Chapter 2

Purification, Biochemical characterization and bioinformatics analysis of a lectin from Solanum tuberosum

2.1 Summary

A protein with sugar affinity and positive activity in hemagglutination assay was isolated from the tubers of *Solanum* tuberosum. It was purified to homogeneity by ammonium sulfate precipitation (80% saturation), followed by affinity chromatography using sepharose4B-fetuine and ion exchange chromatography steps using DEAE cellulose and It is a heterodimer of molecular mass ~15 and ~5 kDa as identified through SDS-PAGE. The yield of the protein was about 2mg per 1000g of the tubers with a specific activity of 640 HUmg-1 in the assay and named PotHg. De novo protein sequencing and MS/MS analysis identified it as serine protease inhibitor (UniProt id: P58515), which consist of 186 amino acids and it is a heterodimer of two chains with a large chain of 151 amino acids and short chain of 35 amino acids. PotHg has a theoretical pI of 10 and is stable in the pH range 3-9, up to 24 hours at 25^0C as well as up to 30 minutes at 80^0C. Acid-Schiff staining method shows it to be glycosylated, and gel filtration analysis shows that it has no higher oligomerization in solution. PotHg inhibited the Trypsin with a specific activity of 351 (TUI/mg) thus proving its protease inhibitor activity. The hemagglutination activity of PotHg is inhibited only by glycoprotein fetuin and chemical modification of tryptophan by NBS leads to loss of hemagglutination activity suggesting the involvement of tryptophan in sugar binding directly or indirectly. Homology search through BLAST analysis identified well-known Kunitz-type serine protease inhibitors as close homologs. The modeled structure of the protein showed 12 β-sheets stabilized by two disulfide bonds. Phylogenetic analysis of the sequence of 24 reported Kunitz-type protease inhibitors from *Solanum tuberosum* showed three different clusters and PotHg belong to the cluster of serine protease inhibitors.

2.2 Introduction

Potato (*Solanum tuberosum*) is a member of family Solanace, which includes economically important plants such as tomato, pepper, aubergine (eggplant) and tobacco. Potato is the world`s fourth most important food crop after rice, wheat, and maize but

despite its worldwide importance, every year there is a huge crop loss due to infections by microorganisms and pests (Prakash, 2008). Similar to other plants, potatoes have developed a robust defense mechanism employing defense proteins like lectins, proteases, protease inhibitors (PIs), etc., to overcome infection.

Potato tubers contain about 2% proteins by weight which consist of patatin, storage protein, and low molecular weight PIs (Pots et al., 1999). PIs are ubiquitous, small proteins which are particularly abundant in plant reproductive and storage organs such as seeds and tubers and can account for 1 to 10% of the total protein in storage tissues and can be classified into at least ten different families. (Habib & Fazili, 2007; Ryan, 1990). In potato, the most abundant PI's are inhibitors of serine proteases from families known as a Kunitz-type inhibitor (KTI), potato trypsin inhibitors type I, potato trypsin inhibitors type II and Bowman-Birk inhibitors (Ryan, 1990; Zavala et al., 2004; De Leo et al., 2002). KTIs are one of the best-characterized inhibitors found abundantly in potato tubers and represent a highly diverse group of proteins. Previous studies have revealed that KTIs have plasticity in their structure, which allows them to interact with different partners, simultaneously (Heibges et al., 2003; Azarkan et al., 2011).

Over the past few years there have been reports of proteins possessing both, protease inhibitor and lectin-like activity and which have been isolated from plants and animals and some examples are a protein purified from the seeds of *Peltophorum dubium* (Macedo et al., 2004), a protein isolated from the seeds of *Labramia bojeri* (Fernanda et al., 2003) and a protein from the skin mucus of the Japanese eel, *Anguilla japonica* (Saitoh et al., 2005).

2.3 A review of the techniques used in the analysis

2.3.1 MALDI-TOF analysis for protein identification

MALDI (Matrix-assisted laser desorption ionization time of flight mass spectrometry) mass spectrometry is used for the determination of molecular weight of intact proteins, or it can be used to identify the protein based on peptide mass fingerprinting. A common procedure for protein identification involves protein digestion

with trypsin to obtain peptides which are separated and their molecular masses determined accurately. The masses are compared against databases, and probability based scoring systems are used to determine the closely matching proteins (Domon et al., 2006)

2.3.2 De novo protein sequencing

Peptide de novo sequencing is the analytical process that obtains a peptide's amino acid sequence from its tandem mass spectrum (MS/MS). In a tandem mass spectrometer, the protein is fragmented and depending on the fragmentation methods used; different fragment ion types can be produced. The most widely used fragmentation methods today are Collision-Induced Dissociation (CID) and Electron-Transfer Dissociation (ETD). CID produces mostly b and y-ions, and ETD produces mostly c and z-ions. The mass difference between two fragment ions is used to calculate the mass of an amino acid residue on the peptide backbone, and the mass can usually uniquely determine the residue. For example, the mass difference between the two residues y_7 and y_6 ions is equal to 129, which is the mass of residue E (Glutamic acid). Similarly, the next adjacent residue between y_6 and y_5 can be determined as L by the mass difference. Such a process can be continued until all the residues are determined. Thus, if one can identify either the y-ion or b-ion series in the spectrum, the peptide sequence can be determined (Ma et al., 2003).

2.3.3 Bioinformatics Analysis

2.3.3.1 BLAST for homology search

A sequence similarity or homology search is often used as the first step to finding out the information hidden in DNA or protein sequences. These similarities often help to gain insight into the probable function(s) from similar sequences. There are many ways to do this homology search, but one the most widely used method is “Basic Local Alignment Tool” (BLAST). BLAST identify the perfect or near perfect matching between the biological or query sequences against the already known sequences present in the database of amino acids or nucleotides (Altschul et al., 1990). As the name suggest BLAST performs local alignments, and the quality of a pairwise sequence alignment is evaluated by "substitution matrix," like PAM or BLOSUM, which assigns a score for aligning any

possible pair of residues (Altschul. S.F., 1991). Sequence alignment scores are reported as E-values which is a measure of the probability of a match between the query sequence and the template sequence. In our studies we used protein BLAST (pBLAST) and the query sequence was searched against Non-redundant protein sequence from *Viridiplantae*. Blast search was done using the BLOSUM62 matrix with an expected threshold of 10.

2.3.3.2 Multiple sequence alignment

A Multiple Sequence Alignment (MSA) can be defined as an alignment between three or more biological sequences, and analysis of data obtained from MSA can be helpful in find out conserved regions in sequences. Data obtained from pairwise sequence alignment can be useful to determine whether genetic sequences from two species are evolutionarily related. MSA is useful in constructing phylogenetic trees, protein modeling, understanding substitution rates, etc. The main software for multiple sequence alignment in present work is ClustalW2 at EBI servers and MEGA program. (Larkin et al., 2007; Goujon et al., 2010; McWilliam et al., 2013).

2.3.3.3 Conserved domain identification

The conserved domains can be defined as part of protein sequence and tertiary structure that can evolve, and function independently and is also responsible for the specific function of protein. Several databases are available for identifying the domain region present in a particular protein sequence. Few of the widely used databases are Conserved domain database (CDD), Simple Modular Architecture Research Tool (SMART) and Pfam (Marchler-Bauer et al., 2004; Schultz et al., 1998; Sonnhammer et al., 1997). In this work, we have used Conserved Domain Database (CDD) (Marchler-Bauer et al.2004) which is available at web server NCBI (http:// www.ncbi.nlm.nih.gov/cdd) and Protein family database (Pfam) available at (https:// http://pfam.xfam.org/) to find out conserved domain present in the purified novel lectins.

2.3.3.4 Phylogenetic analysis

Phylogenetic analysis is use to understand the evolutionary history and relationships among individuals or group of organisms. The information obtained from the

phylogenetic analysis is usually portrayed as tree-like diagrams known as dendrograms, that represents an estimated lineage of the inherited relationships among the sequences, organisms or both. Various algorithms like distance matrix method, maximum parsimony, maximum likelihood and Bayesian phylogeny can be utilized for such studies. The maximum parsimony method available in Molecular Evolutionary Genetics Analysis (MEGA) suite, (Tamura et al., 2011) was used for the phylogenetic analysis studies in this work. Maximum parsimony method is also known as minimum evolution method because it predicts the evolutionary tree that minimizes the number of steps required to generate the observed variation in the sequences from common ancestral sequences. One of the methods to test the validity of such trees is Bootstrapping. Bootstrapping tests the reliability of an inferred tree by Felsenstein's bootstrap test (Felsenstein. J., 1985), which uses Efron's bootstrap resampling technique (Efron et al., 1996). In this test, once the construction of a phylogenetic tree with the selected method is done, a preselected number of residues are randomly shuffled to generate a new set of sequences and generate a new phylogenetic tree. The topology of this new tree is compared to that of the original tree and each interior branch of the original tree that is different from the bootstrap tree. The sequence partitions are given a score of 0, and all other interior branches have a value 1. This procedure of resampling the sites and the subsequent tree reconstruction is repeated several hundred times. The percentage of times each interior branch is given a value of 1 which is known as the bootstrap value. As a general rule, if the value of internal branch is 95% or higher than the topology at the branch is correct.

