BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 8(24)

Identification and annotation of abiotic stress responsive candidate genes in peanut ESTs

Archana Kumari, Ashutosh Kumar, Aakanksha Wany, Gopal Kumar Prajapati & Dev Mani Pandey*

Department of Biotechnology, Birla Institute of Technology, Mesra, Ranchi, Jharkhand-835215, India; Dev Mani Pandey – Email: dmpandey@bitmesra.ac.in; Phone: +91 651 2276223; Fax: +91 651 2276052; *Corresponding author

Received October 30, 2012; Accepted November 11, 2012; Published December 08, 2012

Abstract:
Peanut (*Arachis hypogaea* L.) ranks fifth among the world oil crops and is widely grown in India and neighbouring countries. Due to its large and unknown genome size, studies on genomics and genetic modification of peanut are still scanty as compared to other model crops like Arabidopsis, rice, cotton and soybean. Because of its favourable cultivation in semi-arid regions, study on abiotic stress responsive genes and its regulation in peanut is very much important. Therefore, we aim to identify and annotate the abiotic stress responsive candidate genes in peanut ESTs. Expression data of drought stress responsive corresponding genes and EST sequences were screened from dot blot experiments shown as heat maps and supplementary tables, respectively as reported by Govind *et al.* (2009). Some of the screened genes having no information about their ESTs in above mentioned supplementary tables were retrieved from NCBI. A phylogenetic analysis was performed to find a group of utmost similar ESTs for each selected gene. Individual EST of the said group were further searched in peanut ESTs (1,78,490 whole EST sequences) using stand alone BLAST. For the prediction as well as annotation of abiotic stress responsive selected genes, various tools (like Vec-Screen, Repeat Masker, EST-Trimmer, DNA Baser, WISE2 and I-TASSER) were used. Here we report the predicted result of Contigs, domain as well as 3D structure for HSP 17.3KDa protein, DnaJ protein and Type 2 Metallothionein protein.

Keywords: *Arachis hypogaea*, EST, Gene annotation, Stress, Contigs.

Background:
Peanut (*Arachis hypogaea* L.) is an important source of protein and vegetable oil. Most of the species among the genus Arachis are wild and are not domesticated. Cultivated peanut (*A. hypogea* L.) is an allotetraploid (2n = 4x = 40) species which is widely cultivated around the world in tropical, sub-tropical and warm temperate climates **[1]**. Thus, peanut is an internationally important crop for human consumption and oil production. Abiotic stresses such as drought, cold, salinity and sodacity are major factors affecting the crop yield. There are many genes or gene networks are involved in response to abiotic stresses. Some of the major challenges for the breeding programs are developing abiotic stress tolerant, improved oil quality and flavoured varieties of peanut. As peanut can be cultivated in arid and semi-arid regions and drought or dehydration is most frequent stress reduced the yield. Therefore, study on drought stress responsive genes and its regulation is very much important. Use of molecular markers and marker- assisted selection for crop improvement and evolution showed little variation which resulted in limited application. Therefore, the development of transgenic peanut has enormous potential for enhanced productivity, trait improvement and identification of agronomically useful genes.

Expressed sequence tags (ESTs) are considered as most appropriate tool for discovering a gene, mapping a gene and analyzing various traits necessary for crop improvement. They are generated from large scale sequencing of clones from cDNA libraries prepared from mRNA isolated at a specific stage of cell/tissue development. Many EST libraries have been constructed and are available in sequence databases. As the functions and regulation of most of the newly identified genes are not known, but can be determined by multiple sequence alignment with known genes/ ESTs available in the databases.

EST approach developed during the establishment of genomics era and in HGP (Human Genome Project) has proved to be the most high throughput technologies ever invented **[2]**. Use of EST maps in major crop genome projects such as *Zea mays* **[3]**, *Glycine max* **[4]** and *Triticum aestivum* **[5]** have been successfully applied. Study on identification and functional validation of some drought induced genes differentially expressed during gradual water stress in peanut has been conducted **[6].**

In the present study, we aim to identify the drought stress responsive candidate genes in peanut ESTs available in databases. Experimental data of expression analysis of drought stress responsive genes and its corresponding EST sequences as heat maps and supplementary tables reported previously **[6]** were screened. Furthermore, a phylogenetic analysis was performed to find a cluster of utmost similar ESTs for each selected gene. Individual EST of the mentioned cluster were further searched in peanut ESTs (1, 78,490) using stand alone BLAST. To predict as well as annotate abiotic stress responsive selected genes various tools like Vecscreen (a system that is developed by NCBI for quickly identifying and removing segments of a nucleic acid sequence that may be of vector origin and act as contaminant, before sequence analysis), RepeatMasker (a program that screens DNA sequences for interspersed repeats and low complexity DNA sequences), EST trimmer (a tool used for trimming EST sequences concerning ambiguous sequences, removing of distal oligoN series from either the 5' or 3' end and size cut off), DNA Baser (that allows to assemble a set of contiguous sequences, contigs) and WISE2 (that compares a protein sequence to a genomic DNA sequence, allowing for introns and frameshifting errors) were used. Here we report six best putative candidate genes belonging to three different gene families for drought stress response in peanut. Further, Predicted result of contigs, domain as well as 3D structure for HSP 17.3KDa protein, DnaJ protein and Type 2 Metallothionein protein were also discussed.

Methodology:

Screening of Peanut ESTs

Experimental data for expression analysis of drought stress responsive genes shown in the heat map and supplementary table **[6]** were downloaded. Depending upon the higher expression during drought five gene families was selected. Total twenty two ESTs were screened from the above selected five different gene families. With the efforts of Peanut Genome Initiative (PGI), out of 2, 52,832 ESTs reported from the genus *Arachis* at NCBI, 1, 78,490 ESTs were belongs to cultivated peanut *A. hypogaea*. Total ESTs sequences of *A. hypogaea* and above twenty two screened ESTs sequences were retrieved from the NCBI database.

