CHAPTER-5
DISCUSSION AND CONCLUSION
5. DISCUSSION AND CONCLUSION

The present study entitled “A Comparative Comprehensive Transcriptome Sequence Analysis For Various Stress Responsive Factors In Legumes” conducted with aim to find out stress tolerance genes/factors present in *Cajanus cajan*, *Cicer arietinum*, *Pisum sativum* and *Lens culinaris* and their possible role. Identification of genes present in these legume species is of great importance as much work on these aspect have not been done in these legume species as compared to other model crops like *Medicago truncatula*, *Glycine max* etc. *Cajanus cajan*, being kharif crop, its production is being greatly affected with rainfall and drought condition. *Cicer arietinum*, *Pisum sativum* and *Lens culinaris* are grown during rabi crop, usually without much input under unirrigated conditions. They all are consider to be one of the important pulse crop having drought tolerance capacity and also tolerant to various abiotic and biotic stresses. Legumes are constantly being exposed to biotic and abiotic stresses such as drought, high salinity, high osmolarity, threshold temperature, nutrient deficiency, oxidation and changing light conditions. These environmental stress factors negatively effect growth and productivity. Plant have evolved different mechanisms to respond such challenges. At the molecular level this involves induction of stress-responsive and stress-tolerant genes (Matsui et al., 2008). Having the ability to tolerate various type of biotic as well as abiotic stress conditions, legume contains genes/proteins/factors which help to overcome these stresses. Therefore, in the present study number of stress inducible genes/proteins present in legumes have been identified using *In-silico* analysis of ESTs database from NCBI and further analysis by different software were conducted.

Following software were applied.

- *Cajanus cajan*, *Cicer arietinum*, *Pisum sativum* and *Lens culinaris* ESTs were downloaded from NCBI database.
- EGassembler software used for the preparation of contigs from downloaded ESTs database.
- MEGA 5 was applied for predicting ka/ks values and for constructing phylogenetic tree.
- GC profile software were used for calculating GC content.
- FastAnnotator was applied for GO annotation.
- ORF finder software used for prediction of protein sequences and protein sequence were used for further analysis.
• Removal of signal peptides in the protein sequence by Signal-3L.
• Alignment was done by SIM alignment tool.
• Expasy ProtParam server was applied for prediction of physio chemical properties of protein of interest.
• SOPMA was used for calculating the secondary structure features of protein sequence.
• Prediction of domains in protein by Conserved Domain.
• Homology based modelling of protein structures prediction by Geno3D and SWISS-MODEL server.

Bioinformatics tools freely available on web as listed above were used for further analysis of a total of ~25,578 Cajanus cajan, ~46,441 Cicer arietinum, ~21,838 Pisum sativum and ~10,163 Lens culinaris M were downloaded from NCBI site http://www.ncbi.nlm.nih.gov/. ESTs contain vector sequences and in order to remove these vector sequences from these ESTs EGassembler software was used, having web address http://www.genome.jp/tools/egassembler/ (Ali Masoudi-Nejad et al., 2006). The downloaded ESTs copied on the home page of EGassembler. This software was applied for cleaning, repeat masking, vector trimming, organelle masking and assembling and generate contigs. After some time ESTs were assembled into ~1402 contigs for Cajanus cajan, ~4451 contigs for Cicer arietinum, ~3541 contigs for Pisum sativum and ~1071 contigs for Lens culinaris. Legumes homologues stress responsive genes/factors were identified by sequence similarity search BLAST (BLASTN and BLASTX) from these legume coding sequences. Thirty one genes were identified, responsible for biotic/abiotic stresses in legumes (Table 4.1a and Table 4.1b). Alcoholic Dehydrogenase have been reported in Arabidopsis thaliana (Dolferus et al., 1994). Aldehyde Dehydrogenase have been reported in Barley (Manabu et al., 1995). AP-2 / EREB P Transcription Factor have been reported in Tabacco (Sharoni et al., 2010). Actin, Dehydrin, Glycine Dehydrogenase, Glyceraldehyde-3 Phosphate Dehydrogenase, Histone H2B, Histone H3, Malate Dehydrogenase and 14-3-3 like protein have been reported in Chickpea (Pandey et al., 2008). Aquaporin, Auxin Responsive Factor, Calmodulin, Chlorophyll a/b Binding Protein and Heat Shock Protein have been reported in Arabidopsis thaliana (Tyerman et al., 2002; David et al., 2014; Shinozaki et al., 2003; Yan-Hong et al., 2011, Sun et al., 2001). DREB gene have been reported in plants (Agarwal et al., 2006). Glutamine Synthetase have been reported in Rice (Hoshida et al., 2000). Late Embryogenesis Abundant (LEA) have been reported in Mung bean
Leucine Zipper and W2 Domain Containing Protein have been reported in Chickpea (Rodriguez and Connel, 2006). Lipoxigenase have been reported in *Glycine max* (Bell and Mullet, 1991). Mitogen Activated Protein Kinase have been reported in *Salicornia brachiata* (Agarwal et al., 2014). MYB and NADH Dehydrogenase have been reported in *Arabidopsis thaliana* (Urao et al., 1993; Van Aken et al., 2009). Pathogenesis Related Protein (PR gene) and Proline Rich Protein have been reported in plants (Ebrahim et al., 2011; Ashraf and foolad, 2007). Serine/Threonine Protein Kinase have been reported in *Phaseolus vulgaris* (Hernandez et al., 2008). Transcription Factor WRKY have been reported in *Arabidopsis thaliana* and *Oryza sativa* (Pandey and Somssich, 2009). Zinc Finger Protein have been reported in Rice (Huang et al., 2009). Phospholipase and Stress Induced Protein have been reported in *Arabidopsis thaliana* (Sang et al., 2001; Gilmour et al., 1992).

