Chapter 6

Isolation, Purification, and Characterization of a novel lectin (CotLec) from the seeds of *Gossypium arboreum*

6.1 Summary

A novel lectin was purified from the seeds of cotton (*Gossypium arboreum*) by affinity chromatography on a Sepharose-4B coupled galactose column and designated as CotLec. The results of sodium dodecyl sulfate-polyacrylamide gel electrophoresis and gel filtration showed CotLec to be a monomer of ~15 kDa with no disulfide linkage. Acid-Schiff staining method suggested lack of glycosylation of CotLec. N-terminal sequencing and MALDI-TOF analysis reveal the CotLec belong to a vicilin-like seed storage protein family. In hemagglutination inhibition assay by mono and di-saccharides, galactose was the strongest inhibitor followed by lactose, and the glycoprotein fetuin showed inhibition at 0.1%. Effect of Temperature and pH on hemagglutination activity showed that CotLec is stable up to 60°C and in pH range of 3 to 9. Calculated dissociation constant of CotLec with galactose and lactose through fluorescence spectroscopy was 2.4 ± 0.3 $X10^{-3}$ M and $3.4 \pm 0.5X10^{-3}$ M, respectively. Circular dichroism spectroscopy of CotLec showed it is a mixture of both α- helical (~10%) and sheet structure (~30%) and loops which is in line with known vicilin structures. A literature search showed the presence of vicilin-like protein exhibiting lectin-like activity from *Albizia lebbeck*, *Enterolobium contortisiliquum*, and *Erythrina velutina*. This study is the first report of a vicilin protein with lectin activity from cotton.

6.2 Introduction

Cotton has been used for fiber and fabric production Since 6000 BC and is considered as most important fiber crop across the globe due its importance in textile industry. Cotton belongs to the genus *Gossypium,* which is part of *Malvaceae* or Mallow family. Other members of this family include okra, hollyhock, rose of Sharon (Mauney and Stewart 1986). Besides textile industry, cotton is also an important source of edible oil and nutrient-rich food crop due to the presence of high-quality oil and proteins in cotton seeds. Cotton seeds contain 17-24 % of oil and 40-43% of proteins (Zhou, Sun et al. 2014). India produces about 18% of total world`s cotton production, and it plays a very important role in Indian economy (Annual Report 2014, Cotton Corporation of India Ltd. Government of India). The growth and thus productivity of cotton plants are frequently fraught with biotic and abiotic stress such as pests and pathogens, salinity, heat, and draughts. Compared with most crop plants, cotton adapts quite well to adverse conditions due to sophisticated defense mechanism they have developed over the course of evolution. For example, it is considerably more tolerant to high salinity soils than corn (Mauney and Stewart 1986) and one of the major contributors of this sophisticated defense mechanism is proteins expressed in the seeds.

Seed storage proteins can be defined as the reservoir of metal ions and amino acids which are utilized by organisms to sustain and flourish life. In plants, storage proteins perform essential roles of reserves of nutrition, which is necessary to maintain good health and promote growth which becomes all the more important during embryonic and development stages of plant life cycle (Shewry et al., 1995). They also play a very important role in plant defense by their insecticidal and antimicrobial activities.

The seed contains proteins of several kinds, which are classified into four major categories based on their solubility. i.e. (1) Albumins: soluble in water, (2) globulins: soluble in saline and mostly find in legume seeds, (3) prolamins: Alcohol/water mixture mostly present in cereal seeds and (4) glutelins: alkali/diluted acid present in cereals seeds. Other proteins are used as a source of carbon and nitrogen and are associated with defense mechanism, and members include protease inhibitors, lectin, lectin-like proteins, ribosome inactivating proteins.

Lectins are carbohydrate-binding proteins or glycoproteins of non-immune origin. The selectively and reversibly bind to mono or oligosaccharides without altering their structure. Plant lectins have varied biological roles such as plant defense and cell signaling and are shown to elicit antiviral, antibacterial and antitumor response. They are also abundantly expressed in seeds and are sometimes considered as seed storage proteins (Chrispeels et al., 1991).

Over the last few years, there have been reports characterizing vicilin with lectin activity, and some examples are vicilin showing an affinity for chitin and isolated from *Albizia lebbeck*, *Enterolobium contortisiliquum*, and *Erythrina velutina*.

In this study, we report the isolation, purification, characterization of a vicilin seed storage protein from cotton with lectin activity and hope it can be well developed further for plant defense activity.

6.2.1 N-terminal sequencing

N-terminal sequencing method uses a technique which chemically determines which amino acid is present at the N-terminus of a polypeptide chain. Edman degradation is a widely used method to sequence the amino acid present in the polypeptide chain. The basic principle of Edman degradation method is that; any amino acid residue can be modified chemically in a way that they can be removed from the chain without disturbing the bonds between other amino acids.

The basic procedure for Edman degradation is as follow;

1. The polypeptide which needs to be sequence first immobilized on a Polyvinylidene fluoride (PVDF) membrane by electro blotting.
2. A chemically called phenylisothiocyanate (PITC) under mildly alkaline conditions reacts with an uncharged terminal group on the amino acid chain to form phenylthiocarbamoyl derivative.
3. This phenylthiocarbamoyl derivative amino acid is cleaved by Trifluoroacetic acid generating its anilinothiazolinone derivative.
4. Step 2 and 3 are repeated on next terminal amino acid.

5. Excess buffer and reagents are removed by a wash and ATZ-amino acids are carefully extracted with ethyl acetate and converted to more stable phenylthiohydantoin (PTH)- amino acid derivative.
6. Identification of PTH-amino acid is done through chromatography or electrophoresis by comparing them with standards.

Edman degradation is the most suitable method for amino acid sequencing due to its sensitivity and speed. The major drawback of this method is the length of the polypeptide chain; the procedure tends to fail due to incompletion of cyclic derivatization.

6.3 Materials and Methods

6.3.1 Materials

Cotton seeds were obtained from local markets of Ahmedabad, Gujarat, India. Mannose, galactose, N-Acetyl-D-glucosamine, and N-acetyl-D-galactosamine, were obtained from Himedia Laboratories, Mumbai, Maharashtra, India. Fetuin and other chemicals required were procured from Sigma-Aldrich and protein molecular weight markers were acquired from Thermo Fisher Scientific, Framingham, MA, USA.

6.3.2 Methods

6.3.2.1 Isolation and Purification

Total crude protein extract was isolated from cotton seeds after removal of oil and all the operations were done at 4°C. Briefly, 500 gm of dry seeds were crushed in an electric grinder, and fine powder was soaked with sufficient pre-cooled acetone and left overnight at 4°C. Next day the acetone was removed, and the mixture was air-dried at room temperature to remove the excess acetone. After complete removal of residual acetone, the dry powder was mixed with pre-cooled PBS and stirred in the same buffer at 4°C overnight. The suspension was then filtered through a double layer muslin cloth at 4°C and the filtrate collected was centrifuged at 10,000 *x g* at 4°C for 30 minutes to obtain a clear solution. Total protein was precipitated from this solution by adding 80% ammonium sulfate,

followed by continuous overnight stirring and collection of the precipitate by subjecting the suspension to centrifuge at 10,000 *x g* at 4°C for 45 minutes. The precipitate obtained was dissolved in a minimum volume of pre-cooled PBS and dialyzed against the same buffer using dialysis membrane, to obtain crude protein extract.

The crude protein extract was applied to a sepharose 4B coupled galactose column (5 cm x 1 cm) pre-equilibrated with PBS, washed with sufficient PBS and finally the bound protein was eluted with 0.5 M galactose at 4°C. The collected fractions were then dialyzed against PBS and fractions which showed agglutination with rabbit erythrocytes (RBC) were collected and concentrated and protein purification checked by SDS-PAGE (Laemmli, U.K., 1970)

6.3.2.2 Molecular weight determination

Molecular weight determination of CotLec was carried out according to method described in section 2.5.2.1 of chapter 2 of this thesis.

6.3.2.3 Gel filtration

Gel filtration studies of CotLec was carried out according to method describe in section 2.5.2.2 of chapter 2 of this thesis.

6.3.2.4 N-terminal sequencing

N-terminal protein sequencing was performed on lyophilized protein sample by procise 490 cLC system (Applied Biosystems, USA) and conducted at RIKEN, Japan

6.3.2.5 MALDI-TOF-MS analysis for protein identification

MALDI-TOF analysis of protein identification for CotLec was carried out according to method described in section 2.5.4 of chapter 2 of this thesis.

6.3.2.6 *De novo* protein sequencing

De novo protein sequencing of CotLec was carried out according to method described in section 2.5.3 of chapter 2 of this thesis.

6.3.2.7 Hemagglutination and carbohydrate specificity of CotLec

Hemagglutination and carbohydrate inhibition studies of CotLec was carried out according to method describe in section 2.5.5 of chapter 2 of this thesis.

6.3.2.8 Effect of temperature and pH on hemagglutination activity

Effect of temperature and pH on lectin activity of CotLec was studied according to method described in section 2.5.6 of chapter 2 of this thesis.

6.3.2.9 Fluorescence spectroscopy study

Protein interaction with galactose and lactose were investigated by monitoring the changes in fluorescence intensity and λ_{max} of the protein, resulting from binding of the ligand using a spectrofluorometer (Hitachi F-7000 FL, Hitachi Ltd., Tokyo, Japan). The slit width of excitation and emission monochromators was adjusted at 5nm and the response time, and scan speed were kept at 0.5 seconds and 1200nm/minute, respectively. Titrations were performed by adding aliquots (1µl at each time) of test sugars from stock (1000mM) in 50mM Tris-HCl (pH 7.5) to 1ml of purified protein (0.2mg/mL) in the same buffer. Samples were mixed thoroughly with pipetting, making sure that no bubble formation occurred, and excited at 295nm. The emission spectra were recorded between 300 and 400nm. Each spectrum was an average of three consecutive accumulations and corrected for dilution effect. The dissociation constant for the binding of galactose and lactose with purified protein was calculated by the formula given by Chipman et al. (Chipman, Grisaro et al. 1967).

$$\log [C]f = -\log[Ka] + \log [(F0-FC)/(FC-F\infty)] \quad \text{(Eq. 6.1)}$$

Where C is a concentration of the sugars, F0 is fluorescence intensity of protein alone; Fc is fluorescence intensity of protein in the presence of carbohydrates F∞ is fluorescence intensity of protein in the presence of ∞ concentration of carbohydrates.

