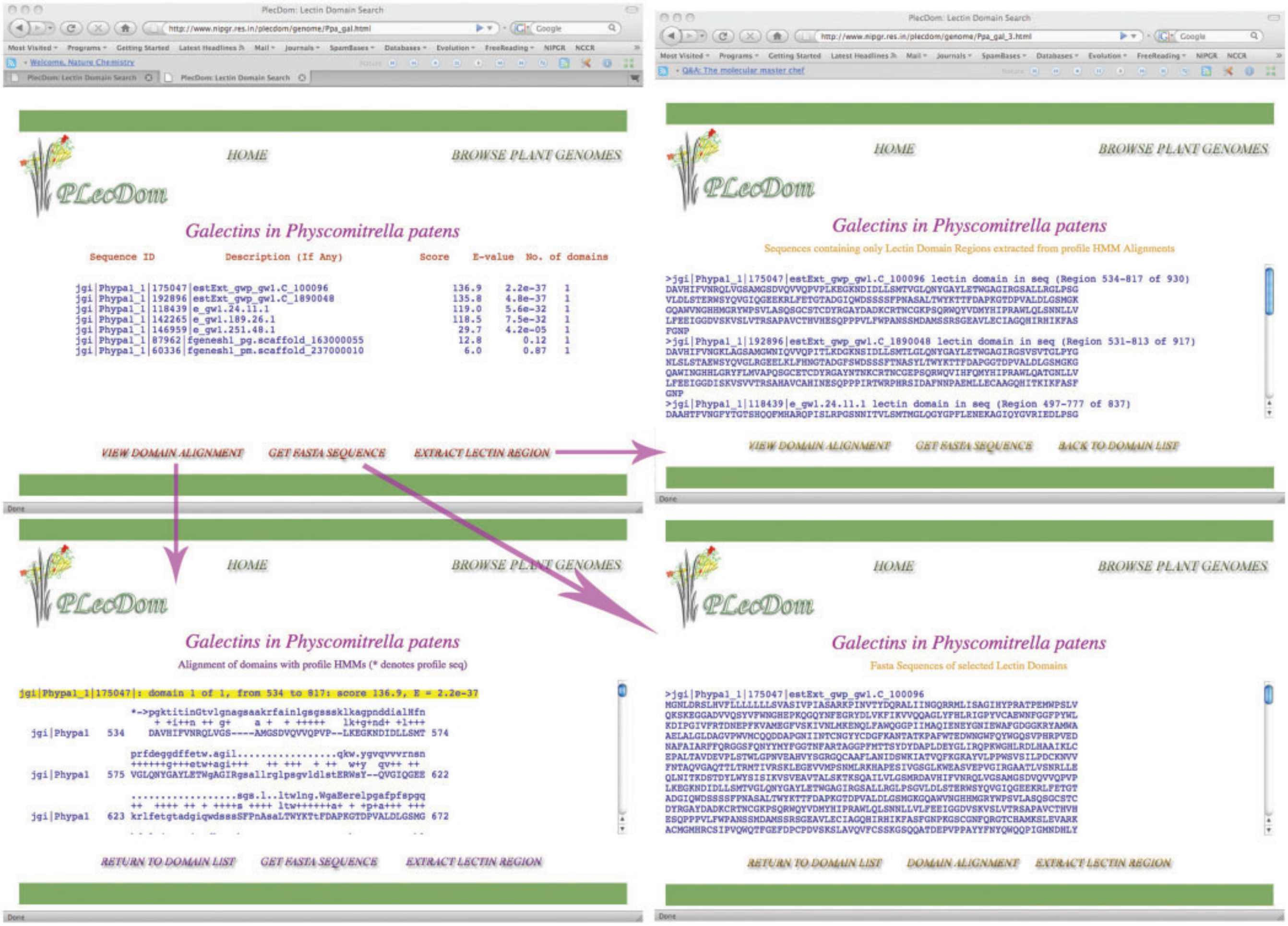
National Institute of Plant Genome Research, New Delhi, India

