Chapter 2

Materials and Methods
In the research accounted here, several methodologies have been used for the gene identification and to study the structure-function relationship of the stress and defense proteins in chickpea. The materials and methods utilized in the entire study are enlisted below.

2.1 Sequence analysis

The protein sequences were retrieved from UniProtKB/Swiss-Prot database and aligned using multiple sequence alignment program ClustalX 2.1 (Thompson et al., 1997) to analyze the extent of conservation among the sequences. The presence and probable location of signal peptide were also predicted using SignalP 4.1 Server (Petersen et al., 2011).

2.2 Phylogenetic analysis

Prior to phylogeny building, protein sequences were aligned using MUltiple Sequence Comparison by Log- Expectation (MUSCLE) (Edgar, 2004), the program that is part of the MEGA. The Unweighted Pair Group Method with Arithmetic Mean (UPGMA) based clustering method was used to generate the alignment keeping the gap open and gap extension penalties 22.9 and 0. This alignment was used to build the phylogenetic tree using the Neighbor Joining (NJ) algorithm of MEGA with the Dayhoff substitution matrix (PAM250) and bootstrap value set to 1000.

A different methodology was utilized to analyze the evolutionary history of chickpea UGTs by using the fast likelihood-based method in order to generate the dendrograms by approximate LRT (aLRT) method (Guindon et al., 2010; Anisimova et al., 2011) in PhyML 3.0 (Guindon et al., 2010). Amino acid sequences were given as input in phylip format keeping LG (Le and Gascuel, 2008) substitution model and proportion of invariable sites and number of substitution rate categories as 0 and 4. Nearest Neighbor Interchanges (NNI) algorithm was utilized in order to improve a reasonable starting tree topology. The phylogenetic trees generated were visualized and analyzed in FigTree v1.3.1 (Rambaut, 2009).
2.3 Recent gene duplication events

The gene pairs involved in recent gene duplication events were identified by calculating the percent identity among them by preparing percent identity matrix in ClustalX and with the help of Pairwise Sequence Alignment utility at EBI (Global alignment). Blosum62 matrix was used for the alignment keeping gap open and gap extension 10 and 0.5. The end gap penalty, end gap open, and end gap extension penalty was set to false, 10, and 0.5. The sequence identity cut off between the two gene pair should be ≥ 90%.

2.4 Molecular modeling

Similarity search using Basic Local Alignment Search Tool (BLAST) algorithm (Altschul et al, 1990) was carried out against the Protein Data Bank (PDB) to select the high resolution homologous protein crystal structures (Berman et al, 2000). The sequence identity cut off was set to ≥ 30% (E-value cut off =1). Comparative modeling of target protein sequences was carried out using the Composite/Chimeric model type of Prime 3.1 [Schrödinger], using homologous template structures to analyze their structural features, binding mode and affinity with the substrates. The structures were visualized in PyMOL software [Schrödinger].

2.5 Validation of homology models and refinement

The molecular models were evaluated using Glide gscore, Emodel value, ERRAT 2.0 (Colovos et al, 1993), PDBsum (Laskowski et al, 2009), ProSA-web Protein Structure Analysis (Wiederstein et al, 2007), Verify3D Structure Evaluation Server (Luthy et al, 1992) and RMSD based on Ca overlap between target and template. GlideScore is an empirical scoring function that approximates the ligand binding free energy and affinity that consist of many terms including force field (electrostatic, van der Waals) contributions and terms rewarding or penalizing interactions known to influence ligand binding. GlideScore should be used to rank poses of different ligands, for example in virtual screening. Emodel value also uses the force field components (electrostatic and van der Waals energies) to comparing conformers, but much less so for comparing chemically-distinct species. Therefore, Glide uses Emodel to pick the best pose among all the generated poses of a ligand and then ranks these best poses.
against one another with GlideScore. The ERRAT program analyzes the environment of the atoms in the protein model. This program plots error values with respect to amino acid position in the sequence by sliding a nine residue window along the sequence. The error function is calculated based on the non-bonded atom-atom interactions in the protein structure. PDBsum is a pictorial database providing a detailed stereochemical analysis of the protein structure in terms of Ramachandran plot statistics, main chain parameters, main chain bond length and bond angles, distorted geometry etc. ProSA is a tool widely used to check errors in 3D protein models. It gives an overall quality score or Z score shown in a plot where scores estimated from experimentally determined structures in PDB are plotted. The Z-score also measures the deviation of total energy of the structure with respect to an energy distribution derived from random conformations. The Verify3D plot showed the overall compatibility of the 3-D structure with respect to the protein sequence by determining the environment of each amino acid in the 3-D model. The RMSD deviation between template and target Ca atoms was calculated by superposing structures in PyMOL. It reveals the quality of the model, lesser is the deviation better is the model.

