# Identification of Synchronized Role of Transcription Factors, Genes and Enzymes in *Arabidopsis thaliana* Under Four Abiotic Stress Responsive Pathways


**A DISSERTATION SUBMITTED TO BRAC UNIVERSITY**
**IN PARTIAL FULFILMENT OF THE REQUIRMENTS FOR**
**THE MASTER OF SCIENCE IN BIOTECHNOLOGY**

**Submitted by-**

**Rabab Mahdi**

**Student ID: 11376001**

**March 2014**

**Biotechnology Program**

**Department of Mathematics and Natural Sciences**

**BRAC University**

**Bangladesh**

*Dedicated*

*To*

*My beloved parents*

# DECLARATION

I hereby declare that the research work embodying the results reported in this thesis entitled "**Identification of Synchronized Role of Transcription Factors, Genes and Enzymes in *Arabidopsis thaliana* Under Four Abiotic Stress Responsive Pathways"** submitted by the undersigned has been carried out under supervision of Dr. Aparna Islam, Associate Professor, Department of Mathematics and Natural Sciences, BRAC University, Dhaka. It is further declared that the research work presented here is original and has not been submitted to any other institution for any degree or diploma.

**Candidate:**

_____

**Rabab Mahdi**

**Certified:**

_____

**Dr. Aparna Islam**

**Supervisor**

Associate Professor

Department of Mathematics and Natural Sciences

BRAC University, Dhaka

# Acknowledgement

It's my honor to write this acknowledgement to express my gratefulness to people who have helped me through this research work.

First of all I would like to thank The Almighty Allah for successfully accomplish my thesis work.

I am grateful to Professor A. A. Ziauddin Ahmad, Chairperson, Department of Mathematics and Natural Sciences, BRAC University, for allowing me to pursue my post graduate studies in the department of MNS and for his constant guidance and help throughout my entire period of study in the department.

My humble respect to my honorable Professor Naiyyum Chowdhury, Coordinator, Biotechnology and Microbiology, Department of Mathematics and Natural Sciences, BRAC University, for his inspiration and prudent advice and also giving me the opportunity of gaining significant experience.

I would like to convey my heightened appreciation to all my respected teachers of the Department of Mathematics and Natural Sciences, BRAC University, for their academic counsel and encouragement.

I also express my special thanks to Dr. Aparna Islam, Associate Professor, Department of Mathematics and Natural Sciences, BRAC University, for her valuable suggestions, unforgettable help and soft behavior also affectionate guidance and for giving me the opportunity to work under her supervision particularly.

I would particularly like to express my deepest thanks to Smasad Razzaque for his hearty, dateless, incessant cooperation and encouragement throughout the study. I gratefully acknowledge him for his advice and crucial contribution throughout my research and thesis writing periods.

Also I thank deeply from my heart to my lovely and helpful seniors Manzur-E-Mohsina Ferdous and Shahana Chowdhury, who provided constant encouragement, sound advice, good company, and lots of good ideas throughout the study and also I would like to thank my friends and juniors for their enthusiastic inspiration and company during my thesis work.

The whole credit of my achievements during the research work goes to my parents. It was their cordial effort that makes me acquaintance with the way of research.