2.3.3.5 Structure Prediction

The polypeptide backbones of proteins exist in particular conformations, called secondary structures. The secondary structure of the proteins was predicted using PSIpred (Buchan et al., 2013) (http://bioinf.cs.ucl.ac.uk/psipred/). The tertiary structure predictions are mainly based on three types of mechanism: homology modeling, (threading) and *ab initio* prediction. In the homology modeling, the query sequence is aligned properly on the known template sequence whose tertiary structure data is available. One of the widely used homology-based structure prediction programs is SWISSMODEL (Arnold et al., 2006). Other methods have fold recognition and prediction algorithms like Phyre and I-TASSER

(Kelley et al., 2015; Yang et al., 2015). Ab initio or de novo protein modeling methods seek to build three-dimensional protein models based on the physical principles rather than previously solved structures and the software used are HMMSTR/Rosetta (Bystroff et al., 2000; Bystroff and Shao, 2002). In this work, I-TASSER server (http://zhanglab.ccmb.med.umich.edu/I-TASSER/search.html) which is an on-line tool for protein structure and function predictions have been used to generate 3D models.

2.3.3.6 Structural homologies and Superpose

Structural homology is a very important way to bridge the gap between some protein sequence available in UniProt and number of protein structure available in PDB. The structural alignment and homology search were carried out by DALI server in this work. Dali database is based on all-against-all 3D structure comparison of protein structures in the Protein Data Bank (PDB) (https://rcsb.org) (Berman et al., 2000). Superpose (http://wishart.biology.ualberta.ca/SuperPose/) is a protein superposition server which calculates protein superpositions using a modified quaternion approach. It generates sequence alignments, structure alignments, PDB coordinates, RMSD statistics, Difference Distance Plots, and interactive images of the superimposed structures (Maiti et al., 2004).

2.3.3.7 Molecular Docking

Docking is a well-established computational technique, and the goal of protein-ligand docking is to predict the possible mode of ligand binding with a protein structure. Cavities in Protein 3-D structure are predicted by algorithms such as CASTp (http://sts.bioengr.uic.edu/castp/) (Dundas et al., 2006) and it is usually in such sites that protein-ligand interactions take place. The interaction of two molecules and the best orientation of ligand forming a complex with minimum overall energy are found by calculating various forces such as electrostatic forces, electrodynamics forces, stearic forces and solvent-related forces. The Search algorithm determines all possible optimal conformations for a given complex in an environment, and it also calculates the energy of the resulting complex. The different types of algorithms that used for docking analysis are molecular dynamics, Monte Carlo methods, genetic algorithms and fragment-based

methods. The interactions are evaluated using scoring functions like:

• Empirical scoring function

Fitness = vdW + H bond + Elec

• Binding Energy ΔG bind = ΔGvdw + ΔGhbond + ΔGelect + ΔGconform + ΔG tor + ΔG sol

There are mainly two types of docking: Rigid docking and flexible docking. In rigid docking, both the internal geometry of the receptor and ligand are fixed during docking whereas the flexible docking calculates the energy for different possible conformations which make ligand more reliable to fit into the protein cavity. The results are analyzed by a statistical scoring function which converts interacting energy into numerical values called as the docking score. It also calculates the interacting energy. There are various soft wares available for molecular dockings like SCHRODINGER, DOCK, AUTODOCK, and DISCOVERY STUDIO. In this, the docking studies were done through auto dock software (Morris et al., 2009). AutoDock is an automated docking tool which is designed to predict the binding of small molecules into a receptor of known 3D structure, and all the studies were done with rigid body docking, and the results were visualized through Pymol or Chimera software.

2.4 Materials

Potato tubers were purchased 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 Corporation, Bangalore, Karnataka, India. Protein molecular weight markers were acquired from Thermo Fisher Scientific, Framingham, MA, USA. The β-amylase (200kDa), alcohol dehydrogenase (150kDa), bovine albumin (66kDa), carbonic anhydrase (29kDa), cytochrome c (12.5kDa) obtained from Sigma were used as gel filtration markers. Prepacked gel filtration column (Sephacryl- S-100 column (120 X 2cm) was from GE health science. The dialysis bag with a cutoff 10kDa was from Sigma, and the centricons of 10kDa cutoff were obtained from Millipore.

2.5 Methods

2.5.1 Isolation and purification

Protein was isolated from potato tubers as per reported method with some modifications (Matsumoto et al., 1983). Briefly, 1kg of peeled potato tubers were soaked with sufficient pre-cooled 3M acetic acid for 180 minutes, crushed and further stirred in the same buffer at 4°C for 180 minutes. The suspension was then filtered through a double layer muslin cloth at 4°C, and the filtrate collected and further 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 to Centrifuge the suspension at 10,000 x *g* at 4°C for 45 minutes. The precipitate obtained was dissolved in a minimum volume of pre-cooled 1X phosphate buffer saline buffer (pH 7.4) (PBS) and dialyzed against the same buffer using dialysis membrane, to obtain crude protein extract.

The crude protein extract was loaded onto a fetuin column (10 cm x 1 cm) pre-equilibrated with PBS and the bound protein was eluted with 0.2M acetic acid at 4°C. The collected fractions were dialyzed against PBS and fractions which showed agglutination with rabbit erythrocytes (RBC) were collected and concentrated to get 10ml of 2mg/ml total protein, and the concentrated protein was further dialyzed against 20mM Tris-HCl, (pH 7.5). The dialyzed fractions were applied to diethyl aminoethyl (DEAE) column equilibrated with 20mM Tris-HCl (pH 7.5), and fractions were eluted using increasing concentrations of sodium chloride dissolved in distilled water.

2.5.2 Molecular mass determination

2.5.2.1 Sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE)

The relative molecular mass of the purified protein was obtained by SDS-PAGE performed according to a standard method (Laemmli. U.K., 1970) on a slab of 15% resolving gel in reduced and non-reduced conditions and the presence of 0.1% SDS using a Mini-protean Tetra vertical electrophoresis cell (Bio-Rad Laboratories). More over to

check the homogeneity of the purified protein, native-PAGE was performed under acidic pH 4.3 and basic pH 8.8, using the standard methods **(**Davis B.J., 1964; Riesfeld et al., 1962). To determine whether the purified protein is a glycoprotein, purified protein was separated on a non-reducing SDS-PAGE gel, and the gel was stained with periodic Acid-Schiff instead of coomassie brilliant blue (Matthieu et al., 1973).

2.5.2.2 Gel filtration studies

Gel filtration studies were conducted on ÄKTA prime plus protein chromatography system (GE Life Sciences) using a pre-packed Sephacryl S-100 column equilibrated with PBS and calibrated with standard molecular weight marker proteins: cytochrome c (12.5 kDa), carbonic anhydrase (29.2 kDa), bovine serum albumin (67.0 kDa), alcohol dehydrogenase (150 kDa), β-amylase (200 kDa) and blue dextran (2000 kDa) obtained from Sigma. The partition constant (K) was calculated for these standards by calculating K= (Ve−Vo) / (Vt−Vo) where is Ve is elution volume of each standard; Vo is the void volume of column determined by dextran blue and Vt is the total volume of the column, respectively. The molecular mass of the protein was determined from the calibration curve obtained from a plot of the log of standard mass versus K values.

2.5.3 De novo protein sequencing

De novo protein sequencing was carried out using Mass Spectrometry facility available at Centre for Cellular and Molecular Platforms (C-CAMP), Bangalore, Karnataka, India. Sequencing was performed on the excised band from SDS-PAGE gel, and the samples were prepared according to a standard method (Shevchenko et al., 2006) and subjected for mass determination and the spectra were analyzed using PEAKS software (Ma et al., 2003).

2.5.4 MALDI-TOF analysis for protein identification

The excised band from the SDS-PAGE gel was destained and in-gel digested with trypsin. The extracted trypsinized peptides were then subjected to MALDI-TOF/MS/MS using AB SCIEX QSTAR© elite LC-MS/MS system (AB SCIEX, Framingham, MA, USA). The peptides fragment peak list generated by 4000 Series Explorer software was

searched against the MASCOT search engine (http://www.matrixscience.com) using the protein pilot software to identify them.