Phylogenetic Analysis

To find relationship between ESTs of individual gene family with representative genes, phylogenetic analysis was performed using MEGA 5.1 Beta (a unified tool for showing automatic and manual sequence alignment, concluding phylogenetic trees, estimating rates of molecular evolution, mining web-based databases, inferring inherited sequences, and testing evolutionary hypotheses) software available at http://www.megasoftware.net. It is used for phylogenetic analysis in various research fields **[7]**.

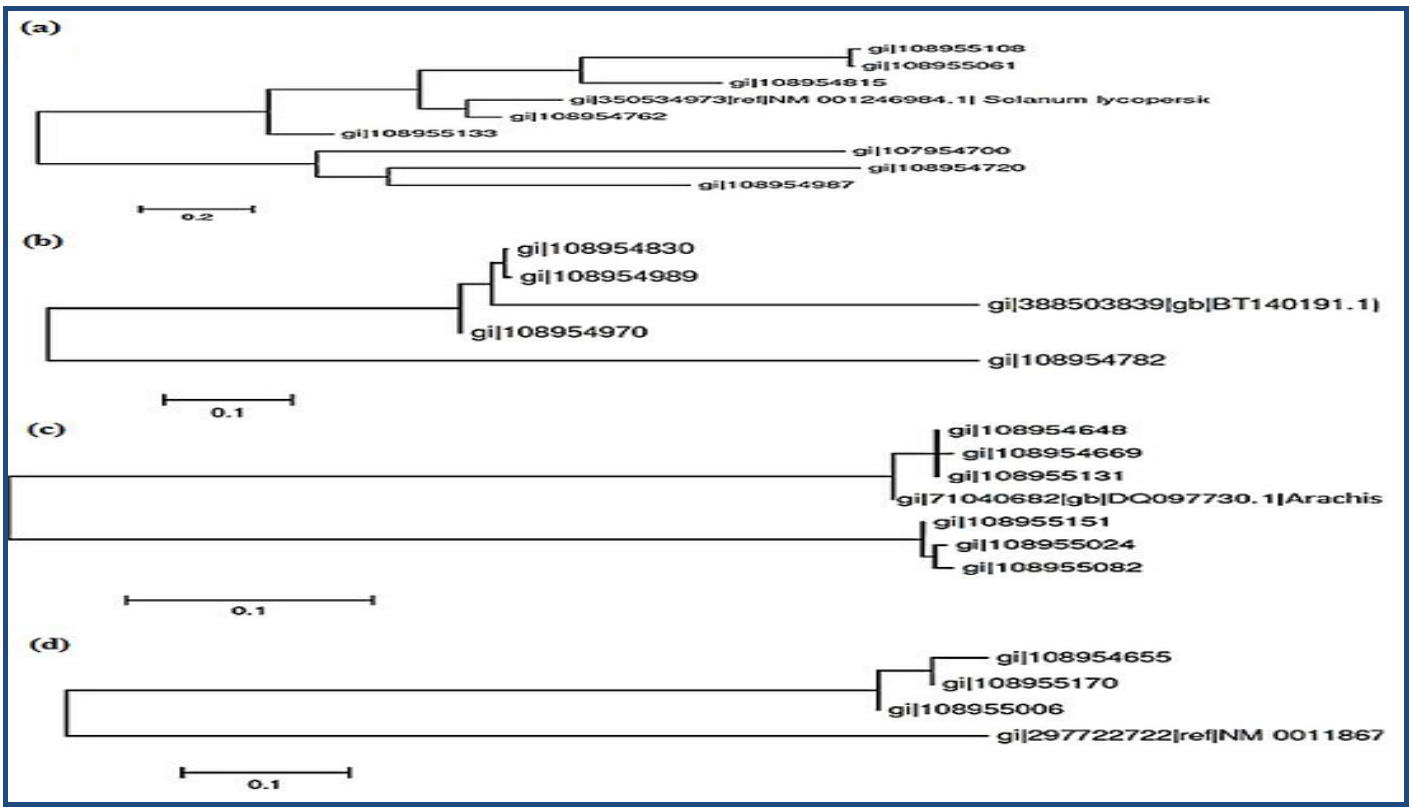

Figure 1: (a) Dendrogram depicting phylogenetic relationship between screened ESTs of 17.3 KDa class I HSP and representative gene of *Solanum lycopersicon* (gi|350534973|ref|NM 001246984.1). The evolutionary relationship is analysed by using MEGA 5.1. The phylogenetic tree represented the nearest gene with respect to reference gene is gi|108954762; **(b)** Dendrogram depicting phylogenetic relationship between screened ESTs of desiccation protective protein and representative gene of *Lotus japonicu*

(gi|388503839|gb|BT140191.1). The phylogenetic relationship is analysed by using MEGA 5.1. The phylogenetic tree represented the nearest gene with respect to reference gene is gi|108954989; **(c)** Dendrogram depicting phylogenetic relationship between screened ESTs of type 2 metallothionein and representative gene of *A. hypogaea* (gi|71040682|gb|DQ097730.1). The evolutionary relationship is analysed by using MEGA 5.1. The phylogenetic tree represented the nearest gene with respect to reference gene is gi|108955131; **(d)** Dendrogram depicting phylogenetic relationship between screened ESTs Hsp70-60 KDa chaperonin and representative gene of *Oryza sativa* (gi|297722722|ref|NM 0011867). The phylogenetic relationship is analysed by using MEGA 5.1. The phylogenetic tree represented the nearest gene with respect to reference gene is gi|108955006.