**Rabab Mahdi**
**Department of Mathematics and Natural Sciences**
**BRAC University, March 2014.**

## ABSTRACT

DNA microarray databases are widely used tool to predict and identify functional genomics and gene expression. This plant genomic recourse can solve various biological questions related to biotic and abiotic stresses at molecular as well as cellular levels. Under stress conditions, various types of proteins and DNA sequences play important role in plant to withstand the adverse conditions. These elements can be transcription factors, regulatory genes and enzymes which independently or in connection with each other attain this task. The present study was conducted to identify these DNA sequences through microarray datasets and find out the connection between their products, and to produce a network in model plant *Arabidopsis thaliana.* Four microarray datasets responding to abiotic stresses, like, heat, cold, drought and abscisic acid were considered here. Preliminary study started with ArrayExpress which gave four groups of DNA sequences for each stress signals. To figure out the common physiological characters between these thousands of genes, 42 common genes were found to be up-regulated during the selected stresses. Among them, 30 found to be closely related. Further bioinformatics study and also literature mining showed that of these 30, eight genes, like, DREB2A, P5CS1, ERD5, CPL1, NHX1, SOS1, SOS2 and SOS3 are highly responsive to the above abiotic stresses. Later their protein-protein networking, protein stability, conserved sequences, interactive domain and individual interectome with other genes were studied. This was done using different web based datasoft, namely, String database, GeneMania, Gene Ontology and InterProScan. The study revealed that these eight genes not only get up-regulated but also they create a connection to each other and produce a tolerance hub. In the present study, the identified genes imply a concurrent defensive role against these abiotic stresses. In future, study needs to be undertaken to validate these findings under *in vivo* condition so that the knowledge can be applied in agriculture to improve crop protection and production under variable climatic conditions.

# Content

# Abbreviation

| | |
|---|---|
| **GRNs** | **Gene regulatory networks** |
| **TFs** | **Transcription factors** |
| **BLAST** | **Basic Local Alignment Search Tool** |
| **GO** | **Gene Ontology** |
| **NCBI** | **National Center for Biotechnology Information** |
| **EBI** | **European Bioinformatics Institute** |
| **ABA** | **Abscisic acid** |
| **TAIR** | **The Arabidopsis Information Resource** |
| **Mb** | **Mega bases** |
| **UTR** | **Untranslated region** |
| **bp** | **Base pair** |
| **mRNA** | **Messenger ribonucleic acid** |

# Introduction

# Introduction

## 1.1 Plant stress and food security:

Plant stresses are the reasons for food insecurity thus threat to mankind. Environmental stress is one of the biggest problems and also already responsible for reduced crop yields worldwide. The effects of climate change have already been noted, with more variable conditions likely to lead to increased exposure to environmental stress. An increase in global temperatures will lead to drought and the expected increase in humidity is likely to increase plant susceptibility to pathogens, which is already a major source of crop spoilage all over the world. These factors are conspiring to greatly endanger food security, leading to social instability and increased poverty, particularly in developing countries. Clearly, this is not just a problem for the developing world, but is a global problem affecting the entire population.

To maintain world food supplies it is essential that we understand the mechanisms by which plants adapt to environmental stress. Plants respond to stress at both cellular and molecular level by altering the expression of many genes via different types of complex molecular signaling networks. The knowledge of these pathways including identification of the regulatory codes will provide opportunities for enhancing the ability of crops to sustain stressful conditions and increase yield through genetic manipulation.

## 1.2 *Arabidopsis thaliana*: a model system to study:

*Arabidopsis thaliana* was the first plant, and the third multicellular organism after *Caenorhabditis elegans* (The *C. elegans* Sequencing Consortium 1998) and *Drosophila melanogaster* (Adams *et al*. 2000) to be completely sequenced (The *Arabidopsis* Genome Initiative 2000). At the time, it was hypothesis that this genome sequence will open-up the deeper understanding of plant development and environmental responses through understanding the function of each and every gene. In line of this thought, present study was conducted to find out the relationship of different up-regulated genes using model plant species *Arabidopsis thaliana* which are responsible during different abiotic stresses and also try to find out the

connection between these up-regulated genes implementing computational approach through World Wide Web.