2.5.5 Hemagglutination and carbohydrate specificity of PotHg

The hemagglutination assay was performed according to standard procedure (Nair et al., 2012). Briefly, serially diluted 50µl of protein solution (2mg/ml in PBS), 50µl of rabbit RBC suspension (1% in PBS) and 50µl of PBS solution was added to the wells in U-shaped microtiter plate and incubated at room temperature for one hour. Visual examination (RBC sediment as button formation) of agglutination was examined, and the hemagglutination unit (H.U.), which is defined as the reciprocal of the highest dilution exhibiting visible hemagglutination was noted while the specific activity was calculated as the number of hemagglutination unit per mg of protein.

For hemagglutination inhibition assay, sugars in the concentration range of 1 to 150mM in PBS were added to the wells in place of 50µL of the buffer. The following sugars were used: D-mannose, D-galactose, lactose, N-acetyl-D-glucosamine, N-acetyl-D-galactosamine, N-acetylneuraminic acid (sialic acid), fucose, fetuin, and D-glucose. The lowest sugar concentration which inhibits visible agglutination was considered as minimal inhibitory concentration (MIC).

2.5.6 Effect of temperature and pH on hemagglutination activity of PotHg

The effect of temperature on hemagglutination activity of purified protein was determined by incubating 50µL of different protein aliquots (1mg/ml in PBS) at temperatures ranging from 30 to 100°C for 30 minutes and then cooling the tubes on ice for 30 min followed by agglutination assay to evaluate activity.

The pH stability of the purified protein was determined by dialyzing protein aliquots in buffers ranging from pH 2 to10 at 4°C for 180 minutes followed by stabilization at 4°C for 60 minutes. Further, the neutralization of pH was done by dialyzing the purified protein against PBS and checking the hemagglutination activity.

2.5.7 Protease inhibition assay of PotHg

Protease inhibition assay was performed, to check inhibition activity of the purified protein, as previously described (Smith et al., 1980). Briefly, a stock protein solution (0.1mg/ml) was diluted to different concentrations and 100µL of diluted solution together with 200µL (20 µg/mL) of bovine pancreas trypsin and 100µL of distilled water were incubated at 37°C for 10 minutes. To this mixture 500µL (0.4mg/mL) of N_{α}-Benzoyl-DL-arginine *p*-nitroanilide hydrochloride (BAPNA; pre-warmed to 37°C) was added and the reaction mixture vortexed and incubated for 10 minutes. The reaction was then terminated by addition of 100µL of 30% acetic acid (v/v) and the reaction mixture centrifuged at 10,000 x *g* for 5 minutes. The optical density of the obtained supernatant was measured at 410nm and the readings corrected for blank. Total inhibitory activity (TIA) and specific activity were calculated according to reported method (Smith et al., 1980).

2.5.8 Tryptophan modification of PotHg

Tryptophan residue modification was performed using tryptophan-specific reagent N-bromosuccinimide (NBS) by slightly modifying the method reported (Swamy et al., 1989). To the purified protein solution (1.5mg/ml) prepared in acetate buffer (pH 4.0) freshly prepared NBS (10mM) was added in increments of 10µl each and after every addition, an aliquot was removed, and the hemagglutination activity was determined after dialysis. The purified protein sample without NBS treatment served as positive control. The NBS-mediated lectin inactivation was also monitored spectrophotometrically by measuring the decrease in absorbance at 280 nm.

2.5.9 Bioinformatics analysis of PotHg

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 "Kunitz-type protease inhibitor AND *Solanum tuberosum*." 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 bootstrapping (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).

2.6 Results

2.6.1 Isolation and purification

A two stage protein purification scheme was implemented to obtain pure protein from the crude protein extract (Figure 2.1). The crude protein extract was prepared from 1kg of potato tubers in PBS and the total protein concentration of crude extract was 12mg/ml and it showed hemagglutination activity which was inhibited by complex sugar fetuin. The fractions collected from the crude extract was loaded onto the sepharose4B-fetuine column and eluted with 0.2 M acetic acid were collected separately and dialyzed against PBS (Figure 2.2). The partially purified fractions showing positive activity on further purification through DEAE ion exchange column eluted with increasing sodium chloride concentration (NaCl) showed one major peak (P2) eluted at 150mM NaCl concentration which showed a single band in SDS-PAGE and positive hemagglutination activity (Figure 2.3), and this was termed as PotHg. In a typical purification experiment, approximately 2mg of purified protein was obtained starting from 1kg of fresh potato tubers with a specific activity of 640H.U./mg. The yield and specificity activity of protein purified at each stage of purification are shown in Table 2.1.

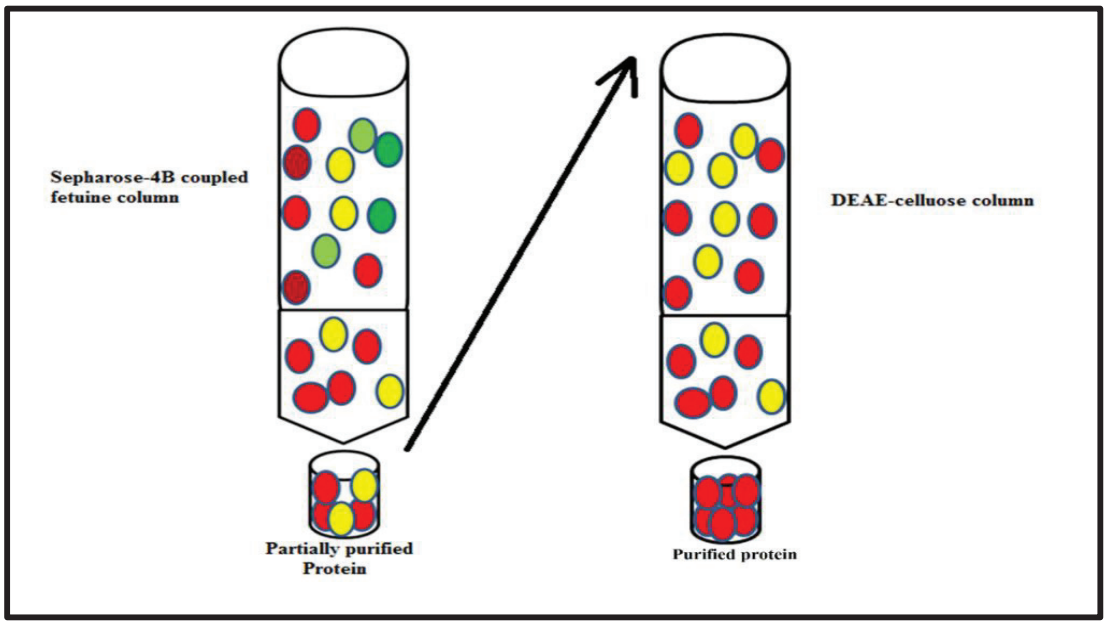

Figure 2.1 Two-step purification scheme for PotHg

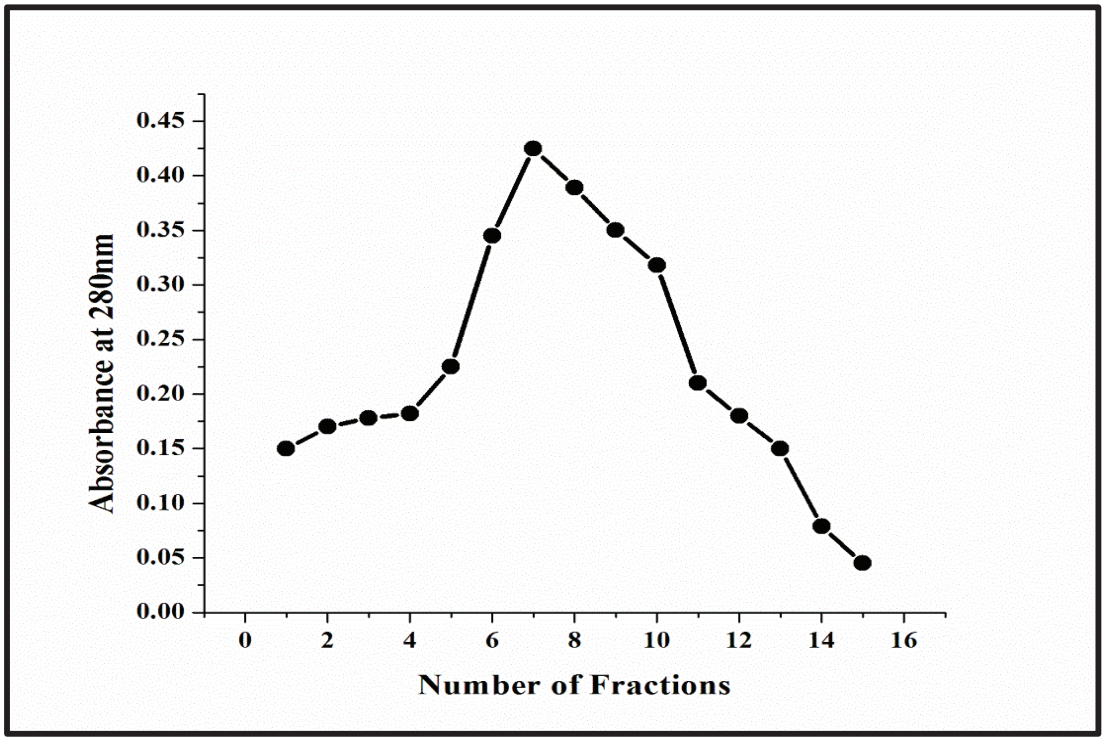

Figure 2.2 Fetuine-sepharose 4B chromatography profile X-axis demonstrate number of fractions collected (2 mL each) and Y-axis the absorbance of each fraction at 280nm