Figure 2. Snapshots of query submission results in PLecDom.

frames of the same sequence are found to have lectin domains or their parts, they can be viewed in the results window of the relevant lectin family. PLecDom has been designed to process simultaneous connection requests from multiple users, and optimized such that huge sets of input sequences can be scanned in a matter of seconds. A detailed tutorial for query submission has been provided in addition to an example set of test sequences for explaining the submission and search procedure. Figure 2 shows a snapshot of the outcome of a successful search. The identified lectin domains are catalogued by the program into distinct sugar-binding families and users can view the number of domains identified in each family. This feature can be useful to narrow down the spectrum of probable substrates and assist in the prediction of fine specificities of newly identified lectin domains. Users can also check if their input sequences already exist in our database. This aspect of PLecDom can be very useful since researchers with partial sequences may find longer pieces, which could help them in cloning full-length sequences. To analyze the identified domains in detail, users can select the lectin family of interest, as depicted in Figure 3. In case of EST sequences, multiple-frame lectin domains, if detected, can be singled out for analysis on this page, as it shows reading frame numbers within the domain list. Furthermore, alignments of the identified domains with family-specific profile HMMs allow users to estimate the divergence of their data from existing homologs. Domain boundaries captured from this alignment allow users to specifically extract the selected lectin regions. In case of EST data, the domain boundaries enable users to analyze flanking regions on the genome for additional information. For example, putative lectin regions identified from EST sequences can be aligned to genomic DNA to carry out intron analysis and for the presence of signal sequences, providing clues to probable biological function and sub-cellular localization. All sequences including source data, extracted lectin regions, and alignments are available for download by a variety of grouping methods.

Performance

Figure 4A shows the relative performance of PLecDom as compared with three major public databases, namely, Pfam, PANTHER and SMART. As can be seen, PLecDom performs equally well or even better than any of these programs, and most remarkably so for the Nictaba lectins. We would like to emphasize that



```
Sequence ID                  Description (If Any)                 Score   E-value  No. of domains

jgi|Phypal_1|175047|estExt_gwp_gw1.C_100096                       136.9   2.2e-37   1
jgi|Phypal_1|192896|estExt_gwp_gw1.C_1890048                      135.8   4.8e-37   1
jgi|Phypal_1|118439|e_gw1.24.11.1                                 119.0   5.6e-32   1
jgi|Phypal_1|142265|e_gw1.189.26.1                                118.5   7.5e-32   1
jgi|Phypal_1|146959|e_gw1.251.48.1                                 29.7   4.2e-05   1
jgi|Phypal_1|87962|fgenesh1_pg.scaffold_163000055                  12.8      0.12   1
jgi|Phypal_1|60336|fgenesh1_pm.scaffold_237000010                   6.0      0.87   1
```