### 1.3 General Physiology of *Arabidopsis thaliana:*

*Arabidopsis thaliana* is a small flowering plant. Its genome is very small 114.5 Mb/125 Mb total spread in 5 chromosomes. It has a rapid life cycle (about 6 weeks from germination to mature seed) and easy cultivation in restricted space and produce enormous seed. For this reason, though it is not of major agronomic significance, but it offers important advantages for basic research in genetics and molecular biology. Hence, it is well established as a model organism in plant researches. *Arabidopsis* is a member of the mustard (Brassicaceae) family. The classification of *Arabidopsis thaliana* is given below,

Scientific classification of *Arabidopsis thaliana*:

| Rank | Scientific Name and Common Name |
|---|---|
| Kingdom | Plantae – Plants |
| Subkingdom | Tracheobionta – Vascular plants |
| Superdivision | Spermatophyta – Seed plants |
| Division | Magnoliophyta – Flowering plants |
| Class | Magnoliopsida – Dicotyledons |
| Subclass | Dilleniidae |
| Order | Capparales |
| Family | Brassicaceae – Mustard family |
| Genus | *Arabidopsis* Heynh. – rockcress |
| Species | *Arabidopsis thaliana* (L.) Heynh. – mouseear cress |

### 1.4 Genome and Genetic studies:

Uncovering the gene and its functional relations in different species is one of the major goals of biological studies now-a-days. The central dogma of molecular biology states that genetic information is stored in DNA, which is a linear sequence of four nucleotides. When needed, this information of the DNA is transcribed into RNA, which in turn is translated into proteins, which are the main cell machinery. The genome sequence along with computational analysis of those genomic data provides us with a relatively complete set of genes and their proteins. In addition to this data analysis of the microarray data gives a picture of when and how genes are transcribed, which is a rough estimation of protein abundance. In the backdrop of this information the natural next step is to study how these proteins perform their desired functions.

Proteins play a major role in cellular processes. Proteins, however, do not act alone; they work together to create various biological processes in a hierarchical fashion. First, multiple proteins physically bind together to form stoichiometrically stable complexes. These complexes interact with each other to form functional modules and pathways that carry out most cellular processes.
In the last couple of decades, large-scale data have been accumulated for many types of interactions, varying from social interactions through links between pages of the World Wide Web and to various types of biological relations between proteins. Visualization of such data as networks and analysis of the properties of these networks has proven useful to explore these complex systems (Alon 2003; Yamada and Bork 2009; Boone *et al*. 2007; Handcock and Gile 2010).

In the post-genomic era, understanding these cellular systems is becoming increasingly important for biologists. Several methodologies are available (Uetz *et al.* 2000; Ho *et al.* 2002; Bhalla *et al.* 2005; Dong *et al.* 2005; Rensink and Buell 2005). But the data generated from these methods covers just a few complexes or pathways and is limited to a handful of model organisms. As a result, computational bioinformatics methods have been developed to integrate this data and to extrapolate from it to provide predictions for proteins and organisms not yet experimentally well characterized (Date and Marcotte 2003; Morett *et al.* 2003; Strong *et al.* 2003; McDermott and Samudrala 2004; Yu *et al.* 2004; Wichadakul *et al.* 2007).

### 1.5 Central dogma: multiple levels of gene expression regulation:

The genomes in higher eukaryotes contain thousands of genes, which encode the proteins, and RNAs that perform all of the structural and biochemical functions within a cell. The expression of these genes must be controlled to ensure the relevant gene products are produced at the appropriate time and place within an organism. Regulation of gene expression is complex and can occur at multiple levels which are depicted in a simplified manner in Fig. 1.1 (Lodish H *et al.* 2000).

Within the nucleus, DNA wraps around histone proteins, which assemble into higher-order structures known as nucleosomes. This coiled like DNA-protein complex- chromatin allows the entire genome to be compacted within a single nucleus which provides the highest level of gene regulation in the system. Chromatin manipulates gene usage by impeding the access of TFs and RNA polymerase to the DNA. To allow transcription to proceed, the condensed chromatin must open up, allowing the DNA to interact with all the factors needed to initiate transcription. Manipulation of chromatin states requires the function of chromatin re-modelling factors, which typically function by chemically-modifying histone proteins. These modifications, such as acetylation and methylation, interfere with positive-charged histones, which disrupt the interaction with negatively charged DNA and therefore lead to more open and accessible DNA.
When free of the restrictions imposed by nucleosomes, gene expression is then regulated at the transcriptional level by regulatory proteins that bind to sites on the open, transcriptionally active DNA and act to increase or decrease the rate of mRNA production.