From the ordinate intercept of the double reciprocal plot of F0/(F0-FC) versus 1/[C], where F0 and FC are the fluorescence intensities of the free protein and the protein plus sugar concentration [C], F∞, the fluorescence intensity upon saturation of all the sugar binding

sites is obtained. In the plot of log[(F0-FC)/(FC-F∞)] versus log[C], the abscissa intercept yielded the Kd value (the dissociation constant) for the protein-sugar interactions, which is the reciprocal of Ka (the association constant). The free energy changes of the association (ΔG) were calculated by using the equation:

$-\Delta G = RT \ln(Ka)$ (Eq. 6.2)

6.3.2.10 Circular dichroism spectroscopy

To record the CD spectra, a Jasco J-810 (Jasco corp., Tokyo, Japan) spectropolarimeter equipped with Peltier thermostat was used. 5µM of purified protein in 20mM Tris-HCl buffer (pH 7.5) was placed in a 2mm path length rectangular quartz cell and spectra recorded at a scan speed of 20nm/min with a response time of 4s, a slit width of 2nm and the spectra recorded for the far-UV region (250 to 200 nm). Analysis of the CD spectrum was done using three different methods, namely CDSSTR, CONTINLL and SELCON3 (Sreerama et al., 1999; Johnson, 1999; Provencher et al., 1981) and employing the software routines available at CDPro. A basis set containing 43 proteins was used as a reference for fitting the experimental spectrum.

6.3.2.11 Bioinformatics analysis

The sequence homology search was done using the BLAST program available from National Centre for Biotechnology Information (NCBI) (Altschul et al., 1990). Sequences used in the various analysis were downloaded from Uniprot database available at (http://www.Uniprot.org/) using the keyword "Vicilin". Multiple sequence alignment was done using muscle program (Edgar, R.C., 2004) and phylogenetic analysis was carried out using Maximum-Likelihood method in MEGA package (Tamura et al., 2011. A statistical boot-strapping (100 steps) test was used to evaluate the phylogeny tree. Conserved domain database (CDD) was used for domain search (Marchler-Bauer et al., 2004). Molecular modeling and model refinement were done using I-tasser server (Yang et al., 2015) and modrefiner tool (Xu & Zhang, 2011), respectively. Structural homologs were obtained using DALI server (Holm & Rosenstrom, 2010). Chimera was used for molecular visualization and superposing the structures (Pettersen et al., 2004).

6.4 Results

6.4.1 Purification of Cotton lectin

The crude extract obtained from cotton seeds were applied to sepharose-4B bound galactose and after subsequent washing to remove the unbound protein from column it was eluted with 0.5M galactose (Figure 1A) and the purification process is summarized (Table 6.1). The eluted fractions were dialyzed against PBS and pure protein which was named cotton lectin (CotLec), was obtained. In a typical purification experiment, approximately 20mg of purified protein was obtained starting from 500g of cotton seeds with specific activity of 2,048HU/mg and the yield and specificity activity of protein purified at each stage of purification is shown (Table 6.1)

6.4.2 Mass analysis of CotLec

CotLec gave a single band with an apparent mass of ~ 16 kDa on SDS-PAGE and the presence or absence of DTT in sample buffer did not change the migration pattern on

SDS-PAGE (Lane 2; Figure 1B). Gel filtration study of purified CotLec showed that it existed as a monomer in solution, as indicated by its K value of 12.34 which is corresponded to ~16 kDa (Figure 1C). Acid-Schiff staining of SDS-PAGE gel to check the glycosylation of CotLec did not show staining suggesting a lack of N or O-linked glycosylation of CotLec.

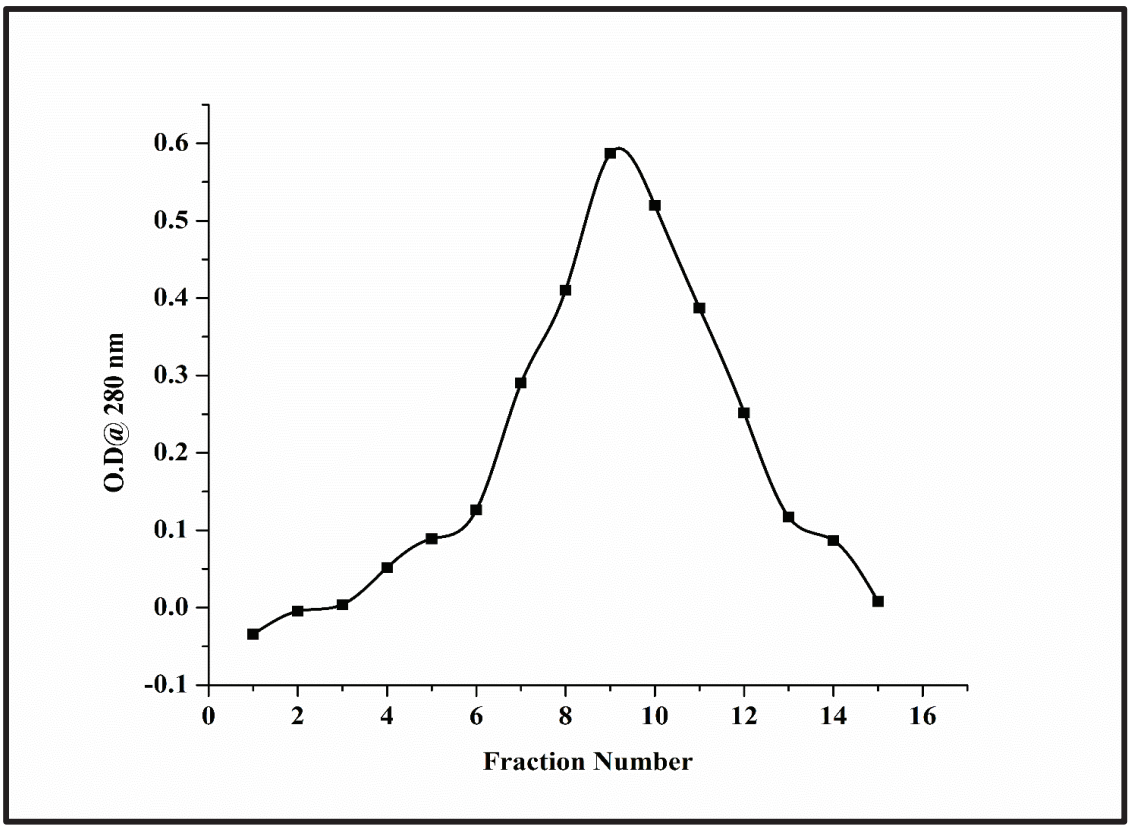


Figure 6.1 A Purification of Cotton lectin Affinity galactose- sepharose 4B linked column profile where X-axis demonstrate number of fractions collected (2 mL each) and Y-axis the absorbance of each fraction at 280nm

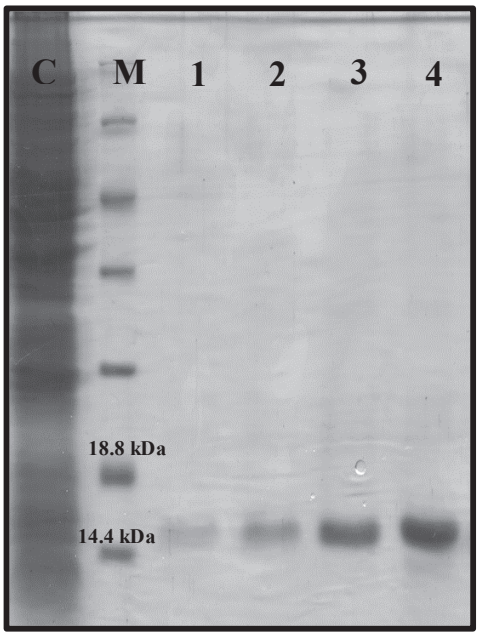


Figure 6.1 B SDS-PAGE analysis of CotLec CotLec was purified through galactose column and homogeneity of purified protein was checked through SDS-PAGE. A sample of purified CotLec was also run and reduced condition to check the existence of possible disulfide bonds. Lane C; Crude protein extract from seeds of cotton; Lane M: Standard Molecular weight markers; Lane 1: Elution fraction 2; Lane 2: Elution fraction 4; Lane 3: Elution fraction 6; Lane 4: Elution fraction 6 in the presence of 1 mM DTT.

Table 6.1 Purification and Specific activity of CotLec from *Gossypium arboreum*

Fraction	Total Volume (mL)	Total Protein (mg)	Total Activity (titer)[a]X volume	Specific activity (titer/ mg protein)	Purification fold
Crude protein extract after ammonium sulfate precipitation	100	1000	25,600	256	1
Sepharose-4B coupled galactose affinity chromatography	10	20	40,960	2,048	8

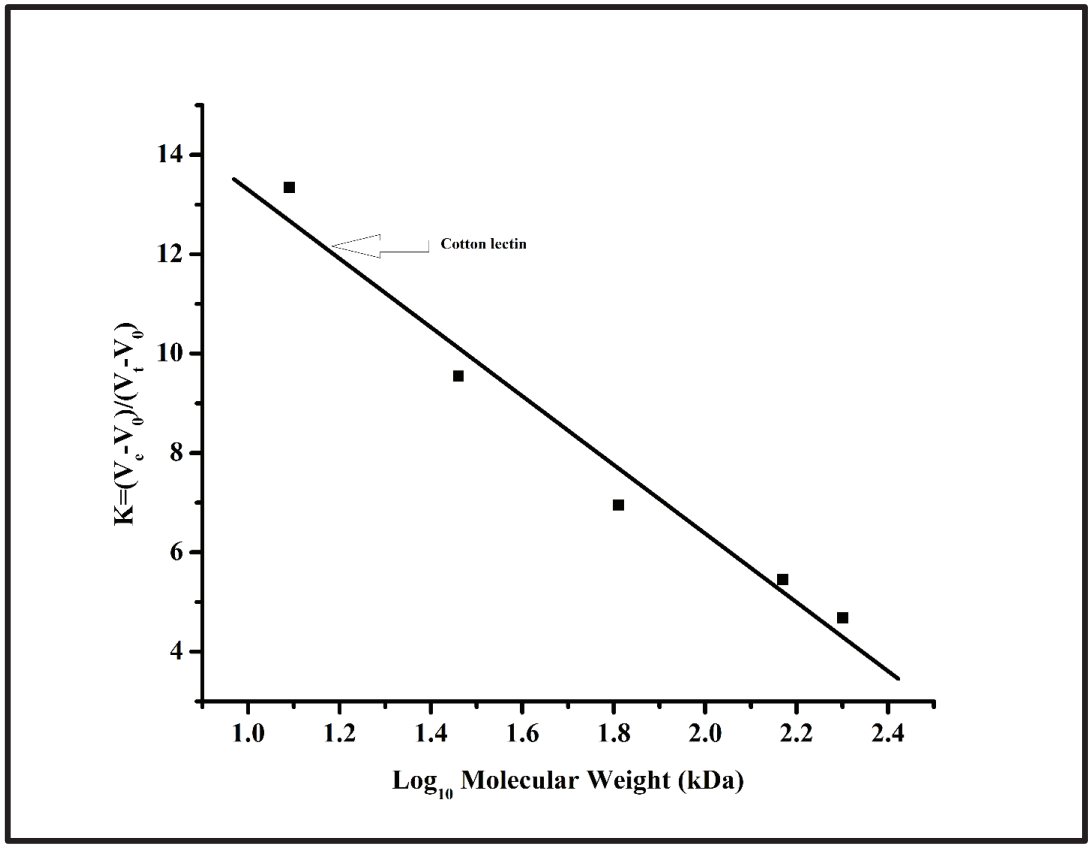


Figure 6.1 C: Molecular weight of CotLec: Plot of K versus log of molecular weight for the estimation of the molecular weight of PotHg and from the plot the mass of PotHg was estimated as ~ 23 kDa. Numbers 1 to 5 suggests different molecular weight markers. 1. β amylase, 2. Alcohol Dehydrogenase, 3. Albumin, 4. Carbonic Anhydrase and 5. Cytochrome C.