PlecDom: Lectin Domain Search
http://www.nipgr.res.in/plecdom/genome/Ppa_gal_3.html
Q&A: The molecular master chef
HOME
BROWSE PLANT GENOMES
PLecDom
Galectins in Physcomitrella patens
Sequences containing only Lectin Domain Regions extracted from profile HMM Alignments
VIEW DOMAIN ALIGNMENT
GET FASTA SEQUENCE
BACK TO DOMAIN LIST
Done

```
>jgi|Phypal_1|175047|estExt_gwp_gw1.C_100096 lectin domain in seq (Region 534-817 of 930)
DAVHIFVNRQLVGSAMGSDVQVVQPVPLKEGKNDIDLLSMTVGLQNYGAYLETWGAGIRGSALLRGLPSG
VLDLSTERWSYQVGIQGEEKRLFETGTADGIQWDSSSSFPNASALTWYKTTFDAPKGTDPVALDLGSMGK
GQAWVNGHHMGRYWPSVLASQSGCSTCDYRGAYDADKCRTNCGKPSQRWQYVDMYHIPRAWLQLSNNLLV
LFEEIGGDVSKVSLVTRSAPAVCTHVHESQPPPVLFWPANSSMDAMSSRSGEAVLECIAGQHIRHIKFAS
FGNP
>jgi|Phypal_1|192896|estExt_gwp_gw1.C_1890048 lectin domain in seq (Region 531-813 of 917)
DAVHIFVNGKLAGSAMGWNIQVVQPITLKDGKNSIDLLSMTLGLQNYGAYLETWGAGIRGSVSVTGLPYG
NLSLSTAEWSYQVGLRGEELKLFHNGTADGFSWDSSSFTNASYLTWYKTTFDAPGGTDPVALDLGSMGKG
QAWINGHHLGRYFLMVAPQSGCETCDYRGAYNTNKCRTNCGEPSQRWQVIHFQMYHIPRAWLQATGNLLV
LFEEIGGDISKVSVVTRSAHAVCAHINESQPPPIRTWRPHRSIDAFNNPAEMLLECAAGQHITKIKFASF
GNP
>jgi|Phypal_1|118439|e_gw1.24.11.1 lectin domain in seq (Region 497-777 of 837)
DAAHTFVNGFYTGTSHQQFMHARQPISLRPGSNNITVLSMTMGLQGYGPFLENEKAGIQYGVRIEDLPSG
```

HOME
BROWSE PLANT GENOMES
PLecDom
Galectins in Physcomitrella patens
Alignment of domains with profile HMMs (* denotes profile seq)
RETURN TO DOMAIN LIST
GET FASTA SEQUENCE
EXTRACT LECTIN REGION
Done

```
jgi|Phypal_1|175047|: domain 1 of 1, from 534 to 817: score 136.9, E = 2.2e-37

                   *->pgktitinGtvlgnagsaakrfainlgsgsssklkagpnddialHfn
                      + +i++n ++ g+    a +  + +++++    lk+g+nd+ +l+++
  jgi|Phypal   534    DAVHIFVNRQLVGS----AMGSDVQVVQPVP--LKEGKNDIDLLSMT 574

                   prfdeggdffetw.agil.................qkw.ygvqvvvrnsn
                   ++++++g+++etw+agi+++   ++ +++   + ++   w+y   qv++ ++
  jgi|Phypal   575 VGLQNYGAYLETWgAGIRgsallrglpsgvldlstERWsY--QVGIQGEE 622

                   ..................sgs.l..ltwlng.WgaEerelpgafpfspgq
                   ++  ++++ ++ + ++++s ++++ ltw++++++a+ + +p+a+++ +++
  jgi|Phypal   623 krlfetgtadgiqwdsssSFPnAsaLTWYKTtFDAPKGTDPVALDLGSMG 672
```

```
>jgi|Phypal_1|175047|estExt_gwp_gw1.C_100096
MGNLDRSLHVFLLLLLLLSVASIVPIASARKPINVTYDQRALIINGQRRMLISAGIHYPRATPEMWPSLV
QKSKEGGADVVQSYVFWNGHEPKQGQYNFEGRYDLVKFIKVVQQAGLYFHLRIGPYVCAEWNFGGFPYWL
KDIPGIVFRTDNEPFKVAMEGFVSKIVNLMKENQLFAWQGGPIIMAQIENEYGNIEWAFGDGGKRYAMWA
AELALGLDAGVPWVMCQQDDAPGNIINTCNGYYCDGFKANTATKPAFWTEDWNGWFQYWGQSVPHRPVED
NAFAIARFFQRGGSFQNYYMYFGGTNFARTAGGPFMTTSYDYDAPLDEYGLIRQPKWGHLRDLHAAIKLC
EPALTAVDEVPLSTWLGPNVEAHVYSGRGQCAAFLANIDSWKIATVQFKGKAYVLPPWSVSILPDCKNVV
FNTAQVGAQTTLTRMTIVRSKLEGEVVMPSNMLRKHAPESIVGSGLKWEASVEPVGIRGAATLVSNRLLE
QLNITKDSTDYLWYSISIKVSVEAVTALSKTKSQAILVLGSMRDAVHIFVNRQLVGSAMGSDVQVVQPVP
LKEGKNDIDLLSMTVGLQNYGAYLETWGAGIRGSALLRGLPSGVLDLSTERWSYQVGIQGEEKRLFETGT
ADGIQWDSSSSFPNASALTWYKTTFDAPKGTDPVALDLGSMGKGQAWVNGHHMGRYWPSVLASQSGCSTC
DYRGAYDADKCRTNCGKPSQRWQYVDMYHIPRAWLQLSNNLLVLFEEIGGDVSKVSLVTRSAPAVCTHVH
ESQPPPVLFWPANSSMDAMSSRSGEAVLECIAGQHIRHIKFASFGNPKGSCGNFQRGTCHAMKSLEVARK
ACMGMHRCSIPVQWQTFGEFDPCPDVSKSLAVQVFCSSKGSQQATDEPVPPAYYFNYQWQQPIGMNDHLY
```