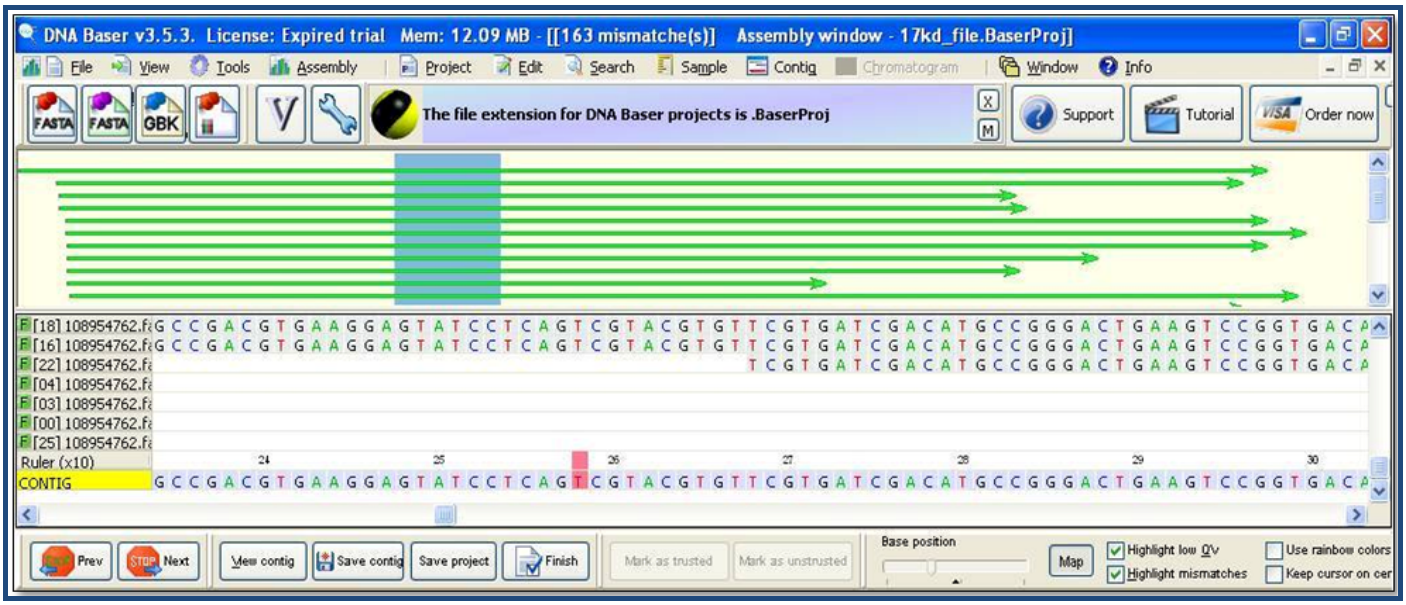

Figure 2: Result showing snapshot of assembled contig map of a gene family analyzed by DNA Baser (v3.5.3). The green arrows represent the aligned EST sequences which were used for EST assembly. The pink color indicated the mismatch in final contig sequence. The side bar indicating the list of input sequences with file extension .fasta ("GI number".fasta).

Standalone BLAST

The standalone BLAST wasperformed against whole peanut ESTs and using screened ESTs as query sequences. BLAST analysis was done on the basis of their highest bit score and lowest e-value.

Data Normalization

After BLAST analysis, the shortlisted ESTs were further checked for the presence of vector sequences, polyA tail, interspersed repeats and low complexity regions in the DNA followed by their removal by using different online available software's like VecScreen [available at www.ncbi.nlm.nih.gov/VecScreen/VecScreen.html and was used for the identification of a nucleic acid sequence that may contain a vector origin. A hash search algorithm was used to remove contaminating sequences **[8].** The algorithm initially uses a hash lookup, followed by a gapless hit extension. Finally, ESTs that passed the filters but possessed an unmasked sequence were discarded], EST trimmer (available at http://imed.med.ucm.es/EMBOSS/runs/tmp.0.xuYbpD/index.html) and can recognize one or more nucleotide sequences with any 3' poly-A tail (or, optionally, 5' poly-T tail) in the input sequences that are at least the specified minimum length **[8]** and Repeat Masker available at www.repeatmasker.org/and is a program that curtains DNA sequences for interspersed repeats and low complexity DNA sequences. The output of the database is a detailed explanation of the repeats that are contemporary in the query sequence as well as a modified version of the query sequence in which all the annotated repeats have been masked. In various contig assemblies, repeat masker has been used **[9]**.

EST Assembly

EST assembly was performed using DNABaser (www.dnabaser.com/) software. Data filtering and assembly of EST sequences were accomplished with the help of DNABaser (v3.5.3), which incorporates quality filtering, clustering, and assembly into one distinct pipeline. The filtering step includes masking of vector sequence and low-quality regions and annotation of low-complexity regions, repeats, and poly (A) regions. The final assembled ESTs were named as contigs.

Identification of putative candidate genes

The contigs of individual gene families were subjected to BLAST analysis with the whole genome of *Arabidopsis* available in TAIR10 (www.arabidopsis.org/) database. The putative candidate gene was identified by WISE2 (www.ebi.ac.uk/Tools/Wise2/) analysis on the basis of lowest e-value and highest WISE2 score (in bits). Wise2 help to perform the comparison of a protein sequence to a genomic DNA sequence which allows understanding of the errors in intron and frame shifting using Gene Wise algorithm **[10]**. It helped to select the putative candidate genes with known functions.