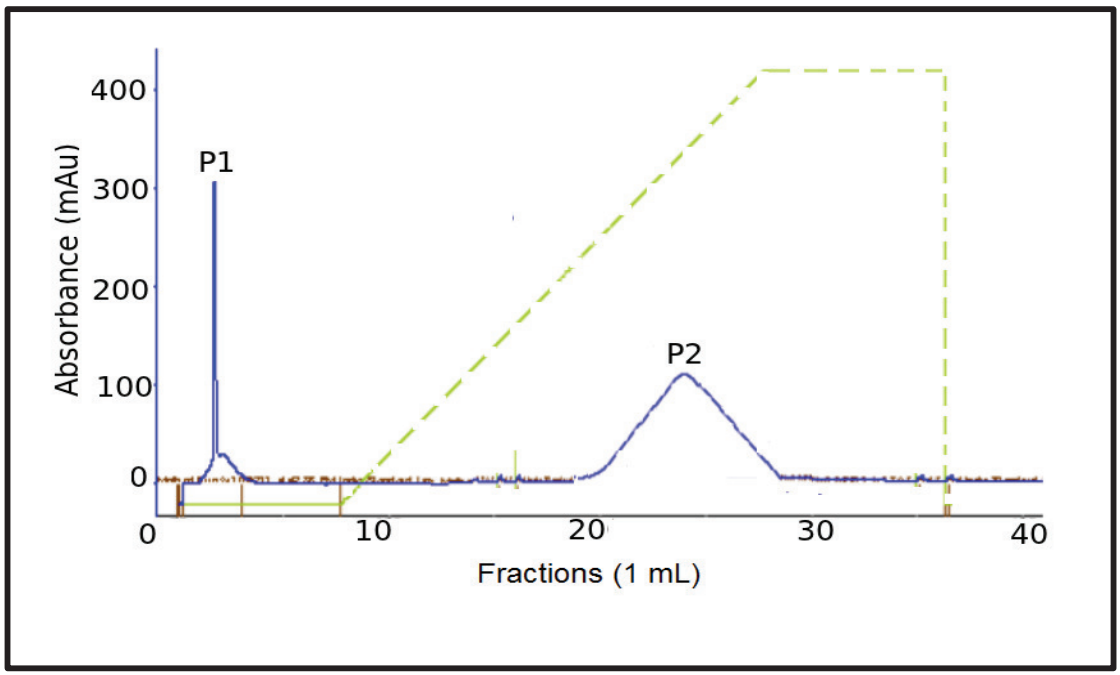

Figure 2.3 DEAE Ion-exchange chromatography profile partially purified protein from fetuin-Sepharose 4B coupled column. The pure protein was eluted at 150 mM NaCl by gradient elution, and green color indicates the concentration of NaCl in the elution buffer

Table 2.1 Purification and Specific activity of PotHg from *Solanum tuberosum*

Purification step	Volume	Total protein	Total activity H.U.	Specific activity (H.U./mg)	Purification fold
Crude Extract (From 1kg fresh tubers)	50 mL	600 mg	25600	43	1
Affinity	10 mL	20 mg	2560	128	3
Ion exchange	5 mL	2 mg	1280	640	15

2.6.2 Molecular weight determination

SDS-PAGE of PotHg under non-reduced conditions (without DTT) exhibited an electrophoretic profile of a single band with an apparent molecular mass of 20kDa but under reducing condition (with DTT), PotHg showed two bands of ~15 kDa and ~5 kDa (Figure 2.4 A, 2.4 B). This indicated that PotHg is a heterodimer composed of a large subunit of ~ 15 kDa and the small subunit of ~ 5 kDa respectively, and both the subunits could be linked through a disulfide bond (Figure 2.4 A). Gel filtration studies showed that PotHg exists as a monomer, as indicated by its K value of 1.34 which corresponded to a molecular weight of ~23 kDa (Figure 2.4 D). Glycosylation of PotHg was verified by periodic Acid-Schiff staining method and presence of staining confirmed that PotHg is a glycoprotein (Figure 2.4 C).

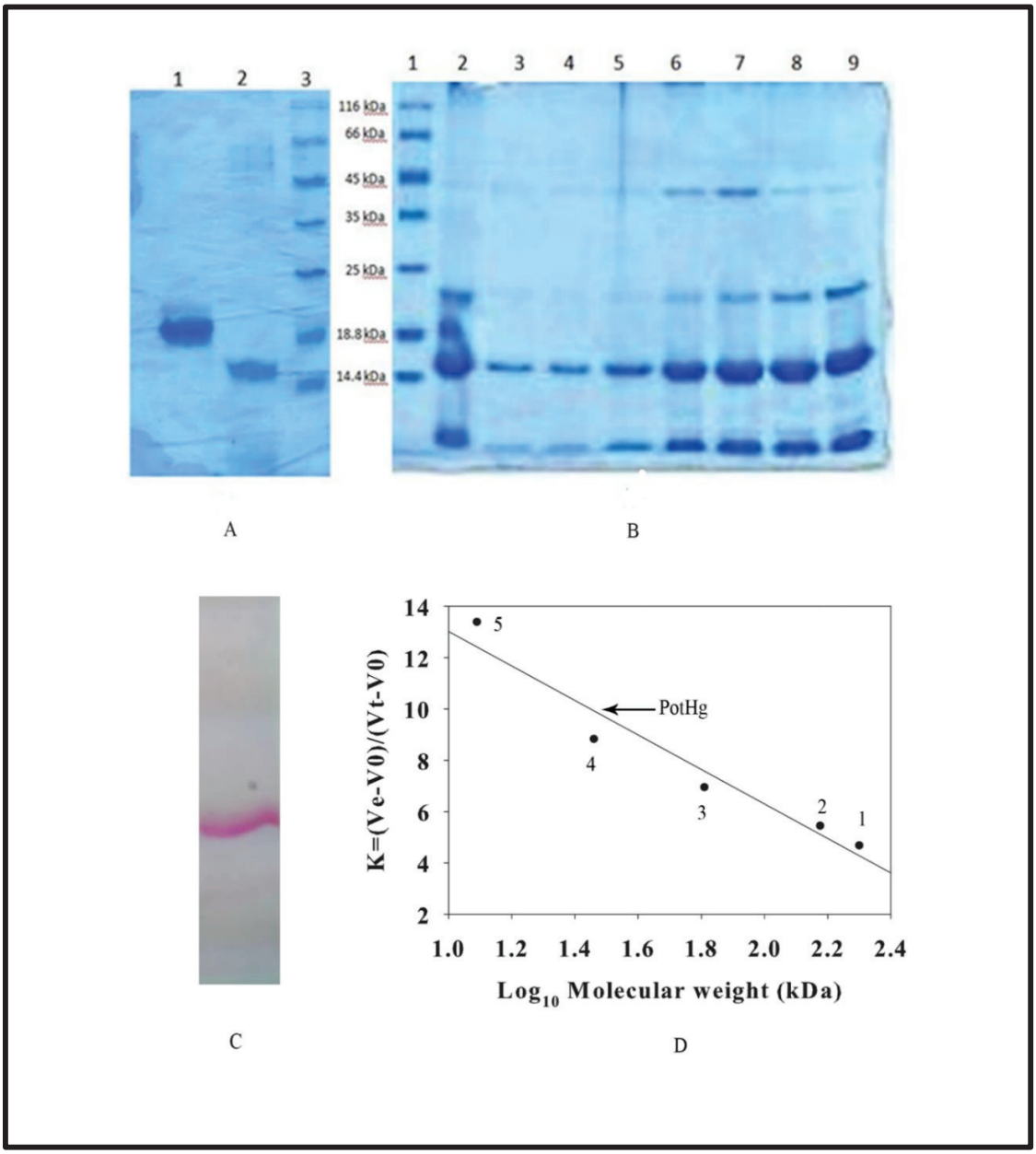

Figure 2.4 Molecular Weight analysis of PotHg (A) 1st lane shows purified protein without BME or DTT and the second lane shows purified protein with DTT. 3rd lane shows molecular weight markers. **(B)**.1st lane shows molecular weight markers. 2nd lane shows partial purification of crude protein passed through fetuin column and eluted with 0.2M acetic acid. Lanes 3 to 9 show different fractions collected from DEAE column. Fractions 3 to 5 Showed hemagglutination with 2% Rabbit RBCs. (C) PotHg stained with periodic Acid-Schiff`s reagent. **(D)** The 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