Figure 3. Genome browsing snapshots for a specific lectin family. Screendumps showing details of Galectins identified by PlecDom in the moss genome.

PLecDom is currently the only program available for detection and characterization of these lectins, since no existing database or annotation program is able to identify them. In case of the Ricin-B family and Hevein domain lectins, PLecDom performs better than PANTHER predictions but marginally less than Pfam.

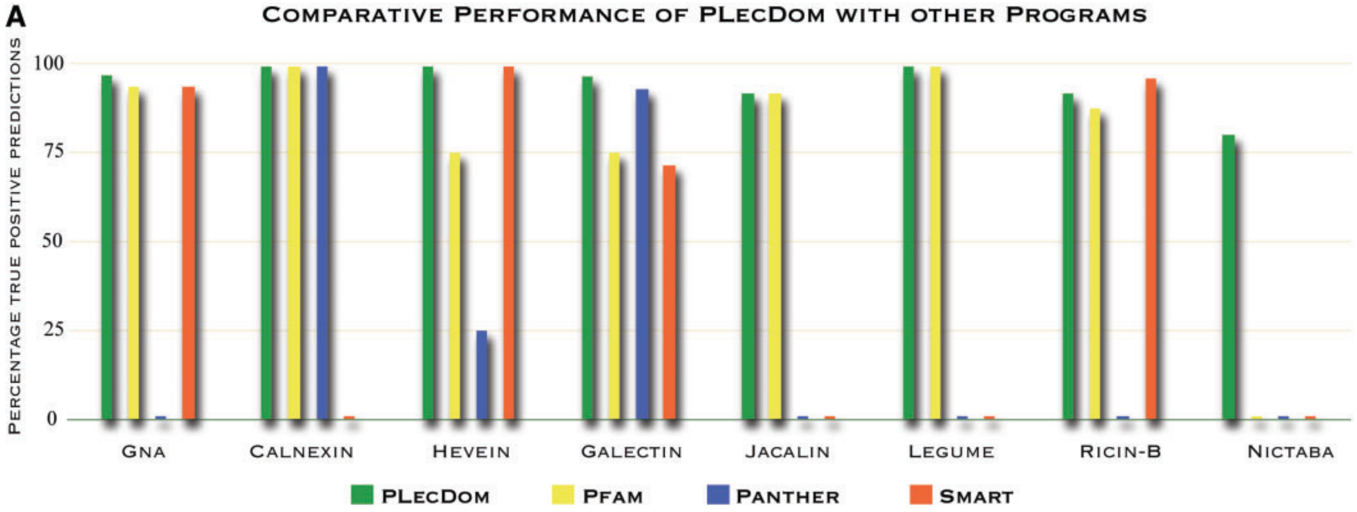
The second test of PLecDom performance was statistical and Figure 4B shows the various measures of prediction accuracy in each of the eight lectin families analyzed. It is interesting to note that all families reveal very high specificities and prediction accuracy, the lowest values being 75% sensitivity in case of Hevein domains and 75.86% precision in case of Ricin-B family. The implications of these observations are discussed in the last section.