After validation, homology models were refined by impref minimization of protein preparation wizard (Sastry et al, 2013) followed by minimization using Impact 5.8 [Schrödinger]. The model refinement phase involved preprocessing of model structures by adding hydrogens, assigning bond order, and filling missing loops and side chains. Later on, the models were subjected to restrained minimization by applying the constraint to converge the non-hydrogen atoms to an RMSD of 0.3 Å using OPLS 2005 (Jorgensen and Tirado-Rives, 1988) force field. Subsequently, the models were subjected to 500 steps of steepest descent energy minimization followed by 1000 steps of conjugate gradient energy minimization using the same force field.

2.6 Retrieval or designing of ligands

The 2-D sketcher utility of Maestro 9.3 was utilized to build the 2-dimensional structures of ligands which were then converted into 3-dimensional structures. The structural coordinates of Z-Pro-prolinal (ZPR; CID 122623) ligand was downloaded from PubChem Compound database (Bolton et al, 2008) in 3-D structure-data file
The sdf file was converted to mol2 file format in order to carry out further structural studies.

2.7 Protein and ligand preparation

Before carrying out the docking studies, water and other hetero atom groups from the protein structures were removed using protein preparation utility of Maestro. Later on the hydrogens were added to perform restrained minimization of the models. The minimization was done using impref utility of Maestro in which the heavy atoms were confined such that the strains generated upon protonation could be relieved. The RMSD of the atomic displacement for terminating the minimization was set as 0.3 Å. Similarly, ligands were refined with the help of LigPrep 2.5 [Schrödinger] to define their charged state and enumerate their stereo isomers.

2.8 Docking studies

2.8.1 Protein-ligand docking

The molecular models generated as already described were used to dock the small molecule ligands in the respective active site pocket by employing Glide 5.8 (Schrödinger) (Friesner et al., 2004). A grid was made either by taking the reference ligand or by selecting the active site residues crucial for the substrate binding. Flexible ligand docking was carried out using the standard precision option. The ligands were docked in the active site by creating a grid around the bound reference ligand or by drawing the grid around the catalytic residue. A total of 20 poses generated were scored on the basis of their glide score and E-model values. Out of these poses, the most favorable one was chosen based on glide gscore, glide Emodel value and essential interactions required for the stable substrate binding.

2.8.2 Protein-protein docking

Protein-protein docking studies were performed by using ZDOCK 3.0.2 (Pierce et al., 2014) server by defining the interface residues between the two protein chains. A total of 10 poses generated were scored on the basis of their ZDOCK score. The most favorable pose is the one which had high ZDOCK score value and stable binding mode. The hydrogen bond interactions between the protein and their cognate substrates were visualized using PyMOL.
2.9 Molecular dynamics simulation studies

2.9.1 Molecular dynamics simulation in Gromacs

The docked complexes were subjected to molecular dynamics simulations using the GROningen Machine for Chemical Simulations V4.5.4 (GROMACS) (Van Der Spoel et al., 2005; Berendsen et al., 1995) using GROMOS96 43a1 force field. The docked complexes were enclosed at the centre of the dodecahedron box solvated in water using SPC216 water model keeping 10 Å distance between the solute and the box. Topology files and other force field parameter files for the ligands were generated with the help of PRODRG2 server (Schüttelkopf et al., 2004). The system was initially energy minimized by steepest descent minimization for 50,000 steps until a tolerance of 10 kJ/mol in order to avoid the high energy interactions and steric clashes. Net charges on the docked structures were neutralized by adding equal number of counter ions to make the whole system neutral using genion program of GROMACS. After addition of ions, the system was again energy minimized by steepest descent minimization keeping identical parameters. The V-rescale, a modified Berendsen thermostat, temperature coupling (Berendsen et al., 1984) and Parrinello-Rahman pressure coupling (Martonâk et al., 2003) methods were used to keep the system stable at 300 K temperature and pressure of 1 bar. The Particle Mesh Ewald (PME) method (Darden et al., 1993) was selected to deal with long range electrostatic interactions. A distance cut off of 9 Å and 14 Å was set for Coulombic and van der Waals interactions. LINCS algorithm (Hess et al., 1997) was used to handle the rotational constraint to bonds. No positional constraints were applied on the system and periodic boundary conditions were applied in all three directions. The trajectories were visualized using Visual Molecular Dynamics program (VMD) (Humphrey et al., 1996).