2.6.3 De novo protein sequencing results

Analysis of *de novo* sequencing of the large chain of 15.5 kDa showed the presence of 122 peptide sequences spanning 20 proteins, in SwissProt database. The identified similar protein was serine protease inhibitor from *Solanum tuberosum* (UniProt id P58515) with maximum coverage of 81% and the highest confidence score of 232.03. A total of 47 different peptides were found covering the sequence of P58515. The final sequence of both the chains is shown in Table 2.2. Mutations at three particular positions in the sequence were detected when the comparison is made between the sequence of P5815 and the sequencing result of PotHg. At position 33, 34 and 63 glycines are observed to be mutated to asparagine, but in its nearest structural homolog (PDB: 3TC2) these glycines are conserved. The mutated amino acids in bold and underlined letter are shown in Table 2. For the second chain (~5 kDa), a total of 74 peptides sequences were generated, spanning 30 proteins in SwissProt database. Protein P58515 showed 100% coverage with 6 different peptides covering the sequence. No mutation was detected in this chain of PotHg. Thus, these data indicated that the isolated protein was similar to the as Kunitz-type protease inhibitor (UniProt id: P58515), and this was further confirmed by MALDI-TOF/MS studies.

Table 2.2 *De novo* sequencing of PotHg

	Uniprot ID	**Sequence obtained from *De novo* sequencing**
Chain 1	P58515	LPSDATPVLDVTGKELDSRLSYRIISTFWGAL**NN**DVYLG KSPNSDAPCANGIFRYNSDVGPS**N**TPVRFIGSSSHFGQGI FENELLNIQFAISTSKLCVSYTIWKVGDYDASLGTMLLE TG GTIGQADSSWFKIVKSSQLGYNLLYCPVTS
Chain 2	P58515	SSDDQFCSKVGVVHQNGKRRLALVNENPLDVLFQEV

2.6.4 Hemagglutination activity and sugar specificity

PotHg agglutinated rabbit RBC with a specific activity of 640H.U/mg and the total activity, specific activity and purification fold of protein at each stage, of purification, is mentioned in Table 2.3. The hemagglutination activity was inhibited only by complex glycoproteins like fetuin, at 0.1% concentration w/v, and not by mono-, di-, and oligosaccharides examined up to a concentration of 150 mM.

Table 2.3 Inhibition of hemagglutination activity of PotHg by different carbohydrates

Sugar	Concentration	Inhibition
Mannose	50 mM-150 mM	No Inhibition
Galactose	50 mM-150 mM	No Inhibition
Lactose	50 mM-150 mM	No Inhibition
Glucose	50 mM-150 mM	No Inhibition
N-Acetyl-D-galactose amine	50 mM-150 mM	No Inhibition
N-Acetyl-D-glucose amine	50 mM-150 mM	No Inhibition
Methyl-α-D-mannopyranoside	50 mM-150 mM	No Inhibition
Methyl-α-D-glucopyaranoside	50 mM-150 mM	No Inhibition
Maltose	50 mM-150 mM	No Inhibition
Fetuin	1%	**Inhibition**

2.6.5 Effect of pH and temperature on hemagglutination activity

The hemagglutination activity of PotHg was retained between pH 3.0 and 9.0 (Figure 2.5 A) and was stable until 80°C and heating above 80°C lead to a complete loss of activity (Figure 2.5 B).

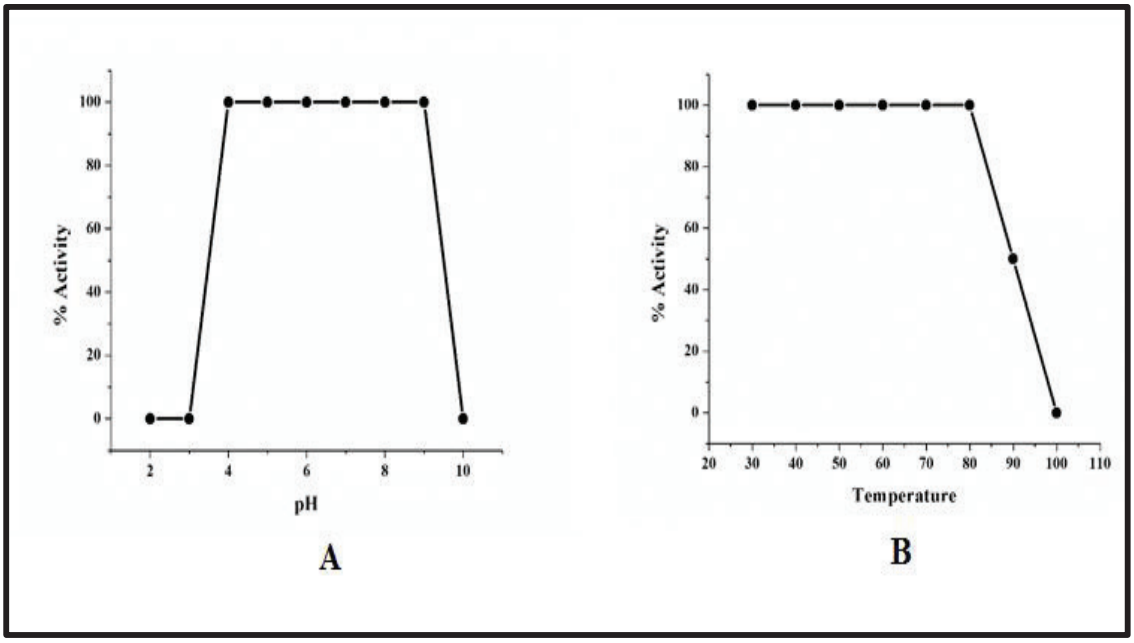

Figure 2.5 Effect of pH and temperature on hemagglutination activity of PotHg (A) Effect of pH on PotHg. (B) Effect of Temperature on PotHg

2.6.6 Protease inhibition assay

The protease inhibitor activity of PotHg was confirmed by incubating trypsin with varying concentrations of PotHg (Figure 2.6). The concentration of PotHg which inhibited the total inhibition activity (TIA) of trypsin by 50% was considered as one unit of protease inhibitor. Trypsin inhibitor activity (TIA) is defined as the number of trypsin units inhibited (TUI). PotHg inhibited trypsin with specific activity of 351 (TUI/mg) (Table 2.4).

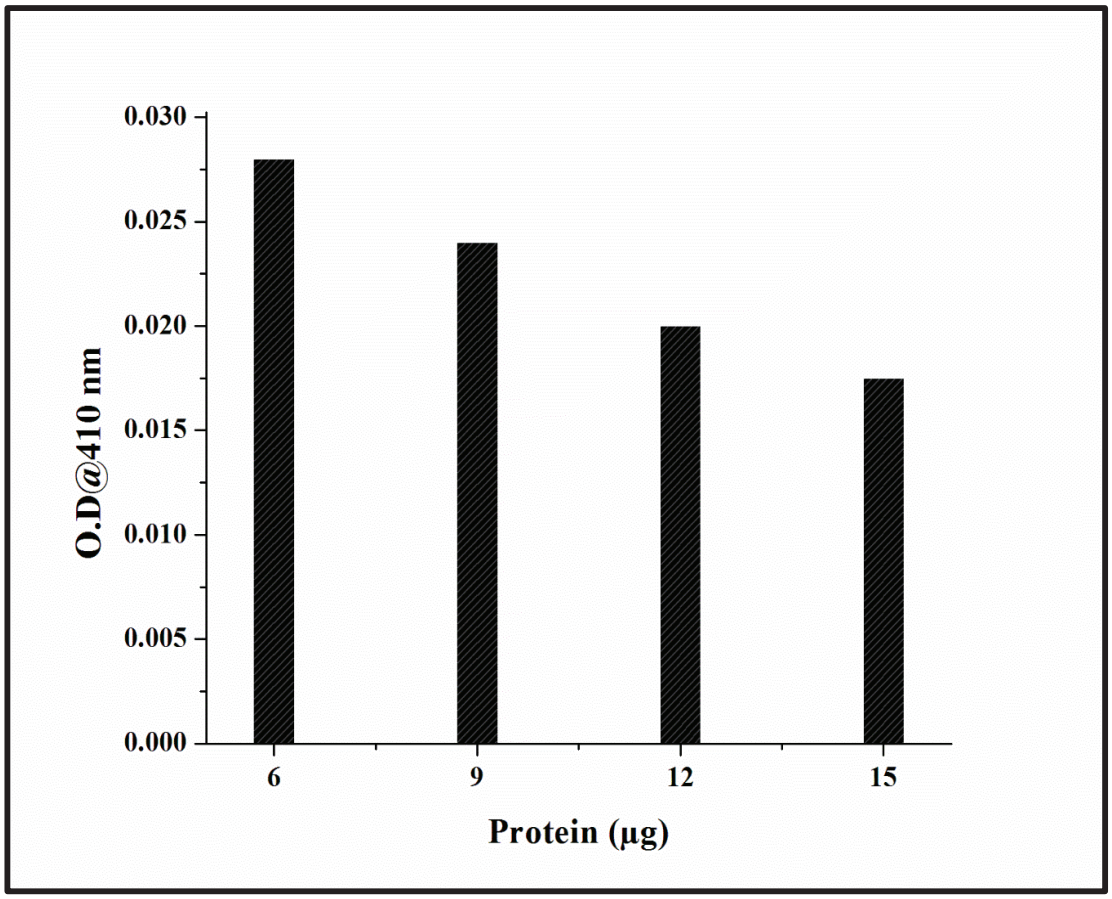

Figure 2.6 Protease inhibition assay of PotHg Serine protease inhibitor activity of PotHg with different concentrations. The concentration dependence of inhibition on trypsin is shown with different concentration of PotHg