In addition to regulation at the DNA level, a gene can be regulated post-transcriptionally at the RNA level. The mRNA transcript can be subject to alternative splicing, polyadenylation and degradation. A major form of post-transcriptional regulation is by the action of miRNAs, small (22 bp) non-coding RNAs which repress translation or target mRNAs for degradation (Sunkar *et al*. 2007; Fujii *et al*. 2005). Mature miRNAs interact with proteins to form the RNA-induced silencing complex (RISC), which guides this complex to mRNA by binding to complementary sites usually located within the 3'-untranslated region (UTR) of mRNA sequences. When targeted to mRNA via miRNA, RISC then uses RNase functionality to cleave the mRNA.

Following translation, the resulting peptide can be subject to extensive regulation that fine-tunes behavior further. Interactions with other proteins can modify protein function (Weltmeier *et al*. 2006; Djamei *et al*. 2007). Chemical modification such as phosphorylation can induce conformational changes which alter functionality or result in protein re-localisation (Mao *et al*. 2011). Phosphorylation can also increase protein stability (Lopez-Molina *et al*. 2001), while ubiquitination targets a protein for degradation (Hardtke *et al*. 2000).

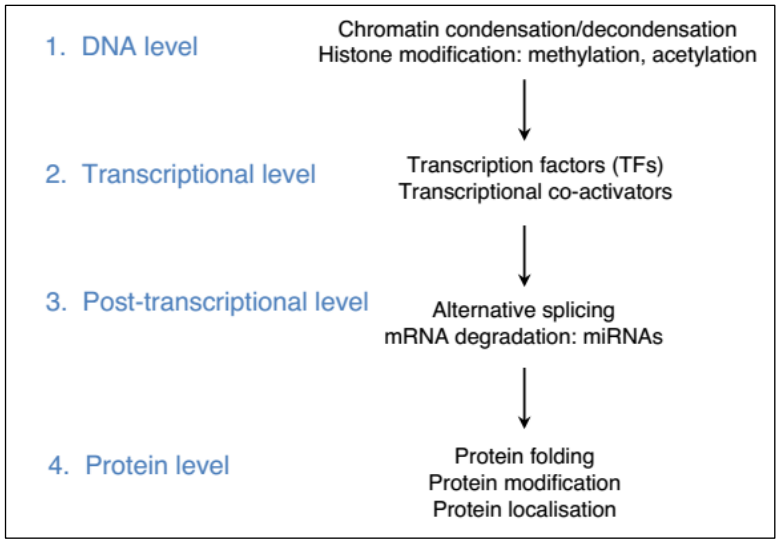


**Figure 1.1:** The multiple levels of gene regulation

### 1.6 Transcriptional gene expression regulation:

While regulation exists at multiple levels, regulation of transcription initiation is perhaps the most important (Lodish H *et al.* 2000). It is the level of control that is most efficient for the cell as it minimizes energy wastage that may result from generating unnecessary proteins and RNAs. The initiation of transcription is controlled by DNA-binding proteins called transcription factors (TFs) that bind to short sequence elements often situated in the non-coding intergenic DNA that exists between genes. TFs interact with this cis-regulatory DNA in a sequence specific manner and induce or repress transcription. The regulatory effect a TF has on transcription is often influenced by interactions with other TFs and/or co-activators that act to facilitate the assembly of the core transcriptional machinery to allow gene expression.

TFs preferentially interact with specific patterns of nucleotides known as motifs. These motifs are typically short (5-15 bp long) and can consist of any pattern of the four nucleotides, often containing degenerate positions (Harbison *et al*. 2004). TFs contact positions within the motif through specific DNA-binding domains, which interact with specific residues of closed-form DNA. Protein-DNA binding is due to non-covalent chemical interactions between the DNA-binding domain and chemical groups present on different nucleotides, and it is these specific properties of the nucleotide that are recognized by the TF. Because individual nucleotides share chemical properties, several nucleotides can be recognized by a DNA-binding domain, leading to degeneracy in the TF binding site. Moreover, not all of the positions within a motif explicitly interact with the TF, resulting in an additional source of degeneracy.