2.9.2 Molecular dynamics simulation in Desmond

The docked complexes mentioned in table were prepared first using protein preparation wizard to check for any errors in the structure. Later on the processed complexes were subjected to molecular dynamics simulations using desmond 3.1 (Guo et al., 2010) of Maestro. OPLS2005 force field was applied on docked
complexes placed in the centre of the orthorhombic box solvated in SPC water model. Total negative charges on the docked structures were balanced by equal number of counter ions to make the whole system neutral (Table 2.1). The system was initially energy minimized for maximum 2000 iterations of the steepest descent (500 steps) using limited memory Broyden-Fletcher-Goldfarb-Shanno (LBFGS) algorithm with a convergence threshold of 1.0 kcal/mol/Å. The short- and long-range Coulombic interactions were taken care by Cutoff and Smooth particle mesh ewald method keeping the cut off radius of 9.0 Å and ewald tolerance of 1*10^{-9}. Periodic boundary conditions were applied in all three directions. The final run of 10 ns was applied on the relaxed system with a time step of 2.0 fs using NPT ensemble by employing a Berendsen thermostat at 300 K temperature and atmospheric pressure of 1 bar. The energies and trajectories were recorded after every 2.0 ps. The Cα atom RMSD of the complexes in each trajectory were calculated and plotted with respect to simulation time.

Table 2.1: Details of box size and number of water molecules and ions added during simulation process.

<table>
<thead>
<tr>
<th>Serial number</th>
<th>Chickpea protein Substrate/Inhibitor</th>
<th>Box dimension</th>
<th>Number of water and ions added</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>POP ZPR</td>
<td>59.69<em>82.42</em>64.98</td>
<td>19019, 18 Na⁺</td>
</tr>
<tr>
<td>2</td>
<td>Cysteine protease inhibitor Papain from Carica papaya</td>
<td>59.23<em>90.80</em>70.65</td>
<td>10614, 8 Cl⁻</td>
</tr>
<tr>
<td>3</td>
<td>Bowman-Birk inhibitor Bovine trypsin</td>
<td>112.25<em>69.25</em>70.51</td>
<td>9304, 4 Cl⁻</td>
</tr>
<tr>
<td>4</td>
<td>Potato inhibitor- I Bovine trypsin</td>
<td>58.60<em>79.52</em>71.99</td>
<td>15134, 17 Cl⁻</td>
</tr>
</tbody>
</table>

2.10 Gene identification

Draft genome of *C. arietinum* was downloaded from Legume information system (http://cicar.comparative-legumes.org/) database. Chickpea has an estimated genome size of around 740 Mb which consists of 28,269 gene models and 7,163 scaffolds covering 544.73 Mb (over 70% of estimated genome size) (Varshney *et al*, 2013). The genes involved in biotic and abiotic stress response in chickpea were identified by following different methodologies namely blastp search, hidden Markov model
profile search, and Position-Specific Weight Matrix search explained in respective chapters. Predicted proteome of chickpea was searched for the presence of stress genes by screening using HMM-profiles of Pfam 27.0 (Pfam family: PF00201.13) (Punta et al., 2012) with the help of HMMER 3.0. (Eddy, 1998; http://hmmer.org/) selecting E-value cut off of $10^{-4}$.

2.11 Detection of orthologs

Orthologs of all the identified chickpea stress proteins were searched in dicot plant genomes of Medicago truncatula, Glycine max, Vigna angularis, Medicago sativa, Vitis vinifera, Lotus japonicas, Phaseolus vulgaris, and Arabidopsis thaliana employing Blast2Go (Conesa et al., 2005) tool keeping E-value cut off 0.001 and sequence similarity $\geq 80\%$. These dicot plants were selected for analysis based on their reported chickpea homologous genomes and Blast2Go hits.

2.12 Analysis of gene structure

The gene architecture of stress genes was analyzed using the Gene Structure Display Server (Guo et al., 2007; GSDS) using the gene sequences and coding sequences. The generated output depicts the exon/intron arrangement, gene length, intron phases and position, length and position of exon and introns, and 3’ and 5’ untranslated regions (UTRs). The three intron phases were assigned as 0 for introns between two codons, 1 for those between first and second base of codon and 2 for introns inserted between second and third base.

2.13 Gene expression analysis

2.13.1 Tissue specific RNA-seq data analysis

RNA-seq raw read data or expression data was downloaded from Sequence Read Archive (SRA) database available at NCBI (http://www.ncbi.nlm.nih.gov/sra), for 5 different plant tissues namely, germinating seedling (GSM1047862), young leaves (GSM1047863), shoot apical meristem (GSM1047864) (Sam), flower bud (GSM1047865, GSM1047866, GSM1047867, GSM1047868) and flower (GSM1047869, GSM1047870, GSM1047871, GSM1047872) (Singh et al., 2013). Reads were mapped to genomic sequence of chickpea with spliced read mapper,
TopHat (Trapnell et al, 2009). Cufflinks tool was used to estimate and analyze the abundance of reads mapped to genes body and thus calculating fragment kilo base transcript per million (FPKM) value as proxy for gene expression in different plant tissues (Trapnell et al, 2010).