Table 2.4 Protease inhibition activity of PotHg

Purification step	Volume	Total protein	Total activity (TIA)	Specific activity (TUI/mg)	Purification fold
Crude Extract (From 1kg fresh tubers)	50 mL	600 mg	25785	42.97	1
Affinity	10 mL	20 mg	1929	96.40	2
Ion exchange	5 mL	2 mg	710.98	350.84	8

2.6.7 Tryptophan modification by N-bromosuccinimide (NBS)

After treatment of PotHg with NBS, a reagent that specifically modified tryptophan and subsequent examination of hemagglutination activity of PotHg after removal of excess reagent, a total loss of activity was found indicating the involvement of tryptophan in carbohydrate interactions. (Figure 2.7 & Figure 2.8)

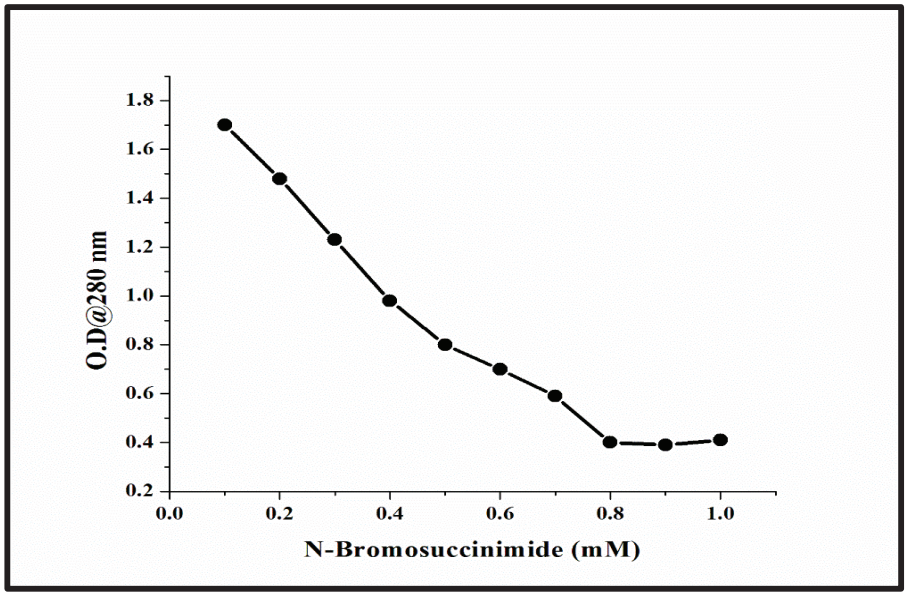

Figure 2.7 Tryptophan modification of PotHg by NBS After each addition of NBS the O.D values at 280 nm were decreased

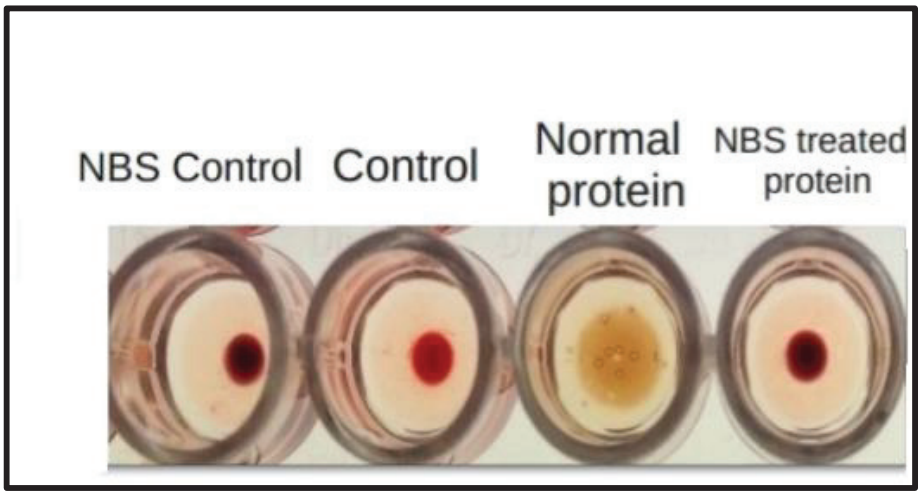

Figure 2.8 Effect of tryptophan modification on hemagglutination activity of PotHg by NBS

2.6.8 Bioinformatics analysis

2.6.8.1 Homology Search

The homologs proteins of PotHg were identified through the BLAST analysis. All the top homologs of PotHg was from *Solanum tuberosum* only, and the nearest homolog of PotHg was another Kunitz-type serine protease inhibitor with 94% identity (Table 2.5).

Table 2.5 BLAST analysis of PotHg

Accession	Description	Query Coverage	E-Value	Max. Ident.
3TC2_A	Chain A, Crystal Structure of Potato Serine Protease Inhibitor [*Solanum tuberosum*]	100%	8e-125	94%
P58514.2	Serine Protease inhibitor1; AltName: Full=PSPI21; PSPI-21-6.3 [*Solanum tuberosum*]	100%	3e-124	93%
P30941.2	Serine Protease inhibitor 7[*Solanum tuberosum*]	100%	1e-123	93%
Q41433.1	Probable Serine protease inhibitor 6 [*Solanum tuberosum*]	100%	6e-121	91%

2.6.8.2 Conserved domain

The conserved domain searches by CDD tool on PotHg sequence showed that PotHg consisted a soybean trypsin inhibitor (Kunitz) family of protease inhibitors (STI) superfamily (Figure 2.9).

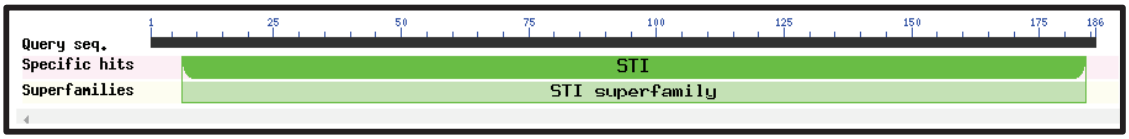

Figure 2.9 Conserved Domain Analysis of PotHg Analysis of PotHg sequence for conserved domain showed the presence of the Kunitz-type serine protease inhibitor domain

2.6.8.3 Multiple sequence alignments

A search of the UniProt database provided 18 non-redundant sequences from *Solanum tuberosum* which are listed as Kunitz-type protease inhibitors. Since Kunitz-type protease inhibitors are highly polymorphic and abundant in potato tubers homology search and subsequent multiple sequence alignment shows sequences which belong to Kunitz-type protease inhibitors only with a high degree of homology (Figure 2.10).