6.4.3 N-terminal sequencing

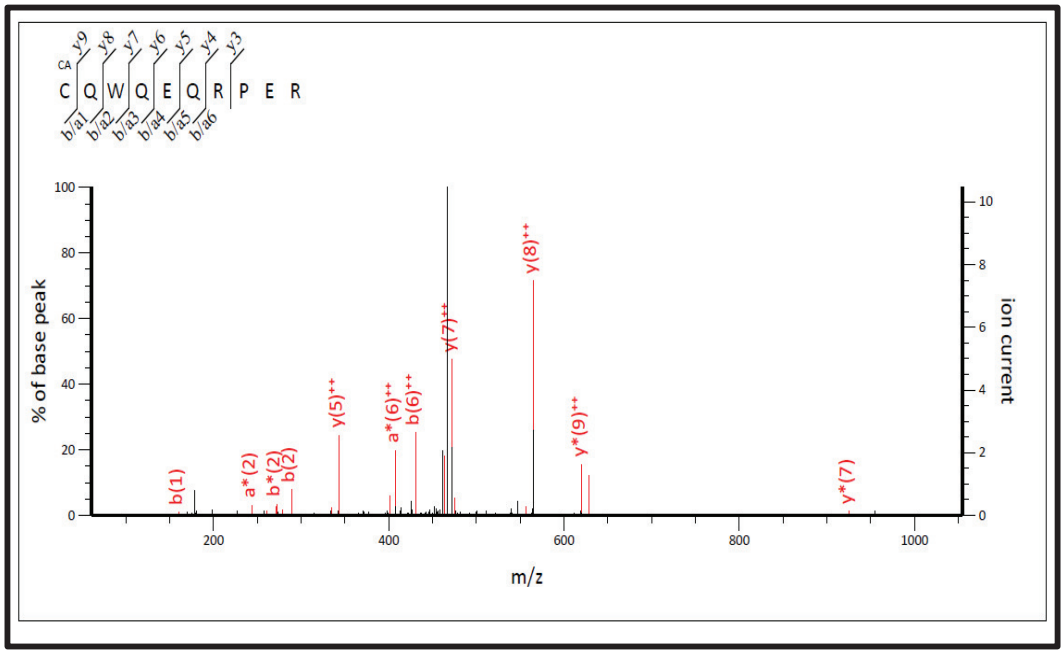
N-terminal amino acid sequence analysis of CotLec gave a sequence of 9 amino acids (Table 6.4). At the 2nd position, there were strong possibilities of the presence of D or E, R, P, F, and K. Similar observation was obtained at position 8th, between R or Y. Based on these observations, a list of possible N-terminal sequences was generated (Table 6.5) and used it for BLAST search. From results of BLAST search, it was concluded that Amino acid D at 2nd position and amino acid R at 8th position showed the highest similarity with cotton sequence (UniProt id: G3M3J3) and thus selected for final analysis.

Table 6.2 N-Terminal sequence of CotLec

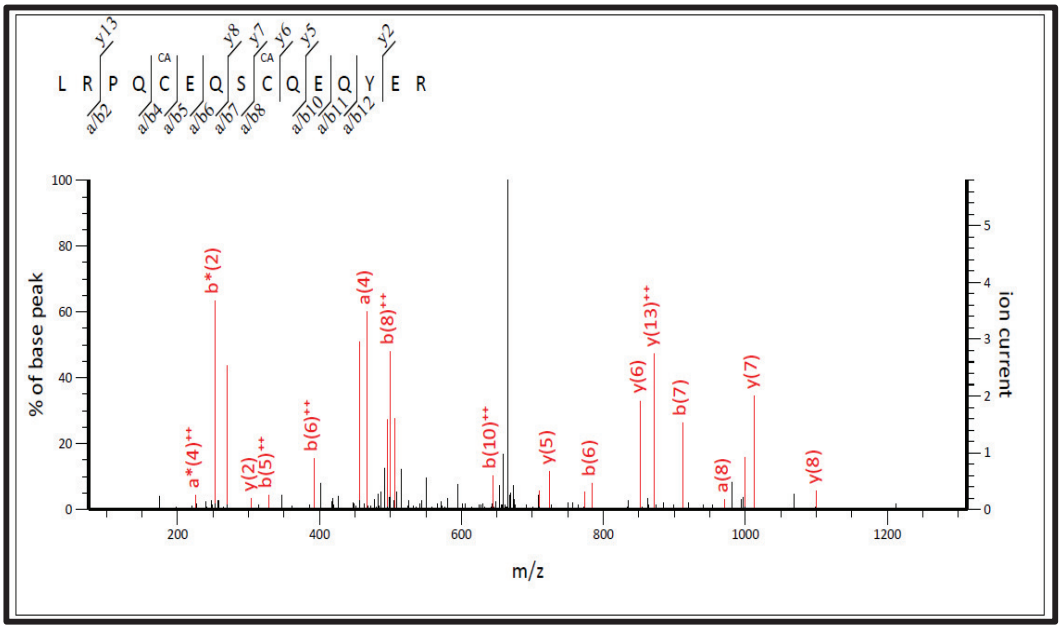
N-terminal Amino acids	Possible combination of N-terminal sequence of CotLec
G, D (or E, R, P, F, K), E, P, Q, R, Q, R (or Y), Y	**GDEPQRQRY**, GEEPQRQRY, GREPQRQRY, GPEPQRQRY, GFEPQRQRY, GKEPQRQRY, GDEPQRQYY, GEEPQRQYY, GREPQRQYY, GPEPQRQYY, GFEPQRQYY, GKEPQRQYY.

6.4.4 Peptide Mass Fingerprinting

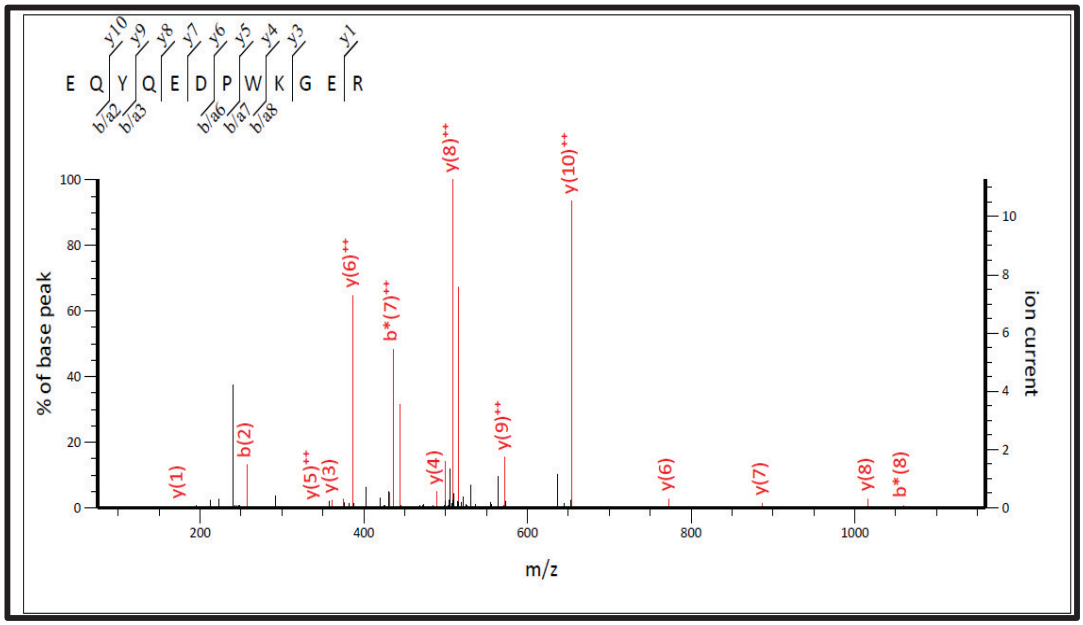
To fully characterize CotLec in terms of sequence, Peptide mass finger printing was carried out. After tryptic digestion of CotLec, peptides were analyzed using MALDI-TOF. After this run, three peptides were selected, and MS-MS (Collision-induced dissociation-CID) experiments were carried out to find out the sequences from a database search. To find out the sequence $[M+H]^+$ ions of m/z 472. 9179, 522.2544, and 670.9851 were fragmented. The analysis of sequence was performed in a survey scan in which analyzer scan in defined mass range and when an intense peak appears analyzer open an additional channel and fragments these ions. In the case of m/Z 472.9179, the sequence CQWQEQRPER was determined (Figure 6.2 A). For m/z 670.9851 and 522.2544 sequences LRPQCEQSCQEQYER and EQYQEDPWKGER were determined, respectively (Figure 6.2 B)


CA
C Q W Q E Q R P E R
y9 y8 y7 y6 y5 y4 y3
b/a1 b/a2 b/a3 b/a4 b/a5 b/a6
% of base peak
ion current
m/z
b(1)
a*(2)
b*(2)
b(2)
y(5)++
a*(6)++
b(6)++
y(7)++
y(8)++
y*(9)++
y*(7)

(A)


L R P Q C E Q S C Q E Q Y E R
CA
CA
y13 y8 y7 y6 y5 y2
a/b2 a/b4 a/b5 a/b6 a/b7 a/b8 a/b10 a/b11 a/b12
% of base peak
ion current
m/z
a*(4)++
b*(2)
y(2)
b(5)++
b(6)++
a(4)
b(8)++
b(10)++
y(5)
b(6)
y(6)
y(13)++
b(7)
a(8)
y(7)
y(8)

(B)



(C)

Figure 6.2 MS-MS analysis of CotLec (A) Fragment spectrum of the $[M+H]^+$ion 472.9179 with the assignment of the peaks to the sequence CQWQEQRPER (B) Fragment spectrum of the $[M+H]^+$ ion 670.9851 with the assignment of the peaks to the sequence LRPQCEQSCQEQYER. (C) Fragment spectrum of the $[M+H]^+$ ion 522.2544 with the assignment of the peaks to the sequence EQYQEDPWKGER.