2.13.2 Analysis of RNA-seq data of drought stressed root tissues

RNA-seq raw read data of drought stressed root tissues from two different genotypes of chickpea, namely ICC37 and ICC506, were downloaded from SRA database available at NCBI. The RNA-seq read data are available for control and stress conditions under the following accession numbers: SRX048918, SRX048919, SRX048915, and SRX048917.

2.13.3 EST data

In addition to RNA-seq data analysis, another methodology was also followed to obtain transcriptional evidence for analyzing the expression pattern of stress genes. A blastn search was carried out using the coding sequences of each identified gene against the NCBI chickpea EST database (http://www.ncbi.nlm.nih.gov/nucest/?term=Cicer+arietinum). The number of chickpea EST till date is 46120 (from GenBank in 14-April-2013; 46120 EST sequences). The sequence identity cut off was set to > 90% to match an EST to map over a gene model keeping Expectation threshold value of 1.

2.14 Domain identification and motif analysis

The various domains present in nucleotide binding site-leucine rich repeat proteins were identified using hmmscan search in HMMER against the pfam database using gathering threshold and default E-value and Bit score parameter (http://hmmer.janelia.org/search/hmmscan). Sequence motifs in the three domains of NBS-LRR proteins were predicted with the help of MEME suite in which the minimum and maximum width of the motif was set to 15 and 50 in order to search for a maximum of 20 motifs with zoops model (Bailey et al. 2009; Bailey and Elkan 1994).
The Pfam domain and signal peptide in the identified proteases were predicted by employing a Simple Modular Architecture Research Tool (SMART) (Jörg et al., 1998; Letunic et al., 2012). The diagram of the protein structure and domain architecture was generated with the Domain Graph 2.0 (DOG) software (Jian et al., 2009).

2.15 Pseudogene analysis

The pseudogenes were identified by performing a tblastn search in NCBI using the consensus sequences of TNL and non-TNL, mentioned in the gene identification section, against the chickpea chromosome assembly with an E-value cut off of 1. A gene is identified as pseudogene if tblastn translation with respect to the consensus sequence has at least one stop codon.

2.16 Promoter analysis

The 2 kb upstream regions of the NBS resistance genes and few other classes of stress genes with major function in plant defense response were selected and then screened against the PLACE database to identify the promoter regions (Higo et al., 1998). Four cis-regulatory elements namely WBOX (TGAC(C/T)), CBF (GTCGAC), DRE (G/A)CCGAC, and GCC (GCCGCC) are known to be involved in regulation during stress condition and disease resistance response were selected for detailed analysis. They were found to be overrepresented in the promoter region of disease resistance NBS genes.

The promoter analysis of other stress proteins responsible for inducing heat and freezing tolerance, water deficit situation, salinity, wounding, oxidative stress, and resistance against biotic agents was carried (Trivedi et al., 2013). The cis-elements analyzed are listed in Table 2.2.
Table 2.2: Various *cis*-regulatory elements present in the different classes of stress genes.

<table>
<thead>
<tr>
<th>Stress protein</th>
<th>Cis elements</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chitinase</td>
<td>WBOX, GCC box, ASF1MOTIFCAMV, GT1GMSCAM4</td>
<td>Pathogenesis, Disease resistance</td>
</tr>
<tr>
<td>Glucanase</td>
<td>WBOX, GT1GMSCAM4, ASF1MOTIFCAMV AGC box, GCC BOX</td>
<td>Pathogenesis, Disease resistance</td>
</tr>
<tr>
<td>Thaumatin</td>
<td>WBOX, GT1GMSCAM4, ASF1MOTIFCAMV</td>
<td>Pathogenesis, Disease resistance</td>
</tr>
<tr>
<td>Heat shock proteins</td>
<td>HSE, CCAATBOX1</td>
<td>Heat response</td>
</tr>
<tr>
<td>LEA</td>
<td>ABRE, MYB1AT, RaV1AAT, MYBCORE</td>
<td>Water deficit, Salinity, Cold response, Drought</td>
</tr>
<tr>
<td>LTPs</td>
<td>WBOX, ASF1MOTIFCAMV, GT1GMSCAM4</td>
<td>Pathogenesis</td>
</tr>
<tr>
<td>Peroxidase</td>
<td>WBOX, G-box, WBOXXNTERF3</td>
<td>Pathogenesis, Oxidative stress</td>
</tr>
</tbody>
</table>