Motifs describe the binding specificities of a TF by summarizing instances of TF binding sites. By observing the sequences with which a TF can interact, a more detailed description of the TF binding specificity can be made. This information can be summarized by storing it as a matrix, which describes how often each of the four nucleotides are observed at each position of the motif. These matrixes are usually referred to as a position specific scoring matrix (PSSM) or a weight matrix (WM). The same information can then be described visually as a sequence logo, where the specificity of each position in the motif is measured in terms of information content (Schneider and Stephens 1990). Approaches to modelling the binding specificity/motif for a given TF are shown in (Figure 1.2).

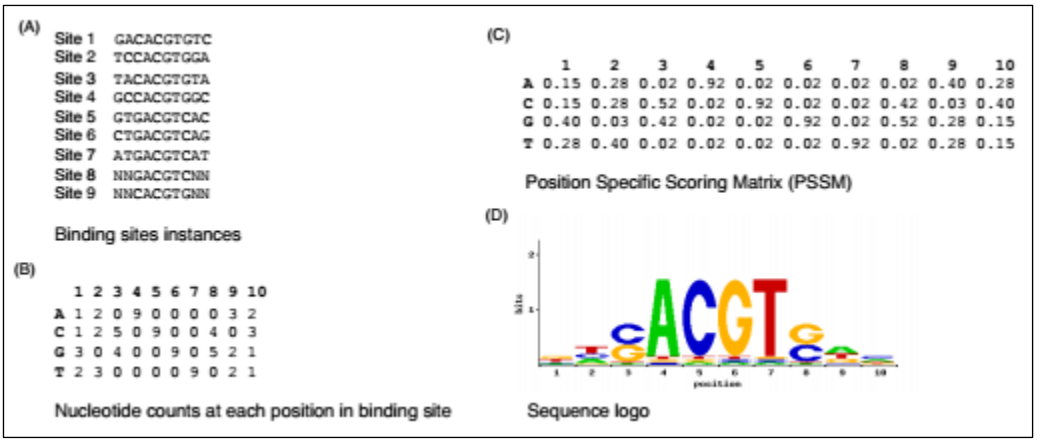


**Figure 1.2:** Modelling TF binding sites. Building models of TF binding site sequence motifs. **(A)** Defining the binding specificity of a TF requires the assembly of instances of TF binding. These sequences are aligned so that it is possible to observe the preferred nucleotides at each position in the motif. **(B)** A matrix is constructed containing the number of observed nucleotides at each position in the motif. **(C)** A PSSM/WM is generated using the matrix of observed nucleotide counts, by computing the ratio of the nucleotide count at a single position in the motif to the total number of nucleotides observed at that position. Pseudocounts are used to correct for small sample size (small number of binding site instances). **(D)** A motif can be visualized as a sequence logo, where the height of a base is proportional to its frequency at that position, while the height of the stack is proportional to the level of conservation at that position.

### 1.7 Importance of finding Genetic Correlation:

Various abiotic stresses, such as, drought, high salinity, variable temperature etc. negatively impact plant growth and productivity of crops. Plants have adapted to respond to these stresses at the molecular, cellular, physiological, and biochemical level, enabling them to survive. Various adverse environmental stresses induce the expression of a variety of genes in many plant species (Xiong *et al*. 2002; Shinozaki *et al*. 2003; Bartels and Sunkar 2005). Numerous stress-induced genes have been identified using microarray experiments (Kreps *et al*. 2002; Seki *et al*. 2002). The products of these genes are thought to promote stress tolerance and to regulate gene expression through signal transduction pathways (Xiong *et al*. 2002; Shinozaki *et al*. 2003).