Browsing PLecDom

Encouraged by the high prediction accuracy of PLecDom (see Figure 4), and its ease with handling high-density data, we applied the program to the available protein predictions and EST sequence data from 48 sequencing projects. This exercise resulted in the identification of more than 7000 lectin domains and their assignation to various known plant lectin families. It may be noted that, approximately 5000 of the identified sequences are novel, i.e. previously un-annotated or, in a few cases, annotated as 'putative' or 'conserved hypotheticals'. This data thus represents a major enrichment of plant lectin domain homologs known till date. The identified sequences have been made available at the web server and can be browsed by species name, followed by the domain family of interest. In each case, the output is similar to that described in Figure 3 so as to maintain consistency with query submission outputs. For incomplete genomes, we have additionally made the EST sequences of detected lectins available for download to users. A genome browsing tutorial has also been provided on the web page to explain the search pages in detail.

DISCUSSION

PLecDom is a web server dedicated to plant lectin domains. The program enables identification and analysis based on a computational exploration of the sequence space of currently available and characterized plant lectin families. PLecDom has been designed to accept EST input in addition to protein sequences to assist researchers with preliminary annotation of newly emerging data. Our objective is to build a robust tool for



B

Statistical Measures of Predictive Performance

	Sensitivity	Specificity	Accuracy	Precision
GNA-related lectins	96.79	100	99.1	100
Calnexin/Calreticulin	100	100	100	100
Hevein domains	75	100	99.1	100
Galectins	96.4	100	99.1	100
Jacalins	91.6	100	99.1	100
Legume Lectins	100	100	100	100
Ricin-B domains	91.6	93.39	93.07	75.86
Nictaba Lectins	80	100	99.2	100

Figure 4. (**A**) Assessment of PLecDom predictions. The chart shows performance of PLecDom as compared to other databases for the validation dataset. Databases having null predictions are shown as base line stubs. PLecDom (in green) shows the best performance in predicting plant lectin domains. Note the Nictaba family predictions where PLecDom is the only program that can detect these domains. (**B**) Statistical performance measures of PLecDom. All values are reported as percentages.

detection of all reported families of plant lectins. However, a few families, mainly the very recently recognized ones, had to be discarded at the data collection stage on account of lack of sufficient number of representative characterized sequences. The families excluded from the current version of PLecDom are the Amaranthins, ABA, EEA, cyanovirins, LysM families, and class V chitinase homologs with lectin activity. In future, as more sequences from these and other families become available, we will update the server to broaden its base and scope.

One of the major achievements of PLecDom is its remarkable predictive success. Unlike pairwise or multiple sequence comparisons, which can often result in capturing false homologs, the optimized profile HMMs in PLecDom give distinct outputs, and are very sensitive, thereby enhancing the likelihood of predicting genuine homologs. A comparison of PLecDom predictions with those obtained from general annotation databases currently available, showed that our program performs very well, and is better than several existing programs, most strikingly so in case of the Nictaba lectin family. We believe our program is a more reliable tool for plant lectin annotation. All families except Ricin-B showed 100% specificity and very high precision values (see Figure 4B). The Ricin-B family profiles returned seven false positive identifications, thereby lowering its precision, although the specificity and accuracy of the family remains significantly high. A comparatively low sensitivity was observed in case of PLecDom predictions for the Hevein domains (75%), but the Hevein domain family profiles showed very high specificity and precision (100%), revealing that although a distant homolog of this family may sometimes fail to be recognized, a positive identification would nevertheless be conclusive. Overall the positive predictive value of all families in PLecDom is very high, thereby making it highly suitable for annotation and assignation of domain family to test sequences.