Three dimensional structure prediction of identified domains

3D structure of the identified domains of the protein sequences were predicted using I-TASSER (http://zhanglab.ccmb.med.umich.edu/I-TASSER/). I-TASSER

(The iterative threading assembly refinement) server determines 3D structures of protein based on the sequence-to-structure-to-function paradigm algorithm. It predicts secondary structure, tertiary structure and functional annotations on ligand-binding sites, enzyme commission numbers and gene ontology terms. The accuracy of prediction is based on the confidence score of the modelling **[11, 12]**.

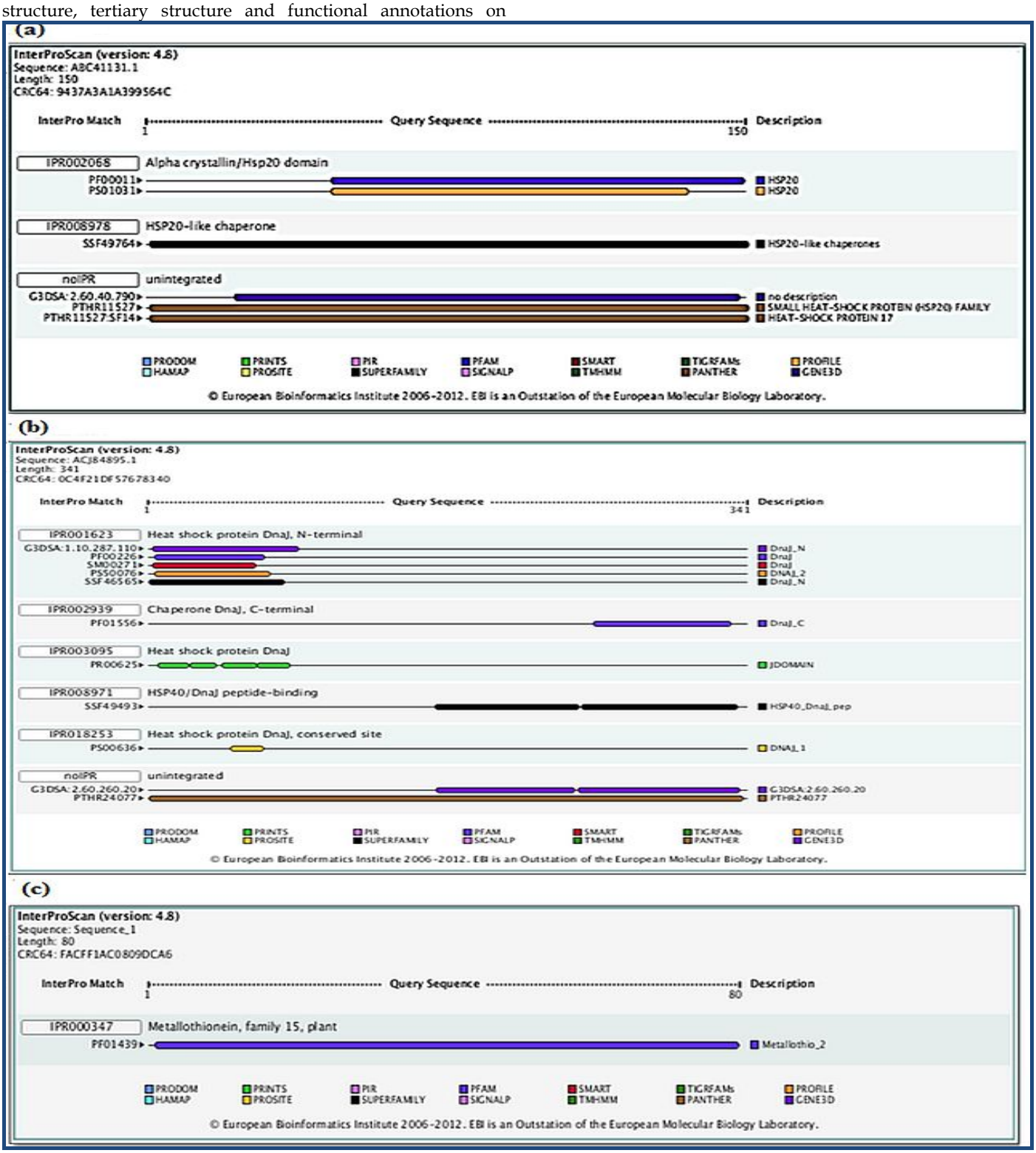

Figure 3: (a): Result of HSP 17.3KDa protein for domain prediction/identification by InterProScan. The blue colour bar is showing the Alpha crystalline/HSP20 domain having interpro ID PF00011 (IPR002068). The sequences of HSP 17.3KDa protein contain the highly similar domain with Alpha crystalline/HSP20 domain. The black bar is representing the HSP20 like chaperone (i.e. IPR008978 from SUPERFAMILY). These sequences also contain the unintegrated domain (i.e. noIPR). In unintegrated domain the blue bar is from GENE3D with no description, the next brown bar (PTHR11527) is for SMALL HEAT SHOCK PROTEIN FAMILY and the last brown bar is for HEAT SHOCK PROTEIN 1.7 (these brown bars from PANTHER); **(b)** Result of DnaJ protein for domain prediction/identification by InterProScan. In IPR001623 it represents the Heat shock protein Dnaj, N-terminal. The violet,

blue, red, orange and black bar is representing the domain of Dnaj_N, Dnaj (PF00226 i.e. from PFAM), Dnaj (SM00271 i.e. from SMART), DNAJ_2 (i.e. from PROFILE) and Dnaj_N (SSF46565 i.e. from SUPERFAMILY) respectively. The next blue bar is for Dnaj_C i.e. Chaperone Dnaj, C-terminal (IPR002939) from PFAM. The green bar is for Heat shock protein Dnaj from PRINTS. The Dnaj peptide binding domain is represented from black bar (i.e. IPR008971 from SUPERFAMILY). The Heat shock protein Dnaj conserved site represented by yellow bar (i.e. IPR018253 from PROSITE). The unintegrated domain is containing two bars of different colour, the violet one is from GENE3D and the brown one is from PANTHER; **(c)** Result of Type 2 Metallothionein protein for domain prediction/identification by InterProScan. The blue colour bar is showing Type 2 Metallothionein domain of PFAM. It has no unintegrated region in its domain part. It only contains the unique PFAM domain.