6.4.5 *De novo* sequencing of CotLec

Analysis of the *De novo* sequence of CotLec showed the presence of 136 peptide sequence spanning 35 proteins in SwissProt database. Out of 136 peptides, only two (LRPHCEQSCR and CQWQEQR) were covering sequence G3M3J3 with a confidence score of 77 and 75, respectively.

6.4.6 Hemagglutination and carbohydrate specificity of CotLec

Agglutination was observed after adding the purified CotLec to 2% solution of RBCs, the minimal concentration of lectin required to obtain hemagglutination was 125µg/ml. Preliminary carbohydrate inhibition assay showed that hemagglutination activity of CotLec was inhibited by galactose, lactose, and fetuin (Table 6.3).

Table 6.3 Inhibition of Hemagglutination activity of CotLec

Carbohydrate	Concentration	Inhibition
Mannose	100 mM	No Inhibition
Galactose	10 mM	**Inhibition**
Lactose	20 mM	**Inhibition**
Glucose	100 mM	No Inhibition
N-Acetyl-D-galactose amine	100 mM	No Inhibition
N-Acetyl-D-glucose amine	100 mM	No Inhibition
Methyl-α-D-mannopyranoside	100 mM	No Inhibition
Methyl-α-D-glucopyaranoside	100 mM	No Inhibition
Maltose	100 mM	No Inhibition
Fetuin	0.1%	**Inhibition**

6.4.7 Effect of pH and Temperature on hemagglutination

The hemagglutination activity of CotLec was retained till 65°C (Figure 6.3 A), at 70°C the hemagglutination was not complete (partial sedimentation of RBCs), above these temperatures, it did not show hemagglutination activity. It retained its activity between pH 3-9 (Figure 6.3 B), below pH 3 and above pH 9 protein was precipitated, after centrifugation of precipitated samples, it did not show any activity indicating aggregation

of protein due to denaturation.

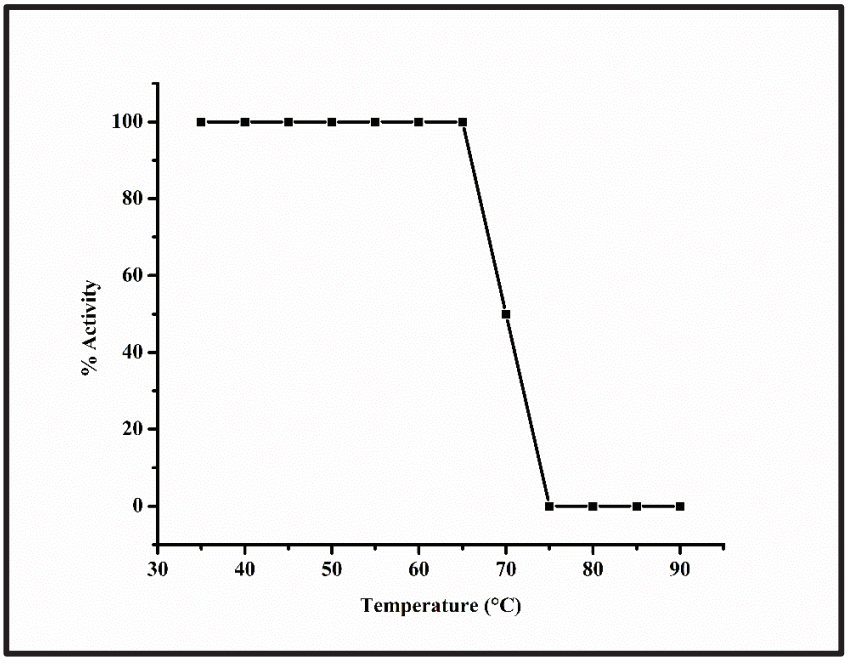


Figure 6.3 A Effect of Temperature on hemagglutination activity of CotLec.

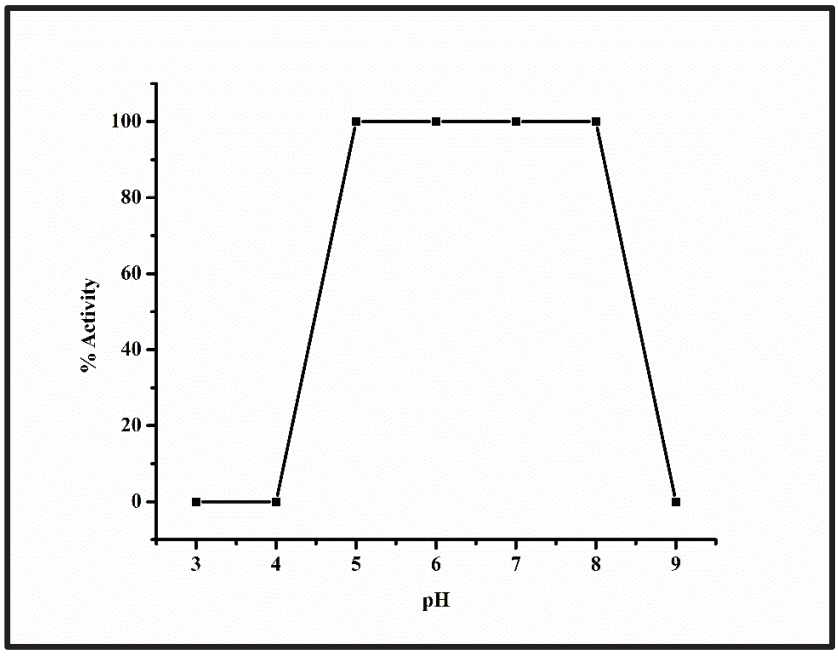


Figure 6.3 B Effect of pH on hemagglutination activity of CotLec

6.4.8 Fluorescence spectroscopy of CotLec

Both galactose and lactose binding induced a decrease in fluorescence intensity with emission ʎmax centered at about 340 nm. Titration of CotLec with increasing concentrations of carbohydrates resulted in a decrease in the emission intensity by 25% (Figure 6.4 A and 6.5 A). A plot of the change in fluorescence intensity as a function of added ligand concentration represents the binding curve of titration of CotLec with carbohydrates (Figure 6.4B and 6.5B), and further calculations were made on it. Analysis of fluorescence data with CotLec exhibited a dissociation constant (Kd) of $2.4 \pm 0.3 \text{ X}10^{-3}$ M for galactose (Figure 6.4C and $3.4 \pm 0.5\text{X}10^{-3}$ M for lactose (Figure 6.5C). The dissociation, association constant and change in values of Gibbs free energy are reported (Table 6.4).

Table 6.4 Dissociation constants, Association constants and corresponding Gibb`s free energy values for CotLec with different carbohydrates

Carbohydrate	Dissociation Constant (Kd)	Association Constant (Kb= 1/Kd) M^{-1}	ΔG°b (Kcal.mol-1)
Galactose	$2.4 \pm 0.3 \text{ X}10^{-3}$	$4.166\text{X}10^{2}$	-15.38
Lactose	$3.4 \pm 0.5\text{X}10^{-3}$	$2.943\text{X}10^{2}$	-14.49