The application of PLecDom to protein and EST sequence data from 48 sequencing projects resulted in a considerable enrichment of the currently annotated plant lectins. This wealth of data can now be used to carry out a comprehensive functional analysis and highlights potential targets for biochemical characterization in many species. Several interesting insights have been gained from the PLecDom outputs. For example, the data show that, contrary to previous assumptions, legume lectins do occur in several non-leguminous species, and the monocot mannose-binding lectin family has many homologs in dicots as well. Preliminary studies reveal that these atypical homologs have different intronic and exonic features, and this may potentially lead to the recognition of new functions or families (G.Yadav unpublished data). Further, the almost complete 'lectinome' that has been extracted in this study, for 10 fully sequenced plant genomes, reveals a mixed set of lectin family combinations in each species rather than taxon-specific lectin families probably signifying a unique 'lectin signature' of individual genomes.

Taken together, these observations provide fascinating new insights into the diversity of plant lectins and a huge potential to investigate their evolutionary ramifications. To summarize, we believe that PLecDom would be of immediate interest to glycobiologists and researchers involved in the identification, annotation and characterization of plant lectins, as well in the study of plant stress responses.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The JGI sequence data for lower plants used in this work were produced by the US Department of Energy Joint Genome Institute http://www.jgi.doe.gov/in collaboration with the user community. Laboratory resources provided under BTIS grant of DBT, India, are gratefully acknowledged. Authors thank Dr Suresh Babu, CEMDE, University of Delhi for constructive criticism of the manuscript. Mr Sachin Pundhir's assistance with server monitoring is acknowledged.

FUNDING

Innovative Young Biotechnologist grant to GY by the Department of Biotechnology, Government of India [BT/BI/12/040/2005] and [BT/PR/5360/Agr/16/483/2004]. Funding for open access charge: USD 2670.

Conflict of interest statement. None declared.