Discussion:

Peanut (*A. hypogaea*) is an important oilseed crop having major nutritional as well as economic value. It is the third major oilseed crop in the world next to soybean and cotton. Production of peanut at global level plays a significant role as far as nutrition, fuel and energy, sustainable agriculture and enhanced productivity is concerned. To enhance crop production, genetic improvement through genomic research is more beneficial **[2]**. Numerous applications and recent developments from the EST concept have been made after it was proposed by Adams *et al* **[13]**. EST sequencing was aimed to clone the full length sequences of genes with agronomic value. Large-scale EST sequencing offers an access into the genome of an organism. The ESTs gave significant evidence about its coding content and expression patterns in different tissues of plants in different environmental conditions. The EST resource of *A. hypogaea* will play major role for genome sequencing and gene expression analysis in different abiotic stresses. Therefore, to get necessary information about genes involved in abiotic stresses, ESTs were selected which currently encompass more entries in the public databases than any other form of sequence data. Thus, EST datasets provide a vast resource for gene identification and expression profiling. By doing EST assembly, researchers will be able to learn about tissue specificity and expression profiling of peanut genes **[14]**. From introduction of Peanut Genome Initiative, till March 2012, the peanut research community has deposited 2, 52,832 ESTs in the public NCBI database **[2]**. Out of total ESTs of peanut, 1, 78, 490 ESTs belongs to cultivated peanut (*A. hypogaea*) only.

For identifying drought stress responsive putative genes in peanut (*A. hypogaea*) ESTs, five drought stress responsive gene families and its corresponding EST sequences were shortlisted from the experimental data of the heat map **[6]**. For each shortlisted gene family, their corresponding ESTs were selected. Total 22 ESTs belonging to five drought stress responsive gene families were selected from the supplementary data described by Govind *et al* **[6]** and shown in **Table 1 (see supplementary material)**. The whole ESTs of *A. hypogaea* (1, 78,490 ESTs) including 22 ESTs screened during this study were downloaded from the public NCBI database. For the identification of similar ESTs of each screened 22 ESTs in whole peanut ESTs, Standalone BLAST (2.2.26) analysis was performed. The screened ESTs were taken as a query sequence and the whole peanut ESTs as a database. Similar EST sequences were shortlisted on the basis of lowest e-value and highest bit score. We found that total of 3073 ESTs in which 826, 286, 17, 66 and 1878 ESTs belongs to Hsp17.3 KDa, Desiccation protective protein LEA5, DnaJ heat shock protein, Hsp70-60 KDa chaperonin and Type 2 metallothionein, respectively **(Table 1).**

Phylogenetic analysis was performed to find the appropriate evolutionary relationship between ESTs of individual gene family with representative genes. Our phylogenetic result showed that individual gene family has close relation with respect to their representative genes. We noticed that HSP 17.6 KDa family, out of 8 ESTs, one EST (gi|108954762) is closely related with gi|350534973 of *Solanum lycopersicon*. In desiccation protective protein family, out of 4 ESTs, two ESTs (gi|108954830 and gi|108954989) are closely related with gi|388503839 of *Lotus japonicus*. In Type 2 metallothionein family, out of 6 ESTs, three ESTs (gi|108954648, gi|108954669 and gi|108955131) are closely related with gi|71040682 of *A. hypogaea* itself. Similarly in Hsp70-60 KDa chaperonin family, out of 3 ESTs, one EST (gi|108955006) is closely related with gi|297722722 of *Oryza sativa* **(Figure 1a, 1b, 1c & 1d).**

The shortlisted ESTs were further normalized by removing vector sequences, Poly-A tail and masking of low complexity region using different online available software's. We found that out of 3073 ESTs, 2931 EST sequences did not contain vector sequences were used for further analysis. As reported by Joshua et al **[15]**, a number of quality features are evaluated for EST assembly such as: (a) the occurrence of chimeric contigs; (b) the constancy of mate-pairs in the same contig; (c) phylogenetic analysis using known genes and their relationships; and (d) the extent of redundancy among the assembly's contigs and singletons. In an ideal assembly, for the generation of a single contig, ESTs transcribed from a single gene are conjoined together. A phylogenic approach was also used to assess EST assembly quality **[16]**. After normalization, EST sequences were assembled for finding the contigs using DNABaser (v3.5.3) **(Figure 2).** These contigs were further BLAST (v2.2.8) with Arabidopsis Whole Genome (TAIR v10) **Table 2 (see supplementary material).** On the basis of lowest e-value, out of five, three gene families (i.e. 17KDa Class II Heat Shock Protein, DnaJ protein and Type 2 Metallothionein) were selected. WISE2 (v 2.1.20) was performed to select putative candidate genes with known functions as well as intron and exon identifications **[17, 18]**. Based on highest WISE2 score (in bits) six putative candidate genes were found and their respective families with their roles are mentioned **(Table 2).**