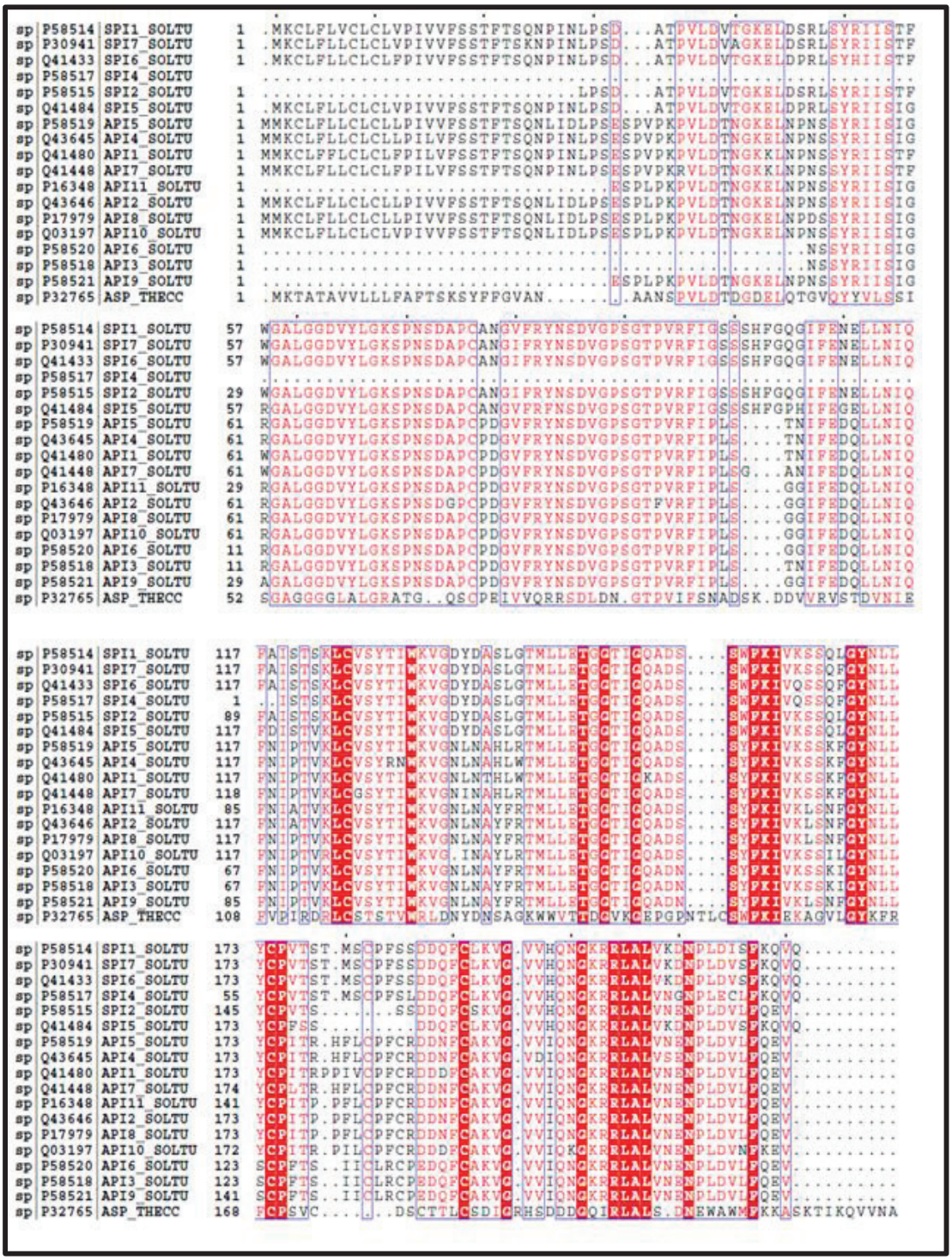

Figure 2.10 Multiple sequence alignment of PotHg Uniprot Id: P58515 (PotHg) along with other Kunitz-type protease inhibitors from *Solanum tuberosum*

2.6.8.4 Phylogenetic analysis

The phylogenetic analysis was carried out among all the Kunitz-type protease inhibitors from *Solanum tuberosum* and three clusters consisting of Aspartate protease inhibitors, serine protease inhibitors, and cysteine protease inhibitors were obtained. The PotHg amino acid sequence ID-P58515) showed a distinct clan amongst the Kunitz-type serine protease inhibitors (Figure 2.11).

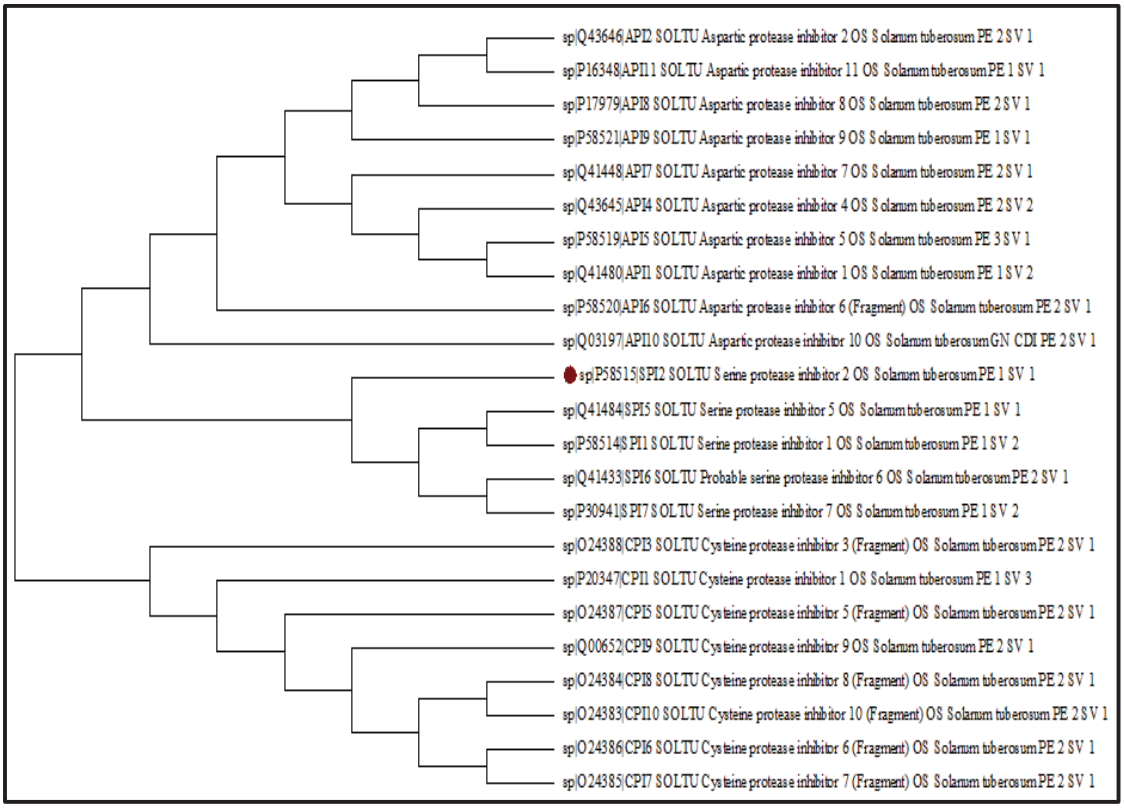

Figure 2.11 The phylogenetic analysis of Kunitz-type protease inhibitors (18 sequences of *Solanum tuberosum* Kunitz-type protease inhibitors downloaded from Uniprot database) shows three distinct clusters. Cluster 1 shows all the Aspartic protease inhibitors; Cluster II consists of Serine protease inhibitors, P58515 marked by red dot forms a distinct clade among this cluster. Cluster III contains Cysteine protease inhibitors

2.6.8.5 Molecular Modeling

In ordered to generate a reliable model for *in silico* studies, the proteins sequence of PotHg (UniProt id: P58515) was submitted to automated protein structure modeling using I-TASSER pipeline. The modeled 3D structure was compromised of 12 β-strands connected by β-turns (Figure 2.12).

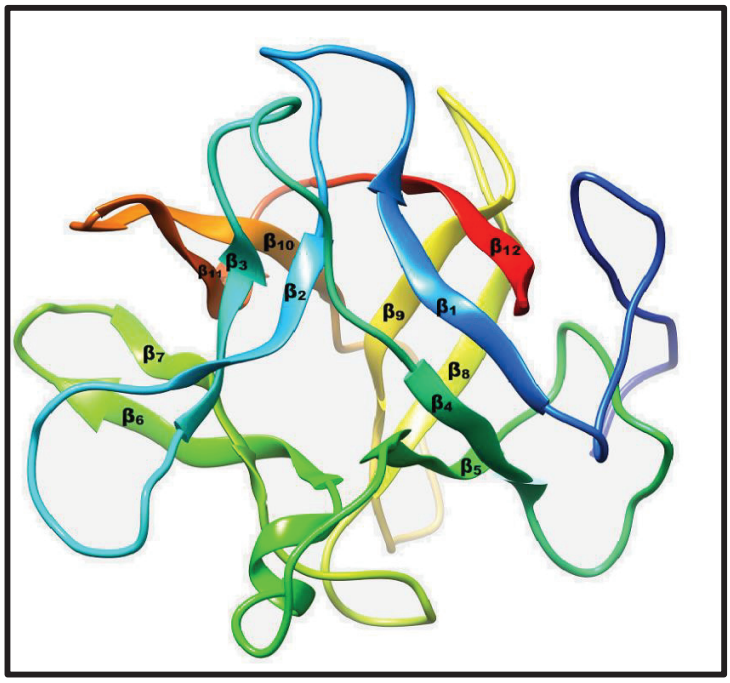

Figure 2.12 Homology model of PotHg The modeled structure of PotHg from I-TASSER server

This structural fold is very common among protease inhibitors and lectins. The refined 3D model of PotHg predicted by I-tasser showed 12 β sheets and superposed well with PDB:3IIR (serine protease inhibitor), and PDB:1WBA (winged bean albumin) with r.m.s.d. of 0.8 and 0.8, respectively (Figure 2.13). Since, this is a homology model it is not surprising to see it superimpose well with template proteins.