REFERENCES

1. Sharon,N. and Lis,H. (2004) History of lectins: from hemagglutinins to biological recognition molecules. *Glycobiology*, **14**, R53–R62.
2. Peumans,W.J. and Van Damme,E.J.M. (1995) Lectins as plant defense proteins. *Plant Physiol*., **109**, 347–352.
3. Rudiger,H. and Gabius,H.J. (2001) Plant lectins: occurrence, biochemistry, functions and applications. *Glycoconj. J*., **18**, 589–613.
4. Chrispeels,M.J. and Raikhel,N.V. (1991) Lectins, lectin genes, and their role in plant defense. *Plant Cell*, **3**, 1–9.
5. Sinha,S., Gupta,G., Vijayan,M. and Surolia,A. (2007) Subunit assembly of plant lectins. *Curr. Opin. Struct. Biol*., **17**, 498–505.
6. Van Damme,E.J.M., Rougé,P. and Peumans,W.J. (2007) Plant lectins. In Kamerling,J.P., Boons,G.J., Lee,Y.C., Suzuki,A., Taniguchi,N. and Voragen,A.J.G. (eds), *Comprehensive Glycoscience – From Chemistry to Systems Biology*. Vol. 3, Elsevier, Oxford, UK, pp. 563–599.
7. Hester,G., Kaku,H., Goldstein,I.J. and Wright,C.S. (1995) Structure of mannose-specific snowdrop (Galanthus nivalis) lectin is representative of a new plant lectin family. *Nat. Struct. Biol*., **2**, 472–479.
8. Van Damme,E.J.M., Lannoo,N. and Peumans,W.J. (2008) Plant Lectins. *Adv. Botanical Res*., **48**, 107–209.
9. Dodd,R.B. and Drickamer,K. (2001) Lectin-like proteins in model organisms: implications for evolution of carbohydrate-binding activity. *Glycobiology*, **11**, R71–R79.
10. Leffler,H., Carlsson,S., Hedlund,M., Qian,Y. and Poirier,F. (2004) Introduction to galectins. *Glycoconj. J*., **19**, 433–440.
11. Van Damme,E.J.M., Barre,A., Rougé,P. and Peumans,W.J. (2004) Cytoplasmic/nuclear plant lectins: a new story. *Trends Plant Sci*., **9**, 484–489.
12. Van Damme,E.J.M., Culerrier,R., Barre,A., Alvarez,R., Rougé,P. and Peumans,W.J. (2007) A novel family of lectins evolutionarily related to class V chitinases: an example of neofunctionalization in legumes. *Plant Physiol*., **144**, 662–672.
13. Fouquaert,E., Peumans,W.J., Smith,D.F., Proost,P., Savvides,S.N. and Van Damme,E.J.M. (2008) The "old" Euonymus europaeus agglutinin represents a novel family of ubiquitous plant proteins. *Plant Physio.l*, **147**, 1316–1324.
14. Onaga,S. and Taira,T. (2008) A new type of plant chitinase containing LysM domains from a fern (Pteris ryukyuensis): roles of LysM domains in chitin binding and antifungal activity. *Glycobiology*, **18**, 414–423.
15. Koharudin,L.M., Viscomi,A.R., Jee,J.G., Ottonello,S. and Gronenborn,A.M. (2008) The evolutionarily conserved family of cyanovirin-N homologs: structures and carbohydrate specificity. *Structure*, **16**, 570–584.
16. Neumann,D., Lehr,C.M., Lenhof,H.P. and Kohlbacher,O. (2004) Computational modeling of the sugar-lectin interaction. *Adv. Drug Deliv. Rev*., **56**, 437–457.
17. Kerzmann,A., Fuhrmann,J., Kohlbacher,O. and Neumann,D. (2008) BALLDock/SLICK: a new method for protein-carbohydrate docking. *J. Chem. Inf. Model*, **48**, 1616–1625.
18. Fujimoto,Y.K., Terbush,R.N., Patsalo,V. and Green,D.F. (2008) Computational models explain the oligosaccharide specificity of cyanovirin-N. *Protein Sci*., **17**, 2008–2014.
19. Aoki,K.F., Ueda,N., Yamaguchi,A., Kanehisa,M., Akutsu,T. and Mamitsuka,H. (2004) Application of a new probabilistic model for recognizing complex patterns in glycans. *Bioinformatics*, **20(Suppl 1)**, i6–i14.
20. Sharon,N. and Lis,H. (2002) How proteins bind carbohydrates: lessons from legume lectins. *J. Agric. Food Chem*., **50**, 6586–6591.
21. Chandra,N.R., Kumar,N., Jeyakani,J., Singh,D.D., Gowda,S.B. and Prathima,M.N. (2006) Lectindb: a plant lectin database. *Glycobiology*, **16**, 938–946.
22. Lee,Y., Tsai,J., Sunkara,S., Karamycheva,S., Pertea,G., Sultana,R., Antonescu,V., Chan,A., Cheung,F. and Quackenbush,J. (2005) The TIGR Gene Indices: clustering and assembling EST and known genes and integration with eukaryotic genomes. *Nucleic Acids Res*., **33**, D71–D74.
23. Larkin,M.A., Blackshields,G., Brown,N.P., Chenna,R., McGettigan,P.A., McWilliam,H., Valentin,F., Wallace,I.M., Wilm,A., Lopez,R. *et al*. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.
24. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
25. Rice,P., Longden,I. and Bleasby,A. (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet*., **16**, 276–277.
26. 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.
27. Finn,R.D., Tate,J., Mistry,J., Coggill,P.C., Sammut,S.J., Hotz,H.R., Ceric,G., Forslund,K., Eddy,S.R., Sonnhammer,E.L. *et al*. (2008) The Pfam protein families database. *Nucleic Acids Res*., **36**, D281–D288.
28. Thomas,P.D., Campbell,M.J., Kejariwal,A., Mi,H., Karlak,B., Daverman,R., Diemer,K., Muruganujan,A. and Narechania,A. (2003) PANTHER: A Library of Protein Families and Subfamilies Indexed by Function. *Genome Res*., **13**, 2129–2141.
29. Letunic,I., Doerks,T. and Bork,P. (2009) SMART 6: recent updates and new developments. *Nucleic Acids Res*., **37**, D229–D232.