Sequence distances, ks as well as ka values were calculated for all possible 31 stress responsive genes in legume under the present study surveyed. It is well known that ka is smaller than ks in natural evolution because of conservation of functional coding genes therefore non-synonymous change was less frequent in mutation of nucleotides during evolution (Hurst, 2002; Nekrutenko et al., 2002) for protein coding sequences, the synonymous rate ks is often regarded as a major of underline mutation rate (Miyata et al., 1980), through it may be influenced by other factors (Williams and Hurst, 2002). On contrary, the non-synonymous rate ka or the ratio of ka/ks (which corrects for variations in ks among proteins) is often regarded either as a major of the amount of purifying selection on the protein or as a major of the amount of positive selection; for most genes, non-synonymous rates are lower than synonymous rate and are much variable from genes to genes, this is thought to reflect differences in the extent of selective constraint and purifying selection among proteins (Graur and Li, 2000). To gain insight into the molecular and phylogenetic evolution of stress responsive genes in the four legumes species analyzed, the rate of synonymous (ks, silent mutation) and non-synonymous (ka, amino acid altering mutation) substitution, generated by MEGA 5 analysis and performed the ka/ks test for positive selection of each gene. ka/ks is a good indicator of selective pressure at the sequence level. Theoretically, a ka/ks>1 indicates the rate of evolution is higher than neutral rate. Conversely a gene with ka/ks<1 has a rate of evolution less than neutral rate (Yang and Bielawski, 2000). The estimated ka and ks for these stress responsive genes as the coalescence time for different stress responsive gene in this linage is quite similar, differences in relative rates between protein should reflect in
selection or mutations rates than time elapsed. The correlation between these ka and ks values is estimated (Fig. 4.2). The ka/ks analysis allowed the detection of genes with low ka/ks ratio, such as those in coding protein H3, H2, Lipoxygenase, Stress induced protein, Aquaporin. The majority of these proteins have been shown to be highly conserved and to suffer strong positive selection (Roth and Liberles, 2006). Analyzing the genes with highest ka/ks we identify a factor protein Glyceraldehyde-3 phosphate Dehydrogenase, Glycine dehydrogenase, AP2/ERBP and Dehydrin. These result are in accordance with previous reports which shows that genes acting in response to stress are often positively selected for diversification due to competition with the evolving effectors proteins of pathogens (Roth et al., 2006; Stukenbrock et al., 2009). This study once again underlines the importance and significance of the stress responsive genes for legumes species. Because of this reason, these genes have remained more conserved in speciation and rearrangement during the evolution of legumes species (Table 4.2). The phylogenetic tree for all 31 stress responsive factor have been generated by MEGA 5 software.

The GC content values were evaluated through GC profile (http://tubic.tju.edu.cn/ GC-Profile/) by halting parameter as default option for 1000, noting t_o≥0 (12), taking minimum length 3000bp for eukaryotic default value is 1% of the input sequences. The GC content of all the 31 stress responsive factors were 44.03 in Cajanus cajan, 41.21 in Cicer arietinum, 42.30 in Pisum sativum and 42.05 in Lens culinaris as shown in figures (Fig. 4.3, Fig. 4.4, Fig. 4.5, Fig. 4.6).