The functional characterization of protein depends upon their 3D structure, therefore from the resulting assembled contig sequences, conserved protein domains were assessed from InterProScan (http://www.ebi.ac.uk/Tools/pfa/iprscan) **[19]** which allows to scan protein sequence for matches against the InterPro collection of protein signature databases. The different combinations of domains give rise to the diverse range of proteins found in nature. The identification of domains that occur within proteins can thus provide insights into their function. We also predicted the possible domains in three protein sequences of HSP17.3KDa protein **(Figure 3a)**, DnaJ protein **(Figure 3b)** and Type 2 Metallothionein protein **(Figure 3c)** by using InterProScan. InterProScan of HSP17.3KDa protein

has Alpha crystallin/HSP20 domain (IPR002068) as well as HSP20-like Chaperone and other unassigned HSP20 domains **(Figure 3a).** Likewise, DnaJ protein showed N-terminal HSP domain (IPRO01623), C-terminal HSP domain (IPRO02939) and others **(Figure 3b).** Again, metallothionein type 2 has entirely a single domain of metallothionein family-15 type 2 domains (IPRO00347) **(Figure 3c).**Earlier it was reported that Small heat shock proteins (sHSPs) varies from 15–42 kDa **[20]** that play a significant role in providing resistance to variety of abiotic stresses such as drought, salt, cold and oxidants besides high temperature **[21, 22]**. Tzi *et al* **[23]** reported that DnaJ protein plays an important role in various stress response mechanism. Clement *et al* **[24]** reported that DnaJ protein interacts with Hsp70 that might also be involved with ABA signalling. Metallothioneins are associated in metal detoxification, homeostasis, and protection during oxidative damage **[25]**. It was reported that metallothionein transcripts were present in both cultivated peanut and wild species **[26, 1]**.

Simultaneously, we predicted the 3D structure of corresponding protein sequences of contigs HSP17.3KD protein **(Figure 4a)**, DnaJ protein **(Figure 4b)** and Type 2 Metallothionein protein **(Figure 4c)** by using I-TASSER. Biological role of above three proteins were also noted from I-TASSER result. The quality of the generated models are estimated based on a confidence score (C-score), ranges from -5 to 2 where a high value signifies a model with a high confidence and vice-versa. C-score is highly correlated with TM-score and RMSD thus, TM-score and RMSD are known standards to measure the accuracy of structure modelling thereby measuring structural similarity between two protein structures. RMSD is an average distance of all residue pairs in two structures and is sensitive to local errors (i.e., a mis-orientation of the tail) which occurs inspite of the correct global topology hence, TM-score must be used for solving these errors. A TM-score >0.5 indicates a model of correct topology. Therefore, the best predicted model is selected on the basis of confidence score; TM-Score as well as RMSD value **Table 3 (see supplementary material)**. I-TASSER server predicts and displays various features in different sections for best model studies. We are considering the prediction and generation of the best model based on C-Score, their structural analogs and binding sites. The C-Score value for the best predicted model HSP17.3KD, DnaJ protein and Type 2 Metallothionein protein (0.3), (-3.18) and (-2.98) respectively and furthermore, highly similar structures in PDB (as identified by TM-align) of HSP17.3KD, DnaJ and Type 2 metallothionein were identified and listed in **Table 4 (see supplementary material)**. Template proteins with similar binding sites for HSP, DnaJ and Type 2 metallothionein were listed in **Table 5 (see supplementary material)**. The best binding site is predicted on the basis of $C\text{score}^{LB}$ (Range = 0-1) and BS-Score (>1) values. A higher score $C\text{score}^{LB}$ indicates a more reliable ligand-binding site prediction and BS-score reflects a significant local match between the predicted and template binding site **(Table 5) [11, 12]**.

We found that HSP 17.1 KDa has important role in some molecular function (GO: 0005212) that contributes in the structural integrity of the cell and less role in creating protein oligomers (GO: 0051260). Similarly DnaJ interact selectively and non-covalently with unfolded proteins (GO: 0051082) and is involved in assisting the covalent and non-covalent assembly of single chain polypeptides or multi-subunit complexes into the correct tertiary structure (GO: 0006457). On the other hand, type 2 Metallothionein interact selectively and non-covalently with an adenyl-ribonucleotide, any compound consisting of adenosine esterified with (ortho) phosphate or an oligophosphate at any hydroxyl group on the ribose moiety (GO:0032559) and is thus involved in the chemical reactions and pathways involving the phosphate group, the anion or salt of any phosphoric acid (GO:0006796). Govind *et al* **[6]** described that for cellular mechanisms, groups of gene products are involved that act in coordination in response to stimuli of water withhold.