Therefore, it is easily understood that different transcription factors, regulatory genes and enzymes play an important role in the regulation of gene expression in response to stresses. Interestingly not all of the genes are up-regulated, rather some of the genes are silenced which

might have a better potentials to act than others. Therefore, it is important of know these regulatory networks and functions of these genes for better understanding the stress tolerance mechanisms. This will help in further study in the agricultural sector for better yield in the farm field knowing which crops is better responsive in different stress conditions. Bangladesh as a developing country has issues in the agricultural sector facing different disaster like flood, drought and temperature fluctuations and the most important the salinity in the south belt.

To understand the correlation between these genes in plant science research, this *in silico* experiment focused on finding the connection of different up regulated genes in different abiotic stress conditions through one of the model plant like *Arabidopsis thaliana*. *Arabidopsis* is one of the model organisms for studying plant genetics and development. The genome of *Arabidopsis* is the first to be sequenced in higher plants, is believed to comprise at least 30,700 genes. Of these genes, the function of approximately one-third (9194) remain unknown according to the functional Gene Ontology (GO) category listed by the *Arabidopsis* Information Resource (TAIR) (Jiexun Li *et al*. 2006). Of the remainder, a large proportion lack complete or adequate functional annotation. We are aimed at constructing a genome-wide functional network of *Arabidopsis* by integrating relations extracted from diverse data sources**.**

To achieve continued improvement in plant traits for food security and bioenergy production will require a sophisticated understanding of the networks that control plant growth and differentiation. This research will generate high-resolution datasets from which regulatory networks controlling biological processes central to real-world agricultural and bioenergy productivity can be identified and characterized.

### 1.8 Transcriptional regulatory networks:

Differential expression of large sets of genes is required for the initiation of developmental programs and stress responses. These expression programs are primarily regulated by multiple TFs, where each TF can regulate multiple genes by binding to common sequence motifs present in non-coding DNA such as promoters. The interactions between TFs and regulatory DNA form complex network structures and ultimately drive the generation of complex expression patterns. While the end nodes of such networks are typically functional genes such as structural proteins

or enzymes, it is primarily TFs that are responsible for network architecture. The TFs themselves are also highly regulated at the transcriptional level to ensure that the appropriate regulators are expressed at the correct time and place. The pattern of cis -regulatory elements that serve as target binding locations for TFs are what underpin gene regulatory networks (GRNs). The arrangement of sequence motifs within the promoters of genes form a regulatory code that is interpreted by TFs and ultimately controls gene expression during specific conditions.

### 1.9 Methods for elucidating transcriptional networks:

Transcriptional networks can be dissected experimentally using what have been termed either \TF-centred" or \gene-centred" approaches (Walhout 2006). TF-centred approaches focus on the TF and seek to identify sites in the genome with which the TF can interact. Gene-centred methods take the opposite approach and seek to identify the TFs that interact with a specific DNA sequence. By focusing on non-coding DNA such as promoter sequences it is possible to identify TFs that may directly regulate a gene. The two types of approach complement one another and as each has its own caveats, a combination of both is needed to comprehensively map transcriptional networks operating during complex biological processes such as stress responses.

Chromatin immunoprecipitation (ChIP) based technologies are the most common and powerful TF-centred methods and have been used to identify hundreds of target genes for certain plant TFs (Morohashi *et al*. 2009). In this approach, the TF is chemically cross-linked to the DNA in vivo followed by immunoprecipitation of the TF together with the associated DNA fragment. The location of the resulting DNA fragments within the genome can then be identified using next generation sequencing (ChIP-SEQ) or microarrays (ChIP-CHIP) (Kaufmann *et al*. 2010). The drawback of ChIP-based methods, however, is the reliance on antibodies to immunoprecipitate the TF of interest. Due to the high sequence similarity amongst many TF families, generation of a specific antibody is often difficult. Even if a suitable antibody is available, if the levels of TF within the sample are low, it may be difficult to isolate sufficient levels of chromatin.