FastAnnotator (http://fastannotator.cgu.edu.tw/analysis.php) have been applied for GO annotation scheme. GO annotation results are grouped according to biological processes, cellular component and molecular functions are plotted in three separated horizontal bar chart as shown in Fig. 4.7.

Based on the result of function annotation and GO analysis, functional classification of differentially expressed stress responsive genes, prevalent protein families have been classified into seven categories.

- Genes involved in regulatory pathway (Cell Signaling and Transcription) include Calmodulin, Serine/threonine- Protein Kinase, 14-3-3 like protein, MAP Kinase, Actin.
- Transcription Factors (TF) acting in Stress includes MYB, AP2/EREBP, Leucine Zipper and W2 domain containing protein, Zinc finger and WRKY.
- Gene involved in metabolism include Alcohol Dehydrogenase, GAPDH, Malate dehydrogenase, NADH, Aldehyde dehydrogenase, Phosphopase, Glutamine synthase, Proline rich protein, Glycine dehydrogenase.
- Genes related to Heat Stress include Heat Shock Protein, Stress Induced Protein
- Phytohormone include Auxin Responsive.
- Defense and Stress responsive gene include PR Gene, Aquaporin, H2B, Lipoxygenase.
- Genes related to abiotic stress and detoxification include Late Embroyogenesis Abundant (LEA), Dehydrin, DREB, Histone H3, Chlorophyll a/b binding protein.

Prediction of protein sequence was done by ORF finder [http://www.ncbi.nlm.nih.gov/gorf/gorf.html](http://www.ncbi.nlm.nih.gov/gorf/gorf.html). ORF finder made the process much faster and made it easier to identify the proper reading frame. On the criteria of comparative coding region and its length, in the present study protein sequence of two protein/gene of interest Glyceraldehyde-3 phosphate Dehydrogenase (GAPDH), 14-3-3 like protein have been selected for homology modelling. In GAPDH homology modelling was done by Geno3D and SWISS-MODEL where as in 14-3-3 it was performed by SWISS-MODEL server. The GAPDH genes that encodes an enzyme in the glycolytic pathway (Russell and Sachs, 1989; Yang et al., 1993). 14-3-3 proteins are a family of conserved regulatory molecules that are expressed in all eukaryotic cells. 14-3-3 proteins have the ability to bind a multitude of functionally diverse signaling proteins, including kinases, phosphatases, and transmembrane receptors. In BLASTN sequence analysis identified, the GAPDH in Cajanus cajan (CcContig655), Cicer arietinum (CaContig665), Pisum sativum (PsContig750) and Lens culinaris (LcContig465) and identified the 14-3-3 in Cajanus cajan (CcContig409), Cicer arietinum (CaContig193), Pisum sativum (PsContig2284) and Lens culinaris (LcContig 964). For protein structure prediction these sequences were converted into amino acid sequences by applying ORF finder and Fasta converter as have been shown in (Table 4.3 and Table 4.7).

The physicochemical properties of GAPDH protein in Cajanus cajan, Cicer arietinum, Pisum sativum and Lens culinaris in context of molecular weight, isoelectric point (pI), number of positively and negatively charged residues (+R and –R), extinction coefficient (EC), instability index (II), aliphatic index (AI) and grand average of hydropathy (GRAVY). EXPASY’S ProtParam server (Gasteiger et al., 2005) [http://us.expasy.org/tools/protparam.html](http://us.expasy.org/tools/protparam.html)
was applied and results are shown in Table 4.4. The generated isoelectric point (pI) was found to be less than 7 except in *Lens culinaris* indicating acidic nature of the protein this can be helpful for developing buffer system for purification by isoelectric focusing. The AI values for all GAPDH protein in all legumes were found to be stable for a wide temperature range. Low gravy indices of GAPDH under study indicates the possibility of better interaction in aqua media. The EC indicates the how much light a protein absorb at a certain wavelength. The instability index provide an estimate of the stability of protein in a test tube. Similar pattern result have been observed for 14-3-3 as shown in Table 4.8.

SOPMA ([Geourjon and Deleage, 1995](http://www.ncbi.nlm.nih.gov)) was used for calculating the secondary structure features of protein sequences as have been shown in Table 4.5. The secondary structure elements for protein in *Cajanus cajan* and *Lens culinaris*, the α-helix were found to be more dominate than in *Cicer arietinum* and *Pisum sativum*. The random coil were found to be dominate in secondary structure among all four legume followed by extended strand were in this GAPDH study. In 14-3-3, the predicted secondary structure α-helix were found to be more dominate in *Cajanus cajan*, *Cicer arietinum* and *Lens culinaris* than in *Pisum sativum*. Whereas, the random coil were found to be more dominate in *Pisum sativum* among all four legume as have been shown in Table 4.9.