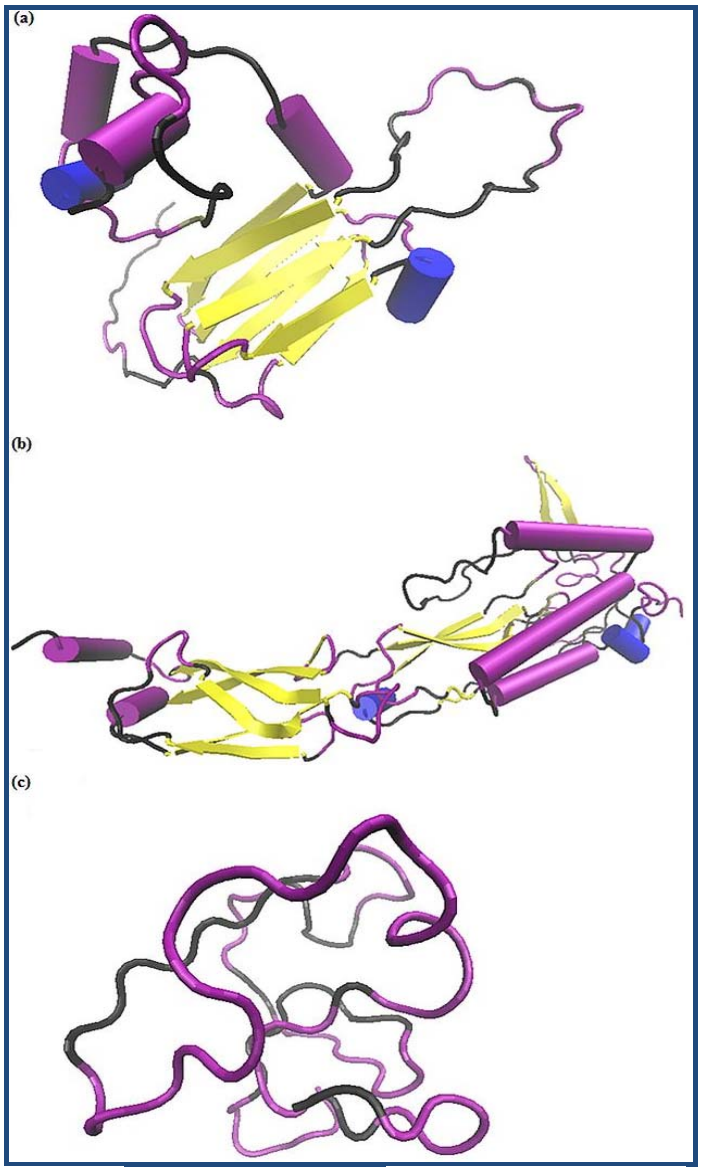

Figure 4: (a) Result showing 3D structure of HSP17.3KDa protein predicted by I-TASSER. The coloring method is based on secondary structure. The cartoon shaped structure representing α helix and the yellow colored arrow representing β-sheets. The black color part of ribbon representing coil and the purple color representing turn in this 3D structure; **(b)** Result showing 3D structure of DnaJ protein predicted by I-TASSER. The colouring method is based on secondary structure. The cartoon shaped structure representing α helix and the yellow colored arrow representing β-sheets. The black

colour part of ribbon representing coil and the purple colour representing turn in this 3D structure; **(c)** Result showing 3D structure of Type 2 Metallothionein protein predicted by I-TASSER. This structure is not containing any α helix and β-sheets. But it is completely made up of only coil (black colour part of ribbon) and turns (in purple colour).

Conclusion:

The present *in silico* study allowed identification of six best putative candidate genes belonging to three different gene families for drought stress response in peanut. Contigs, domains as well as 3D structure of HSP 17.3KDa protein, DnaJ protein and Type 2 Metallothionein protein were also predicted. Thus, EST approach play key role in providing an excellent resource for novel gene discovery and its annotation for drought stresses.

Acknowledgement:

Birla Institute of Technology, Mesra, Ranchi, Jharkhand, India is greatly acknowledged for providing Institute Fellowships to Mr. Ashutosh Kumar and Ms. Archana Kumari. UGC, New Delhi, India is also acknowledged for providing Fellowship to Mr. Gopal Kr. Prajapati. DBT, New Delhi, India is greatly acknowledged for providing Bioinformatics Facility at our Institute.

References:

[1] Luo M *et al*. *Crop Sci*. 2005 **45:** 346
[2] Feng S *et al*. *Comp Funct Genomics*. 2012 [PMID: 22745594]
[3] Gai X *et al*. *Nucleic Acids Res.* 2000 **28:** 94 [PMID: 10592191]
[4] Shoemaker R *et al*. *Genome.* 2002 **45:** 329 [PMID: 11962630]
[5] Lazo GR *et al*. *Biotechniques*. 2001 **30:** 1300 [PMID: 11414222]
[6] Govind G *et al*. *Mol Genet Genomics*. 2009 **281:** 591 [PMID: 19224247]
[7] Trebichalský A *et al*. *Journal of Microbiology, Biotechnology and Food Sciences*. 2012 **1:** 711
[8] Dong Q *et al*. *Plant Physiol*. 2005 **139:** 610 [PMID: 16219921]
[9] Cheng D *et al*. *Science China*. 2012 **55:** 452
[10] Birney E *et al*. *Genome Res*. 2004 **14:** 988 [PMID: 15123596]
[11] Zhang Y *et al*. *BMC Bioinformatics*. 2008 **9:** 40 [PMID: 18215316]
[12] Roy A *et al*. *Nature Protocols*. 2010 **5:** 725 [PMID: 20360767]
[13] Adams MD *et al*. *Science*. 1991 **252:** 1651 [PMID: 2047873]
[14] Siddanna BS *et al*. *American Journal of Plant Sciences.* 2012 **3:** 1169
[15] Joshua AU *et al*. *Genome Res.* 2006 **16:** 441 [PMID: 16478941]
[16] Close TJ *et al*. *Plant Physiol.* 2004 **134:** 960 [PMID: 15020760]
[17] Tamura M & Tachida H, *Mol Genet Genomics*. 2011 **285:** 393 [PMID: 21442326]
[18] Kumar A *et al*. *Int J Bioinform Res Appl.* 2011 **7:** 376 [PMID: 22112529]
[19] Quevillon E *et al*. *Nucleic Acids Res.* 2005 **33** (Web Server issue): W116 [PMID: 15980438]
[20] Vierling E *et al*. *Acta Physiol Plant.* 1997 **19:** 539 [PMID: 12296361]
[21] Banzet N *et al*. *Plant J.* 1998 **13:** 519 [PMID: 9680997]
[22] Sato Y *et al*. *Plant Cell Rep.* 2008 **27:** 329 [PMID: 17968552]
[23] Tzi BN *et al*. *Biopolymers.* 2012 **98:** 268 [PMID: 23193591]
[24] Cle´ment M *et al*. *Plant Physiol.* 2011 **156**: 1481 [PMID: 21586649]
[25] Akashi K *et al*. *Biochem Biophys Res Commun.* 2004 **323:** 72 [PMID: 15351703]
[26] Proite K *et al*. *BMC Plant Biol.* 2007 **7:** 1 [PMID: 17302987]