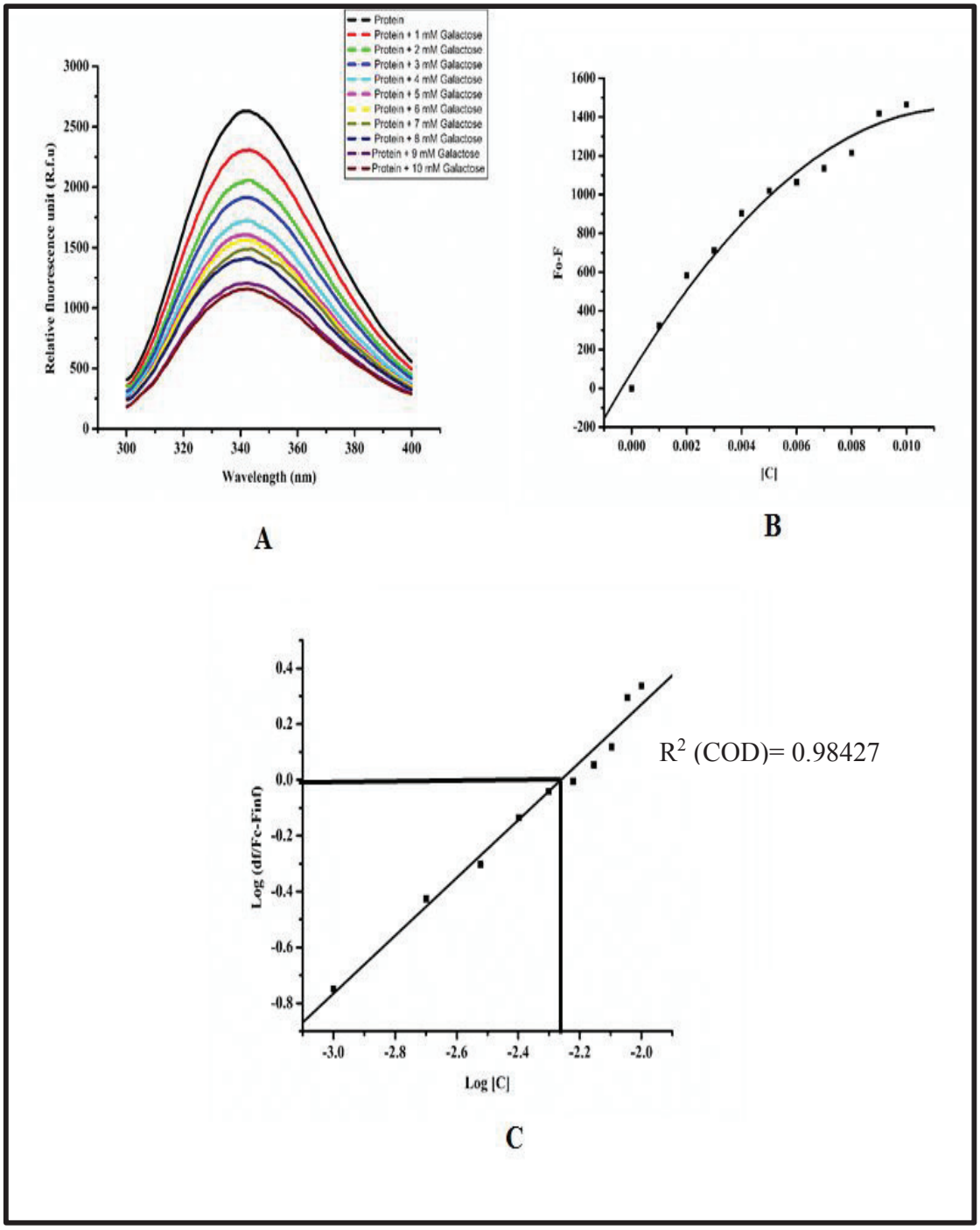


Figure 6.4 Fluorescence spectra of CotLec in the absence and presence of galactose. (A)Represents fluorescence quenching of PotHg on the addition of galactose to the protein solution. (B) The plot of (Fo-F) vs. [C] shows saturation of binding site with increasing conc. Of galactose (C) Double logarithmic plot for binding of galactose with CotLec

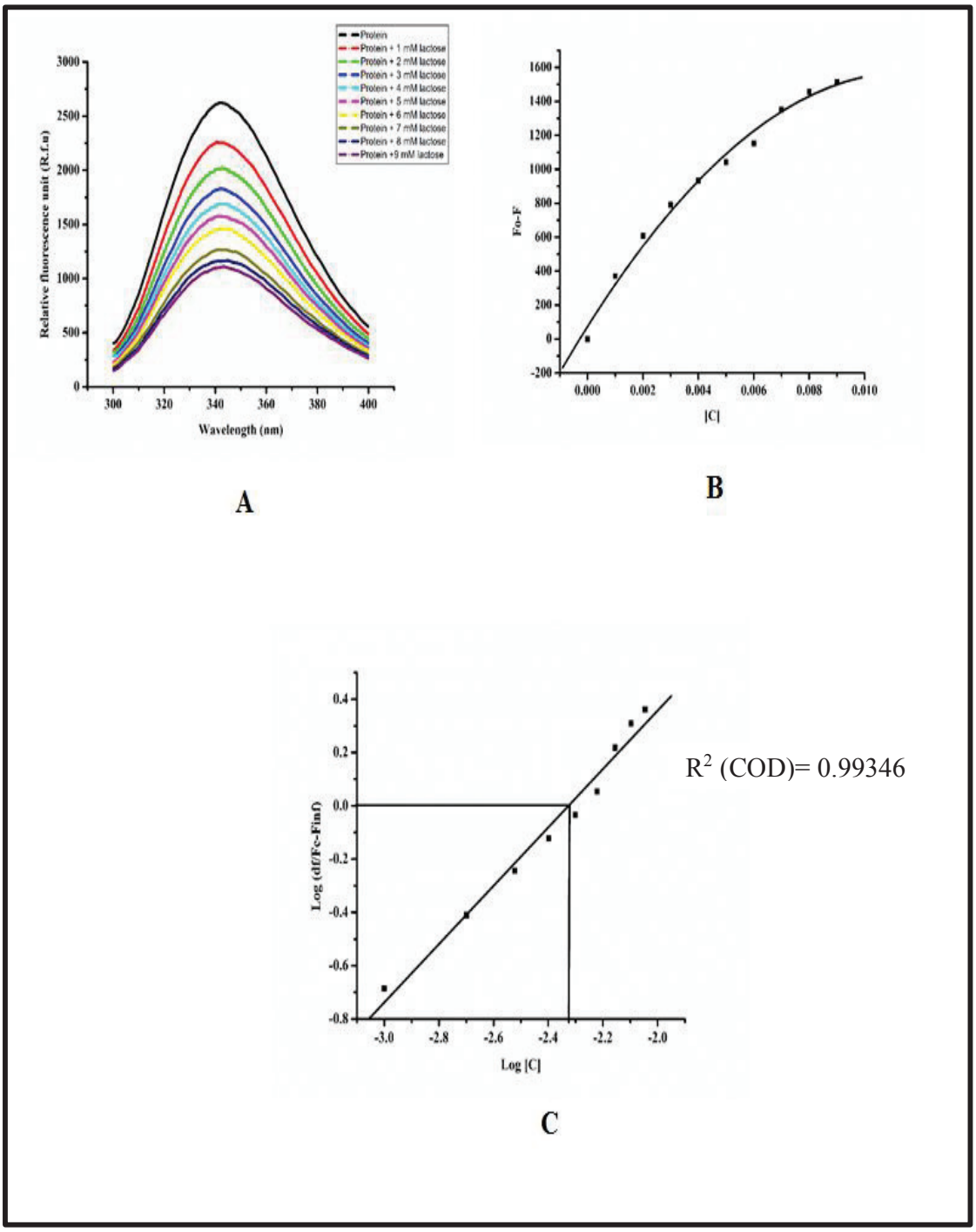


Figure 6.5 Fluorescence spectra of CotLec in presence and absence of lactose (A)Represents fluorescence quenching of PotHg on the addition of galactose to the protein solution. (B) The plot of (Fo-F) vs. [C] shows saturation of binding site with increasing conc. Of galactose (C) Double logarithmic plot for binding of lactose with CotLec.

6.4.9 Circular dichroism spectroscopy

The far-UV (250-200 nm) spectra were recorded at pH 7.4, and it showed typical spectra with zero crossing near 238 nm and 206 nm, a minimum near 220 nm. The overall secondary structure predicted by CDPro suite, which applies three different programs, CDSSTR, CONTILL and SELCON3 are given (Table 6.4). The secondary structural elements of CotLec predicted from CDSSTR are ~16 % α-helices, ~35% β- sheets and ~ 25% β- turns and ~ 25% unordered structure.

Table 6.5 Calculated secondary structure elements of CotLec using CDpro suite

Method	α%	β%	Turns	Unordered
CDSSTR	16.6	35.1	24.8	24.5
CONTINLL	15.3	33.2	29.6	21.9
SELCON3	17.9	30.0	27.1	25.0
Average	16.6	32.7	27.1	23.8

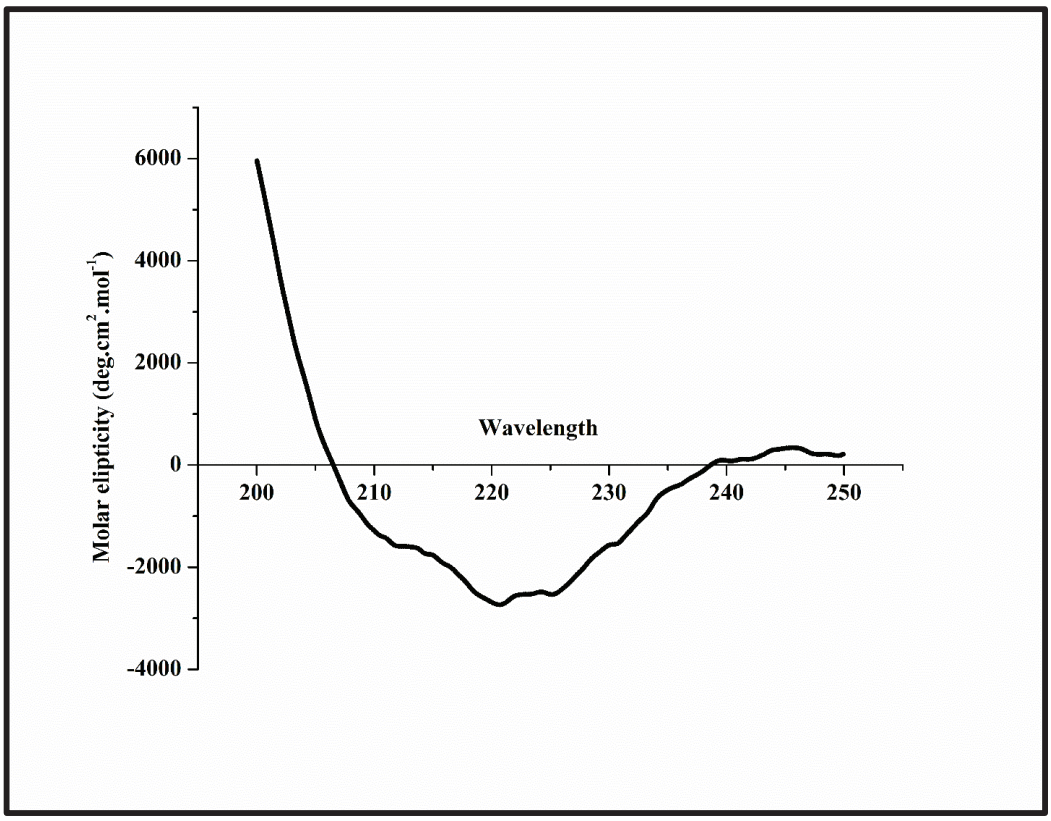


Figure 6.6 Circular dasichroism (CD) spectra of CotLec. Far-UV spectra (250-200 nm) of CotLec (5 µM).

6.4.10 Bioinformatics analysis of CotLec

6.4.10.1 Sequence identification and analysis

BLAST Search using the amino acids obtained in N-terminal sequencing as a query sequence against *Gossypium* species (Id:3633) resulted in output sequences containing seed storage protein vicilin A with a sequence identity of 78% and E-value of 5.3 (Table 6.6). Vicilin sequences of from *Gossypium* family with a sequence identity of ~ 50% with E-values in the range of 4e-^{05} were obtained on further refining the search with peptide fragments obtained through de novo sequencing (Table 6.7). On the basis of this analysis, the sequence id G3M3J3 is predicted as the putative sequence of CotLec.