The protein sequences were further taken for BLASTP ([http://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE =Proteins](http://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins)) with the PDB to find suitable template protein sequence. For GAPDH protein template selected was PDB 3E5R_O (*Oryza sativa*) having identity 85% with *Cajanus cajan*, 84% with *Cicer arietinum*, 82% with *Pisum sativum* and 86% with *Lens culinaris*. The template selected for 14-3-3 like protein was PDB 3AXY_C (*Oryza sativa*) having identity 86% with *C.cajan*, 75% with *C.arietinum*, 78% with *P.sativum* and 84% with *L.culinaris*.

In GAPDH homology modelling was through Geno3D and SWISS-MODEL servers as have been shown in Table 4.6. Geno3D and SWISS-MODEL shows 74.4% and 84.4% of the total residue falls in the most favoured region respectively for *Cajanus cajan* GAPDH (Target_ *Cajanus cajan*) protein, whereas, 1.1% and 0.0% of the residue falls in the disallowed regions.. The SWISS-MODEL was found to be the best for the residues, as the percentage contribution of the residues in the generously allowed and disallowed region were 0.0% and 0.0% for *Cajanus cajan* protein. Similar pattern for the most favoured region was also observed in all other three
legumes under study. In 14-3-3 homology modelling was through SWISS-MODEL servers as have been shown in Table 4.10. For *Cajanus cajan* 14-3-3 protein SWISS-MODEL shows 100% of the total residue falls in the most favoured region where as data for *Cicer arietinum, Pisum sativum* and *Lens culinaris* are 98.7%, 98.5% and 98.3% respectively.

Structure analysis and verification server (SAVS) was employed to keep a check on model quality. The percentage of the residues in the “core” region is the most suitable way to check the stereo chemical quality. 84.4% of residues were in the core region of *Cajanus cajan*, 84.4% of residues were in the core region of *Cicer arietinum*, 83.3% of residues were in the core region of *Pisum sativum*, 83.2% of residues were in the core region of *Lens culinaris*. The models found was the best fitting both the parameters required for quality check i.e having maximum core region covers and less disallowed region with minimum energy (Fig. 4.19). The average Z score, for GAPDH in all four legumes was -1.15 *Cajanus cajan*, -0.1 *Cicer arietinum*, -0.38 *Pisum sativum* and 0.08 *Lens culinaris* as computed in PROVE of Structural Analysis and Verification Server (SAVES). The models generated for 14-3-3 protein were PASS through SAVS with 87.21% *Cajanus cajan* (Fig. 4.25), 80% *Cicer arietinum* (Fig. 4.26), 100% *Pisum sativum* (Fig. 4.27), 83.695 *Lens culinaris* (Fig. 4.28). The average Z score, for 14-3-3 in all four legumes are 0.315 *Cajanus cajan*, -1.506 *Cicer arietinum*, -1.341 *Pisum sativum* and -1.673 *Lens culinaris* as computed in PROVE of SAVS.

Based on the above results, it can be helpful in extracting information about gene content features, transcriptome changes and novel stress responsive genes both abiotic and biotic factors. The results concerning the prevalence of protein related to various stress related genes in all four crop under study. Identifying and mining genes involved in stress response represent a key step to unraveling and manipulating stress tolerance in legumes. Comparative analysis among the legumes within the same species and between species will enable us to identify species specific genes underlying stress response. Despite knowing that comparisons between these legumes species data should be carefully inspected, our initiative established possible transcriptome elements that could guide the legumes specific community in unraveling the molecular mechanism that distinguish these four extremely important legumes species. In addition, the annotation of legumes-specific/stress prominent genes adds new element to genomic initiatives that our searching for traits (factors) that could differentiate legume each species from other.
Comparative homology model generated by SWISS-MODEL was more accurate, precise and acceptable than Geno3D with more than 80% of the most favored region in all four legume species. The homology modelling may provide a good base for functional analysis. The protein structure of GAPDH and 14-3-3 in *Cajan cajan*, *Cicer arietinum*, *Pisum sativum* and *Lens culinaris* obtained through *in silico* based approach were as good as their protein structure obtained by X-ray crystallography or NMR. This study may help in understanding the protein function, number, taxonomic studies, protein–protein interactions, predicting immunogenic portion and evaluation studies.

The data are a valuable aid to the interpretation of legume development, providing insight that could help in legumes reading program and indicating potential targets for functional analysis and biotechnology products of such socially and economically important legume species.