Edited by P Kangueane
Citation: Kumari *et al.* Bioinformation 8(24): 1211-1219 (2012)

BIOINFORMATION

open access

Supplementary material:

Table 1: Identification of selected ESTs of each gene family by standalone BLAST

EST IDs	No. of selected ESTs after BLAST analysis	EST IDs	No. of selected ESTs after BLAST analysis
17.3 KDa class II heat shock protein (Hsp17.3)			
gi|107954700	2	gi|108954720	113
gi|108954815	4	gi|108955108	147
gi|108955133	15	gi|108955061	148
gi|108954762	30	gi|108954987	367
Desiccation protective protein LEA5			
gi|108954782	36	gi|108954830	84
gi|108954989	82	gi|108954970	84
DnaJ heat shock N-terminal domain containing protein			
gi|108954994	17		
Hsp70-60 KDachaperonin			
gi|108954655	22	gi|108955170	22
gi|108955006	22		
Type 2 metallothionein			
gi|108955024	126	gi|108954648	500
gi|108955082	126	gi|108954669	500
gi|108955151	126	gi|108955131	500

Table 2: Selection of putative candidate genes with the help of TAIR and WISE2 result

Family Name	Assembled Contigs	TAIR Result			WISE2 Result	
		Arabidopsis Gene Id	E-Value	Description	Exon position	WISE2 Score (in bits)
17KDa Class II Heat Shock Protein	108954762	AT5G12030.1	5.00E-04	At-Heat Shock Protein 17.6 Kda	96-554	215.62
	108954815	AT5G37670.1	2.00E-08	HSP20-like chaperones superfamily protein	55-489	213.46
	108955061	**AT2G29500.1**	**7.00E-25**	**HSP20-like chaperones superfamily protein**	**238-687**	**283.27**
		AT3G46230.1	2.00E-21	At Heat Shock Protein 17.4 Kda	238-687	266.81
	108955108	**AT2G29500.1**	**7.00E-25**	**HSP20-like chaperones superfamily protein**	**238-687**	**283.27**
		AT3G46230.1	2.00E-21	At-Heat Shock Protein 17.4 Kda	238-687	266.81
DnaJ heat shock protein	**108954994**	**AT5G01390.4**	**1.00E-23**	**DnaJ heat shock family protein**	**205-1023**	**409.42**
Type 2 Metallothionein	**108954648**	**AT3G09390.1**	**7.00E-09**	**At Metallothionein-1**	**115-354**	**146.79**
		AT5G02380.1	0.096	Cysteine-rich protein with copper-binding activity	115-354	130.76
	108954669	**AT3G09390.1**	**7.00E-09**	**At Metallothionein-1**	**115-354**	**146.79**
		AT5G02380.1	0.096	Cysteine-rich protein with copper-binding activity	115-354	130.76
	108955024	AT3G15353.1	1.00E-06	Metallothionein, binds to and detoxifies excess copper and other metals	72-266	89.56
	108955082	AT3G15353.1	1.00E-06	Metallothionein, binds to and detoxifies excess copper and other metals	72-266	89.56
	108955131	**AT3G09390.1**	**7.00E-09**	**At Metallothionein-1**	**115-354**	**146.79**
		AT5G02380.1	0.096	Cysteine-rich protein with copper-binding activity	115-354	130.76
	108955151	AT3G15353.1	1.00E-06	Metallothionein, binds to and detoxifies excess copper and other metals	72-266	89.56

Table 3: Best predicted model with their C-Score, TM Score and RMSD value where C-Score is the confidence score for the predicted model, TM-score is a measure of global structural similarity between query and template protein and Root Mean Square Deviation is the RMSD between residues that are structurally aligned by TM-align.

Best predicted model			
Best Model	**C-Score**	**TM Score**	**RMSD**
HSP best Model 1	0.3	0.75 ± 0.10	4.2 ± 2.8Å
DnaJ best Model 1	-3.18	0.36 ± 0.12	14.2 ± 3.8Å
Type 2 metallothionein 1	-2.98	0.38 ± 0.13	10.0 ± 4.6Å

Table 4: Identified best two structural analogs in PDB where Coverage represents the coverage of global structural alignment and is equal to the number of structurally aligned residues divided by length of the query protein.

Top 2 Identified structural analogs in PDB				
Gene Family	**PDB Hit**	**TM-score**	**RMSD**	**Coverage**
HSP	1gmeA	0.864	1.47	0.908
	2byuI	0.613	0.54	0.621
DnaJ	1nltA	0.62	1.7	0.651
	3agzA	0.484	2.78	0.549
Type 2 metallothionein	1sj8A	0.486	3.97	0.914
	2o8rA	0.467	3.78	0.864

Table 5: Best template protein for similar binding sites.

Template proteins with similar binding site						
Gene Family	**Cscore^{LB}**	**PDB Hit**	**TM-score**	**RMSD**	**BS-score**	**Ligand Name**
HSP	0.04	3l1e0	0.493	2.79	1.03	PEPTIDE
DnaJ	0.18	3agyB	0.461	2.73	1.52	PEPTIDE
Type 2 metallothionein	0.01	3hkeB	0.459	4.07	0.57	T13