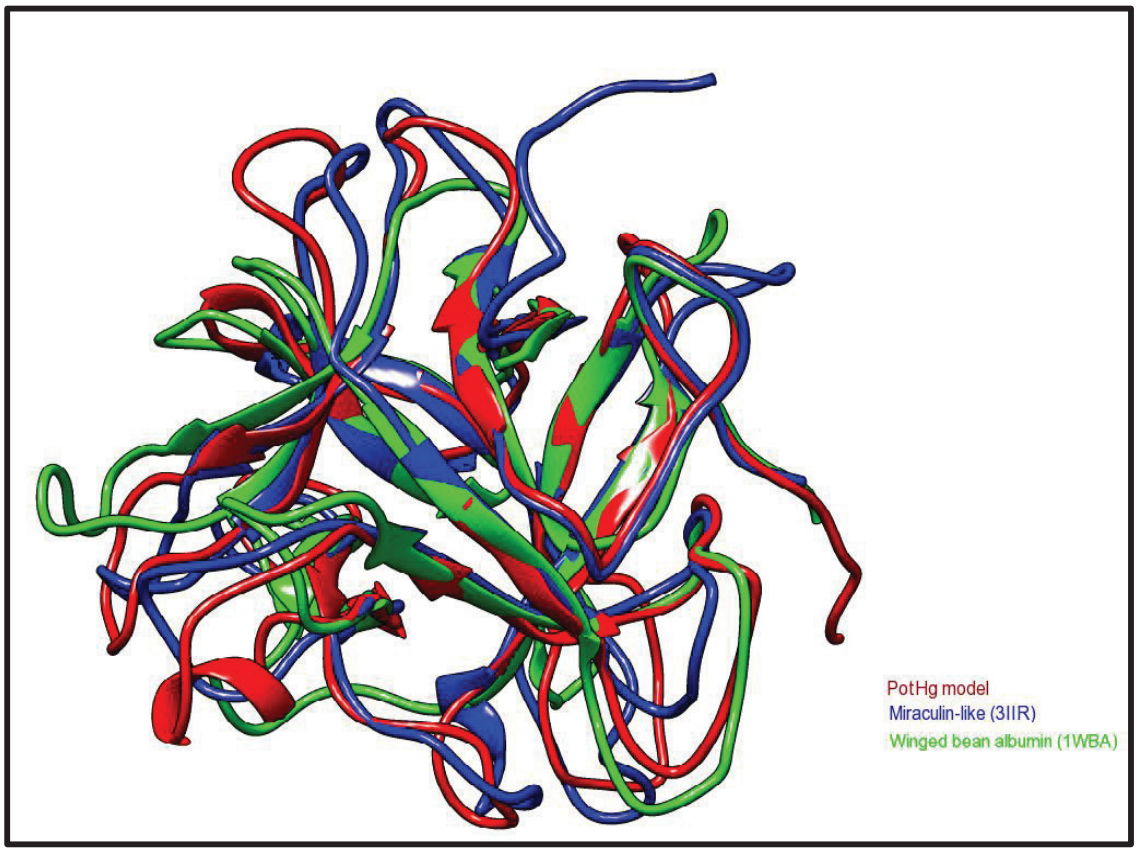

Figure 2.13 Superposition of PotHg with homolog structures Superposed structure of PotHg (Red) over Miraculin like protein (PDB code: 3IIR; Blue) and Winged bean albumin (PDB code: 1WBA; Green)

2.7 Discussion

Potato is an important food crop producing high yields of nutritionally valuable food in the form of tubers. However, it is also susceptible to a wide range of pests and pathogens resulting in low yields (Prakash, A., 2008). During the last two decades, among several plant proteins, important progress has been made in the study of the activity of plant lectins against pathogens, nematodes, and especially insects pests (Lis & Sharon, 1998) Plant lectins are suitable for various biological application including plant defense due to their structural diversity and affinity for several carbohydrates (Vijayan & Chandra, 1999). Previous studies have reported the isolation and characterization of different lectins with varied carbohydrate specificity from potato (Saitho et al., 2005; Marinkovich, 1964; Toda

et al., 1981; Leach et al., 1982; Ciopraga et al., 2000; Hasan et al., 2014) On the other hand, data pertaining to protease inhibitor exhibiting lectin activities are very few.

In accordance with our stated objectives, the present study described the isolation, identification, and characterization of a protein termed PotHg from *Solanum tuberosum.* We obtained 2mg of PotHg from 1kg of fresh potato tubers and this protein when separated on SDS-PAGE in the absence of DTT showed a single band of about 20 kDa which was resolved to two distinct bands of about 15 and 5 kDa in the presence of DTT. Gel filtration studies showed that PotHg existed as a monomer. Subsequently, *de novo* sequencing and MS/MS analysis of PotHg confirmed that the isolated protein was similar to a Kunitz-type serine protease inhibitor (UniProt id: P58515) identified by Valueva et al. (Valueva et al., 2000) who have reported that the protein is a heterodimer of a chain of 150 and 36 amino acid residues length linked by a disulphide bridge. Moreover, it has been reported that serine protease inhibitors are synthesized as a single precursor molecule which might be processed. The protease inhibitor data exhibited strong protease inhibitor activity towards trypsin. A similar finding has been reported by Walsh et al., (Walsh & Twitchell, 1991).

The lectin-like properties of PotHg was evidenced by its hemagglutination activity on rabbit RBC and hemagglutination inhibition assay, which showed that this protein had an affinity for fetuin but did not show any inhibition by simple carbohydrates like mannose and galactose. These observations were further strengthened by thermal shift assay data exhibiting a change in the Tm of protein in the presence of mannose and galactose, though at a very high sugar concentration (>100mM; data not shown). In literature, there are very few reports of protease inhibitors exhibiting lectin activity (Macedo et al., 2004; Troncoso et al., 2003; Saitoh et al., 2005). Similar protease inhibitors with lectin-like activity purified from seeds of *Peltophorum dubium and Labramia bojeri* showed specificity towards complex carbohydrates such as asialofetuin, thyroglobulin, ovalbumin and orosomucoid (Macedo et al., 2004; Troncoso et al., 2003).

Apart from biochemical studies, bioinformatics was also used to analyze the obtained data. Uniprot data for protease inhibitors from *Solanum tuberosum* tubers fetched 54 sequences. Phylogenetic analysis of 24 (non-redundant) Kunitz-type protease inhibitors showed that Kunitz-type protease inhibitors were distributed in three major clusters. These

major clusters are aspartate protease inhibitors, serine protease inhibitors, and cysteine protease inhibitors. PotHg was found as a distinct clade in the second cluster, together with other serine protease inhibitors which are reported to have a mature protein with two chains. On the contrary, proteins present in other clades contained a mature single chain protein.

Structural homology search with DALI server revealed structural similarity with different classes of proteins mainly, serine protease inhibitors, Ricin-B, Ricin-B like lectin and hemagglutinin. All these homologs possessed β-trefoil fold with carbohydrate specificity, except for serine protease inhibitors which lack lectin activity generally. Moreover, phylogenetic analysis of the sequences of the various structural homologs obtained from DALI output showed two different clades: one clade consisted of Ricin-B, Ricin-B like lectin, hemagglutinin and agglutinin and the second clade contained PotHg along with other serine protease inhibitors, miraculin-like protein, and winged bean albumin. Interestingly, winged bean albumin and miraculin-like proteins are seed storage proteins, and latter is also known to be involved in plant defense (Gahloth et al., 2011). Based on DALI output data, PotHg displayed similarity to winged bean albumin rather than with other serine protease inhibitors. Hence, it may be assumed that PotHg may also have a role in plant defense.

Thus, it may be concluded that PotHg is a unique Kunitz-type serine protease inhibitor having lectin-like activity. Further studies are necessary to elucidate the details of serine protease inhibition and regulation by this novel protein, to provide a scientific basis for better pest management strategies.

2.8 Conclusion

In this chapter, we have established the hemagglutination property of a protease inhibitor present in tubers of *Solanum tuberosum*. The lectin exists as a heterodimer in solution, held together by an interchain disulfide bond. It is thermostable, retains its activity in broad pH range, and inhibited only by glycoprotein fetuin. Purified lectin also shows strong trypsin inhibitor activity. Chemical modification of tryptophan residue led to losing of hemagglutination activity, which suggests possible direct role of tryptophan in

carbohydrate recognition and binding. The purified protein showed high sequence homology with other Kunitz-type protease inhibitors and forms a separate clade among other Kunitz-type protease inhibitors. Molecular docking studies supported the possible role of tryptophan in sugar binding. The modeled structure contains 12 β-strands which superpose well over the 3-D structure of a miraculin like protein and a winged bean albumin.