Another approach that attempts to identify the genome wide binding profile of a TF is to first elucidate its binding specificity and then use this to scan the genome for putative binding sites and target genes (Walhout 2006). Many experimental techniques exist that can reveal TF binding specificity, such as bacterial-1-hybrid, protein-binding microarrays and SELEX (Noyes *et al*. 2008; Godoy *et al*. 2011; Oliphant *et al*. 1989). This interaction specificity can then be modeled as a PSSM and used to scan the genome for instances of the motif. Predictions of the genes targeted by the TF can be improved by scanning sequences that are likely having regulatory role, such as the core promoter of a gene.

The yeast-1-hybrid (Y1H) system is one of the most successful and popular gene-centred methods and has been used to isolate many plant TFs that physically interact with regulatory DNA sequences (Tran *et al*. 2007; Chen *et al*. 2010; Zhu *et al*. 2010). This method allows for the identification of multiple TFs from different families to interact with a piece of DNA. Subsequent analysis of interactions identified using Y1H has revealed some to have a regulatory consequence in vivo (Brady *et al*. 2011), indicating that this is a powerful method for understanding transcriptional networks.

### 1.10 Present research objective:

This experiment was done to find out the up regulated genes, expressing in different abiotic stresses like cold, drought, salinity and transcriptional responses to abscisic acid and also find out the connection between these genes which are not directly connected to up regulation but also play role in abiotic stress conditions, by using microarray database from World Wide Web of *Arabidopsis* it is the research goal to find out the correlation of genes using different bioinformatics tools.

# Materials and Methods

## Materials and Methods

Protein-protein interactions can be studied by different techniques utilizing genetic, biochemical, and biophysical properties. However, the acceleration pattern with which protein sequences are now indentified or predicted has created a need for high-throughput methods for understanding the interaction as well as detection of protein function. A variety of experimental and computational approaches took place in the past several years that can solve the problem at large scale, resulting in a vast amount of interaction data in the public domain through internet.

Different types of software and web links are available in the public domains for prediction of protein-protein interaction at DNA and protein level utilizing the available complete genome sequences in the internet. In this experimental study different types of web based information databases and software were used to identified the desired proteins and their function that are required for finding the interaction between particular species in *Arabidopsis thaliana*. The web based sites and the software gives the complete knowledge for finding the interaction between proteins in DNA level analysis.

### 2.1 Work Plan:

Complete genomes are becoming available every now-and-then. Several related methods are being proposed for predicting protein interactions from these DNA sequence information. In this study different types of datasets were used to understand the up-regulation of protein function and their interaction during abiotic stress condition in *Arabidopsis thaliana*. The total work plan of the present study is given in a pictorial form in Fig. 2.1 to understand the objective of the experiments.