Table 6.6 BLAST analysis of N-terminal sequence of CotLec

Accession Id	Sequence Name	Identity	Max. Score	E-value
AEO27684.1	Seed storage protein vicilin A (*Gossypium hirsutum*)	78%	23.5	5.3
AEO27683.1	Seed storage protein vicilin A (*Gossypium arboreum*)	78%	23.5	5.3
AEO27682.1	Seed storage protein vicilin A (*Gossypium herbaceum*)	78%	23.5	5.3

Table 6.7 BLAST analysis of N-terminal sequence and peptide fragments from De novo sequencing

Accession Id	Result	Identity	Max. Score	E-value
XP_012483137.1	Predicted: Uncharacterized protein (*Gossypium raimondii*)	49%	49.7	4e-05
P09799.1	Vicilin GC 72-A Alpha-globulin A (*Gossypium hirstum)*	49%	49.3	4e-05
AEO27682.1	Seed storage protein vicilin A (Gossypium arboreum)	40%	45	4e-05

The hits obtained from BLAST analysis were all in the molecular range of 50-70 kDa. Since it is well known that many vicilin undergoes proteolytic cleavage as part of a post-translational modification to generate vicilin fragments with lower molecular weights. The previous study by Chlan et al., (Chlan et al., 1987) suggested proteolytic cleavage sites of pre-vicilin from *Gossypium hirustum.* In this study, they suggested the presence of two potential cleavage sites, Arg-Glu and Arg-Ser. These sites are immediately preceded by another arginine residue in the sequence. Since chose a sequence from the analysis for this study contained only Arg-Ser site, it was considered as a cleavage site. Also selection of

Arg-Ser site fulfill the criteria of molecular weight of ~ 14 kDa (Figure 6.8). The putative processed sequence of CotLec (Figure 6.9) shows a vicilin_N domain, with C-X-X-X-C-$X_{(10\text{-}12)}$-C-X-X-X-C signature motif.

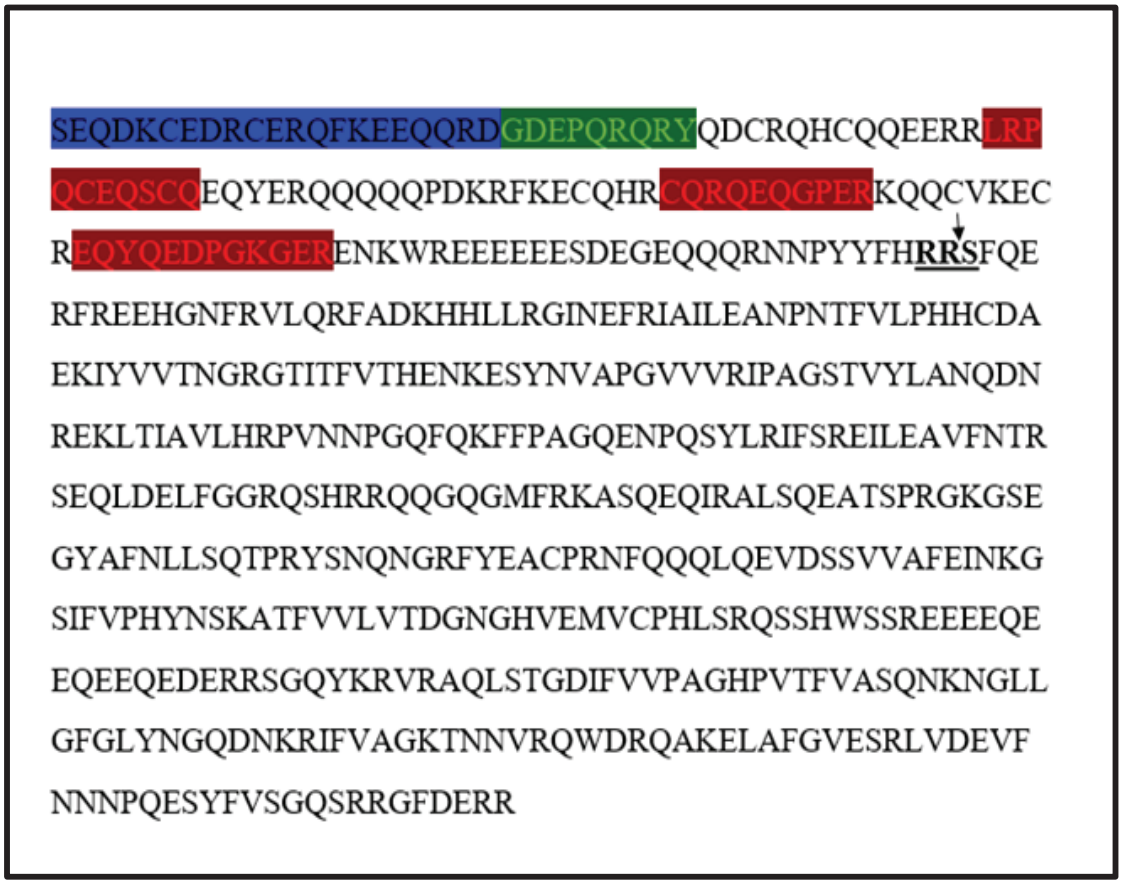


Figure 6.7 Putative processed sequence of CotLec based on the sequence id G3M3J3 identified through BLAST analysis: Precursor sequence of CotLec is in black letters. Amino acids which form the prepropeptide are shown in blue. Amino acids in green are an N-terminal sequence of mature CotLec. The arrow between blue and green amino acid shows the site where protease cleaves and generate N-terminal of the mature polypeptide. Amino acids in red are obtained after tryptic digestion of CotLec. Amino acid shown in bold and underlined character forms the part of C-terminal of CotLec. Arrow shows the site between amino acid Arg-Ser, where protease cleaves the peptide bond to form the mature CotLec

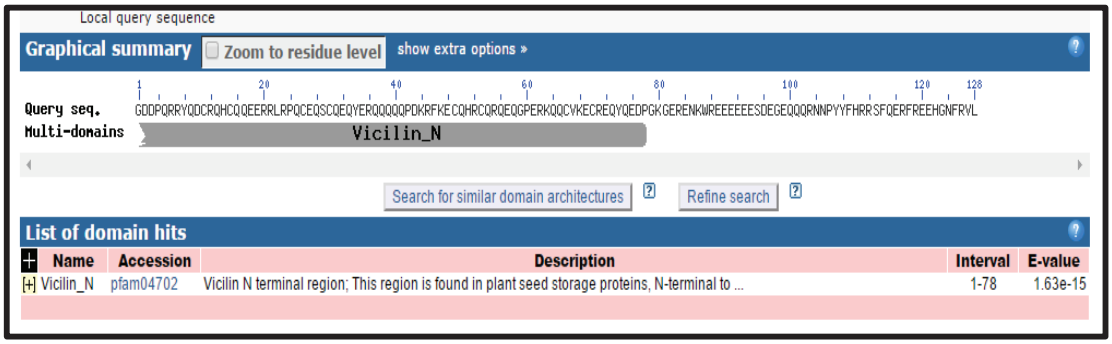


Figure 6.8 Conserved domain analysis of CotLec. Output from CDD search of CotLec: CDD Database search of CotLec as query sequence showed it contained the vicilin-N domain.

6.4.10.2 Multiple sequence alignment

Vicilins in plants can be divided into two categories based on the presence of N-terminal hydrophilic region consisting of the signature motif of C-X-X-X-C-$X_{(10-12)}$-C-X-X-X-C. In this study seven sequences which were processed; *Gossypium hirsutum* (P09799), *Gossypium aerboreum* (G3M3J3), *Gossypium hirsutum* (P09801), *Theobroma cacao* (Q43358), *Macadamia integrifolia* (Q9SPL5), Sucrose binding protein from *Glycine max* (Q04672) and *Lupinus angustifolius* (F5B8W3) were analyzed. Of all the seven sequences only *Lupinus angustifolius* did not contain any signature motif though it had post-translational processing, while sucrose binding protein from *Glycine max* contained only one signature motif, rest of the proteins have two or more such signature motif (Figure 6.9).

Another feature of this analysis was the C-terminal regions of all the proteins contained all most identical amino acids. All the proteins except, vicilin from *Lupinus* angustifolius and sucrose binding protein from *Glycine max* did not have putative processing site of Arg-Ser and rest of the proteins possessed potential processing site Arg-Ser.

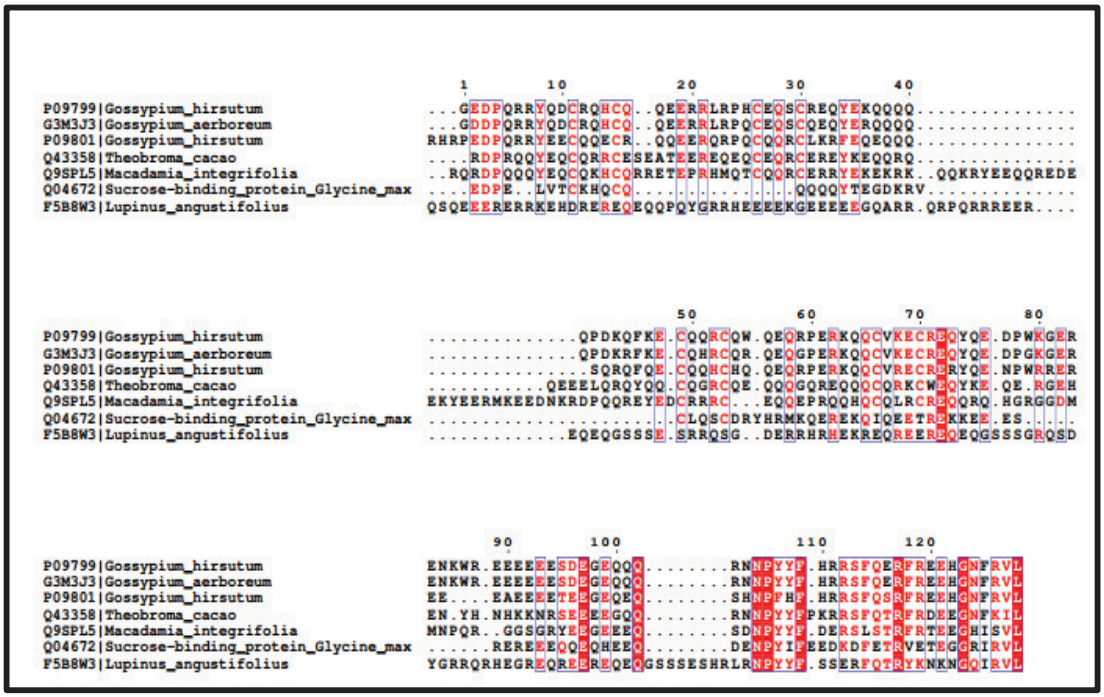


Figure 6.9 Multiples sequence alignment of CotLec CotLec with other vicilin having N-terminal hydrophilic region and signature motif

6.4.10.3 Phylogenetic analyses of CotLec

The phylogenetic analysis was carried out among vicilin having N-terminal hydrophilic regions, vicilin not having such N-terminal hydrophilic region and known vicilin with lectin-like activity. The analysis produced two major clades, one that contained all the vicilin having an N-terminal hydrophilic region and a sucrose binding protein (Figure 6.11). The second clade contained vicilin with lectin-like activity and vicilin not having N-terminal hydrophilic region.

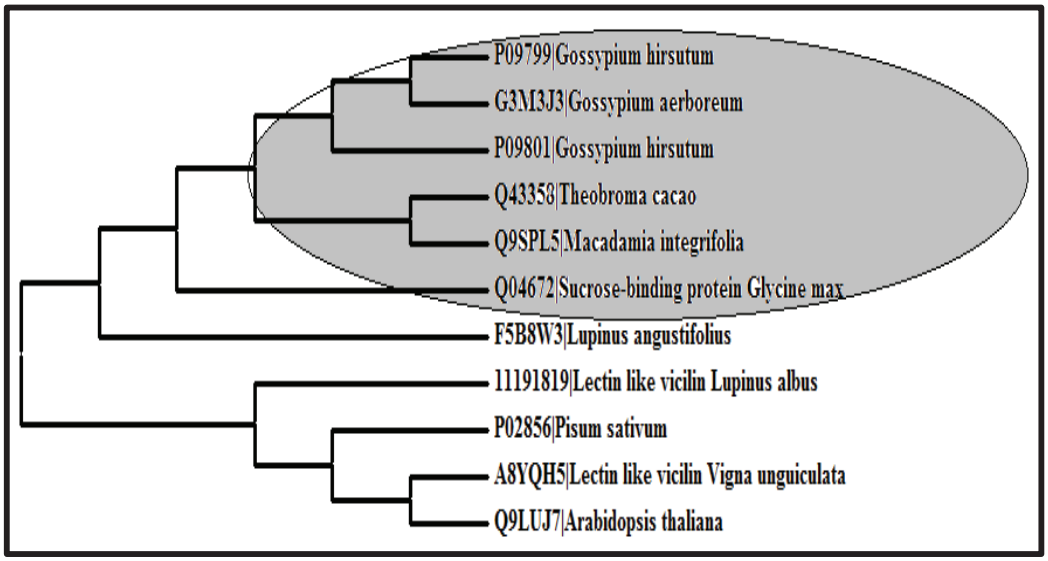


Figure 6.10 Phylogenetic analysis of CotLec The phylogenetic analysis of CotLec showed two distinct clusters. One cluster contained all the sequence with the N-terminal hydrophilic region and signature motif (Shown in gray oval), while rest of the vicilin and vicilin with known lectin-like activity was clustered separately.

6.4.10.4 Molecular Modelling

To get the idea of the 3-D structure of CotLec, full-length sequence of CotLec was submitted to automated protein modeling through I-TASSER server. The Standard output from I-TASSER server contained four models with different C-score, the model with C-score of -1.68 was selected for further analysis. The selected model was checked for quality through Ramachandra plot analysis, which revealed 84.7% residues in the favored region, 11.7% in allowed region and 3.6% in outlier region, respectively. The model contained eight α-helices connected by loops. The structural homology search of CotLec through DALI server showed it shared homology with Apoptosis regulator (PDB Id: 4OYD), Ribosome recycling factor (PBD Id: 1EK8) and spectrin alpha chain (PDB Id: 3LBX) with % identity of 12, 6 and 10, respectively. None of the homolog found in DALI server search belonged to vicilin or any other seed storage family.

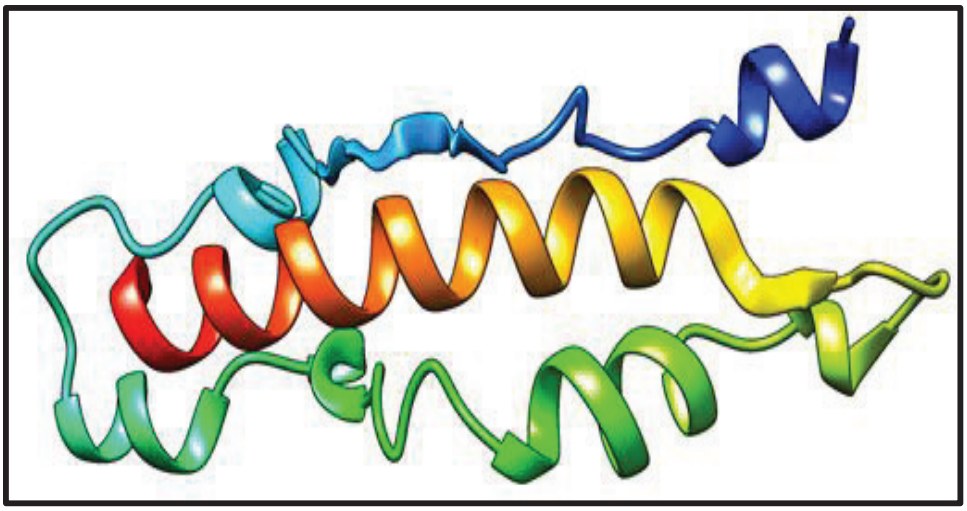


Figure 6.11 Homology modelling of CotLec. A monomeric model of CotLec based on template PDB ID: 4MH6.

6.5 Discussion

Cotton has been used for fiber and fabric production Since 6000 BC and is considered as most important fiber crop across the globe due its importance in textile industry. The growth and thus productivity of cotton plants are frequently fraught with biotic, and abiotic stress such as pests and pathogens, salinity, heat, and draughts and it has developed a sophisticated defense mechanism. One of the major contributors of this sophisticated defense mechanism is proteins expressed in the seeds and cotton seeds contain 17-24 % of oil and 40-43% of proteins (Zhou, Sun et al. 2014).

Seed storage proteins can be defined as the reservoir of metal ions and amino acids which are utilized by organisms to sustain and flourish life, and they also play a very important role in plant defense by their insecticidal and antimicrobial activities. Storage proteins are associated with different plant component responsible for protein synthesis and storage such as seeds, kernel, roots or tubers. These organs consist high percentage of protein as their dry weights some time as high as up to 95% (Candido et al., 2011). Expression of storage proteins is regulated at different levels and stored in protein bodies or vacuoles, which prevents it from premature degradation. When seed germination starts these proteins are rapidly degraded and provides necessary nutrition for growth. Plant

storage proteins can be subdivided into two categories: Seed storage proteins (SSPs) and Vegetative storage proteins (VSPs) and in some plant seeds, these proteins constitute about 40% of its dry weight and provides the major source of proteins for human consumption (Candido et al., 2011).

The seed contains proteins of several kinds, which are classified into four major categories based on their solubility. i.e. (1) Albumins: soluble in water, (2) globulins: soluble in saline and mostly find in legume seeds, (3) prolamins: Alcohol/water mixture mostly present in cereal seeds and (4) glutelins: alkali/diluted acid present in cereals seeds. Other proteins are used as a source of carbon and nitrogen and are associated with defense mechanism, and members include protease inhibitors, lectin, lectin-like proteins, ribosome inactivating proteins. Among these proteins, globulins are the most diversely distributed protein family in seeds, and it can be subdivided into two major class based on their sedimentation coefficient (1) 7S type or widely known as vicilin and (2) 11 S or widely known as legumins (Shwrey et al., 1995).

7S globulins or vicilin are present in both monocots and dicots. They are generally found in the trimeric state with and oligomeric molecular mass of 150,000 to 200,000 with a typical subunit mass of ~ 50 kDa (Candido et al., 2011). The subunit structure revealed by SDS-PAGE is similar in absence and presence of reducing agents which suggest the absence of disulfide linkages. Glycosylation and proteolytic cleavages are often part of post-translational modification process of vicilin and N-terminal and C-terminal sequences of vicilin shows homology which is attributed to gene duplication event. N-terminal consists of 50 amino acid repeat with each of these repeats containing a signature motif of $C\text{-}X_3\text{-}C\text{-}X_{10\text{-}12}\text{-}C\text{-}X_3\text{-}C$ and (Marcus et al., 1999). C-terminal domain of vicilin belongs to cupin superfamily which is one of the most diverse protein families and includes proteins with enzymatic functions, transcription factors, and seed storage proteins (Dunwell et al., 2001).

Lectins are carbohydrate-binding proteins or glycoproteins of non-immune origin. The selectively and reversibly bind to mono or oligosaccharides without altering their structure. Plant lectins have varied biological roles such as plant defense and cell signaling and are shown to elicit antiviral, antibacterial and antitumor response. They are also

abundantly expressed in seeds and are sometimes considered as seed storage proteins (Chrispeels et al., 1991).

Over the last few years, there have been reports characterizing vicilin with lectin activity and some examples are vicilin showing affinity for chitin and isolated from *Albizia lebbeck* (Souza et al., 2012), *Enterolobium contortisiliquum* (Moura et al., 2007), and *Erythrina velutinous* (Macedo et al., 2008).

In continuation of our studies on characterization of novel lectins (Nair et al., 2012; Shah et al., 2016), the present study describes the isolation, purification, and characterization of a lectin termed CotLec from seeds of *Gossypium arboreum*. Sequence analysis proved CotLec to be a vicilin type seed storage protein. There is no previous report on the detailed characterization of a lectin from seeds of cotton barring one, where they have shown the presence of a galactose-specific lectin from seeds of cotton with a molecular weight of 65 kDa but no sequence or structure information available for that purified protein (Zheng et al., 1995). CotLec shares sugar specificity with previously characterized lectin from cotton seeds, but molecular weights and glycosylation pattern are different.

On another hand, there are two reports showing the presence of different vicilin from seeds of cotton. One report showed the presence of ~ 93 kDa vicilin from seeds of cotton containing two subunits of 54 and 48 kDa and are held together by hydrophobic interactions (Marshal, H., 1990). Another report showed vicilin from seeds of *Gossypium hirustum* with anti-fungal activity. Purified vicilin was consisting of four different subunits, and total molecular weight of purified protein was ~85 kDa, and it was basic in nature. Out of these four subunits, two were off ~9-10 kDa; one was ~ 16 kDa, and one was ~ 46 kDa. CotLec shares high percentage (~ 90%) N-terminal sequence similarity with one the ~9-10 kDa fractions (Chung et al., 1997).

We obtained 20 mg of pure CotLec from 500 gm of dry seeds. The purified protein was separated on SDS-PAGE in reducing and the non-reduced condition is showing single band, suggesting a lack of disulfide bonds. Reports suggest that vicilin does not form disulfide bond as they lack cysteine in their sequences (Shewry et al., 1995) and though CotLec sequence contained 8 cysteine molecules and yet they do not form a di-sulfide

bond. Gel-filtration studies showed CotLec exists as a monomer, and its molecular weight is ~ 15 kDa, which is consistent with the molecular weight observed in the SDS-PAGE analysis.

Vicilins are typically trimeric proteins with molecular weight of ~ 150000 Da. Their subunit structure varies greatly due to post-translation modification (glycosylation and proteolysis). Many vicilin from plants which did not undergo proteolytic cleavage as part of their post-translational modification are trimeric with subunit weight of ~ 47000 to 50000 Da, and the associations are mostly by non-covalent interactions. An example of such vicilin is Canavalin from jack bean, Phaseolin from French bean and β-conglycinin from soybean (Ko et al., 2000; Lawrence et al., 1990 & Maruyama et al., 2003). There has also been a report of the presence of a low molecular weight vicilin from different plants. One of the well-characterized vicilin of low molecular weight (~ 11 kDa) is from the seeds of *C. lanatus*. This low molecular weight vicilin exist as a monomer just like CotLec, but at higher concentration, they form aggregates (Yadav et al., 2011). These suggest, smaller molecular weight vicilin lack the ability to associate and forms multimeric protein in solution.

CotLec did not show staining by Acid-Schiff staining method suggesting a lack of glycosylation and analysis of CotLec sequence for a potential site for N-glycosylation via online tool NetNGlyc 1.0 server also indicated no potential site for N-linked glycosylation. These results are supported by the fact that CotLec lacks a consensus sequence Asn-X-Ser/Thr, which is necessary for N-linked glycosylation. Glycosylation is frequently observed in vicilin via the addition of N-linked complex glycan, but it is not necessary for proper folding or export to protein storage vacuoles (Sturm et al., 1987).

N-terminal sequence and peptide sequence obtained from MS/MS and *De novo* protein sequencing analysis confirmed CotLec is a vicilin A seed storage protein from *Gossypium arboreum* (Uniprot Id: G3M3J3). The sequence available in Uniprot database is obtained from cDNA sequence, and it contains 537 amino acids. The vicilin or 7S storage proteins in plants are believed to undergo a series of post-translational modification which includes the addition of N-linked complex glycan during glycosylation, and limited proteolysis of vicilin (Sherwry et al., 1995). As per our analysis, the N-terminus of CotLec started from G^{22} of the amino acid sequence deduced from cDNA sequence (Hu et al.,

2011) (Figure 6.8). First 21 amino acids of deducing sequence do not form signal peptide according to Signal P analysis, yet in post-translational modification, they are removed suggesting they are part of the propeptide sequence. Till date, there is no previous literature showing the presence of such propeptide in other known vicilin and the putative role of this propeptide is still unknown. The processing of C-terminal is based on potential cleavage sites Arg-Arg/Arg and Arg-Arg/ser which is present in hydrophilic N-terminal regions (Chlan et al., 1987). Cleavage at this site generate proteins with ~11 kDa (Arg-Arg/Arg) and ~15 kDa (Arg-Arg/Ser). We consider cleavage site Arg-Arg/ser since the molecular weight of CotLec obtained from SDS-PAGE and gel filtration analysis matches the estimated weight of protein generated by cleavage at this particular site. Also, Arg-Arg/Arg site is missing in CotLec sequence (Figure 6.8).

Vicilin as a protein family is very diverse in nature, and several domains have been recognized in vicilin proteins. One of such domain in many plant vicilin including peanut, barley, maize, cocoa, pumpkin and macadamia nut are cysteine-rich hydrophilic region contiguous to the N-terminal region with the signature motif of C-X-X-X-C-X $_{(10\text{-}12)}$-C-X-X-X-C. In this chapter, we have shown that the vicilin protein from cotton seeds (*Gossypium arboreum*) contains cysteine-rich hydrophilic N- proximal region. CotLec post-translationally processed by proteolysis and has lectin activity.

Many vicilin containing cysteine-rich hydrophilic N-proximal regions like CotLec undergo processing by proteolysis in a different way. In vicilin from plants macadamia nut and pumpkin, proteolysis generates smaller molecular weight (~ 7 kDa) vicilin compared to CotLec. Detailed analysis in previous studies by Yamada et al., showed multiple functional proteins are produced by cleavage at particular site Asn-Gln (Yamada et al., 1999). Multiple sequence analysis of vicilin from macadamia nuts, pumpkin, and CotLec showed that both macadamia nuts and pumpkin contained Asn-Gln in their sequence at multiple positions while CotLec lacks Asn-Gln pairs in its sequence.

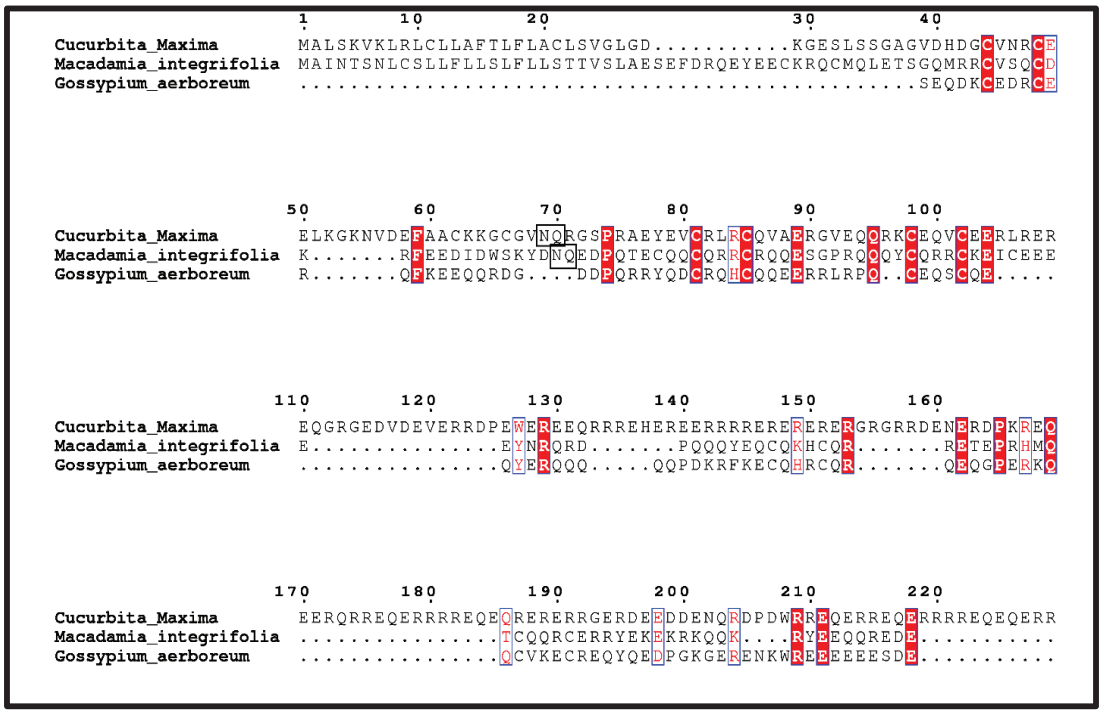


Figure 6.12 Multiple sequence alignment of CotLec with other vicilin showing putative processing site
Sequences of *Cucurbita maxima* (pumpkin) and *Macadamia integrifolia* (Macadamia nuts) contained site Asn-Gln leading to the generation of smaller molecular weight (~ 7 kDa) vicilin. CotLec devoid of this site does not form smaller molecular weight vicilin

Fluorescence spectroscopy of CotLec with galactose and lactose showed maximum intensity centered around 340 nm, and it decreases with increasing concentration of carbohydrates, suggesting quenching of tryptophan fluorescence. This observation shows the direct or indirect role of tryptophan in carbohydrate recognition. The dissociation constant values of galactose and lactose with CotLec is in mili-molar range, suggesting weak or non-specific binding, which is the case for most the lectin binding with mono or disaccharides (Varki et al., 2009).

Analysis of CD spectrum of CotLec through a different program of CDPro suite showed it contained ~ 15% α-helices and ~ 30% β-strands. Analysis of other vicilin structures deposited in PDB showed similar % of α-helices and β-strands in their structure. For example, a recent structure of vicilin from *Solanum melongena* consisted of 22% α-helices and ~ 37% β-strands (Jain et al., 2016).

Molecular model of CotLec generated from I-TASSER server was consisting only of α-helices, which is in contradiction from the results obtained by CD Spectroscopy. CD Spectroscopy data showed, CotLec contained a mixture of both α-helices and β-strands, which is consistent with already solved vicilin structures from other plants. The possible reason I-TASSER server did not select any of the known vicilin as a template to build the model might be due to low sequence identity (~5-10%) of mature CotLec with known vicilin structure in PDB.

Cotton crop plays a very important role in agriculture production and economy across the globe, especially in India. Over the past, decades with the introduction of Bt-cotton and better paste management has led to an increase of production many folds from 13 million bales to 40 million bales. In 2013, a total of 11 million hectares of land was under Bt-cotton cultivation (James Clive, 2014). Bt-cotton has several advantages compared to non bt-cotton including, an increase in yield, reduction in pesticide use, reduction in the cost of cultivation and no adverse effect on beneficial insects (Khadi B.M., 2011). With all the benefits and increase in cultivation area changes in insect pest complex is evident. Outbreaks of mealy bugs and mirid bugs, which wreak havoc and affected the production seen as a major challenge against BT- cotton in India (Khadi B. M., 2011). Technology including newer varieties of BT-cotton is being found, another way to strengthen the cotton own defense mechanism could be the possible way to find the long term solutions in protecting the important crop plant against pests.

6.6 Conclusion

In this chapter, we have Isolated, purified and identify a novel protein from the seeds of Cotton with lectin activity. The lectins showed specificity towards galactose and lactose and it is a monomer of ~ 15 kDa. Despite rich in cysteine in its sequence CotLec does not form disulfide bonds and lack glycosylation. CotLec is generated from the precursor protein after proteolytic cleavage as part of the post-translational modification. It is thermostable, retains its activity in broad pH range. The N-terminal sequencing and peptide mass fingerprinting identified the purified CotLec as seed storage Vicilin protein. Circular Dichroism spectroscopy showed CotLec contains both α- helices and β-strands