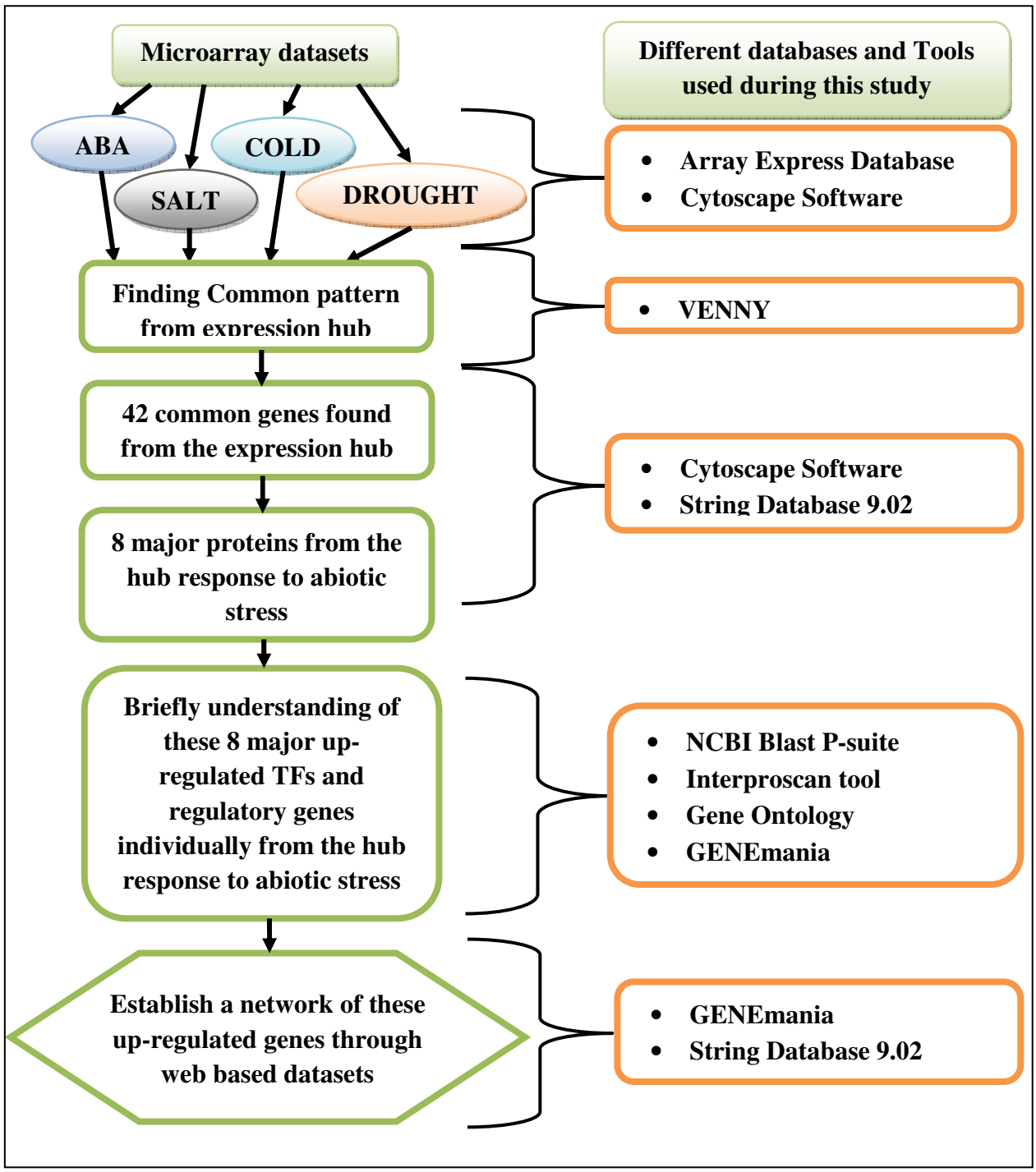


**Fig. 2.1:** Experimental work plan of protein-protein network of *Arabidopsis thaliana*

Computational approaches to understand protein-protein interactions

## 2.2 Description of Databases & Software used in this study:

### 2.2.1 ArrayExpress:

ArrayExpress is a database for analysis of functional genomics. This database includes gene expression data from microarray and high throughput sequencing studies. Two types of standards are maintained in this datasets which are MIAME (Minimum Information about a Microarray Experiment) and MINSEQE (Minimum Information about a high-throughput nucleotide Sequencing Experiment). Different experiment data can be submitted directly to ArrayExpress or imported from the NCBI GEO database. Web address for using this database is http://www.ebi.ac.uk/arrayexpress/ showed in Fig. 2.2. In current study this database was used extensively to retrieve experimental microarray data based on abiotic stresses, like salinity, drought and cold induced gene expression in *Arabidopsis thaliana*. The chosen dataset from these databases were:

- Accession: *E-GEOD-33642*: Microarray dataset after salinity stress
- Accession: *E-MEXP-3714*: Microarray dataset after cold stress
- Accession: *E-GEOD-42290*: Microarray dataset after drought stress
- Accession: *E-GEOD-45543:* Transcriptional responses to abscisic acid

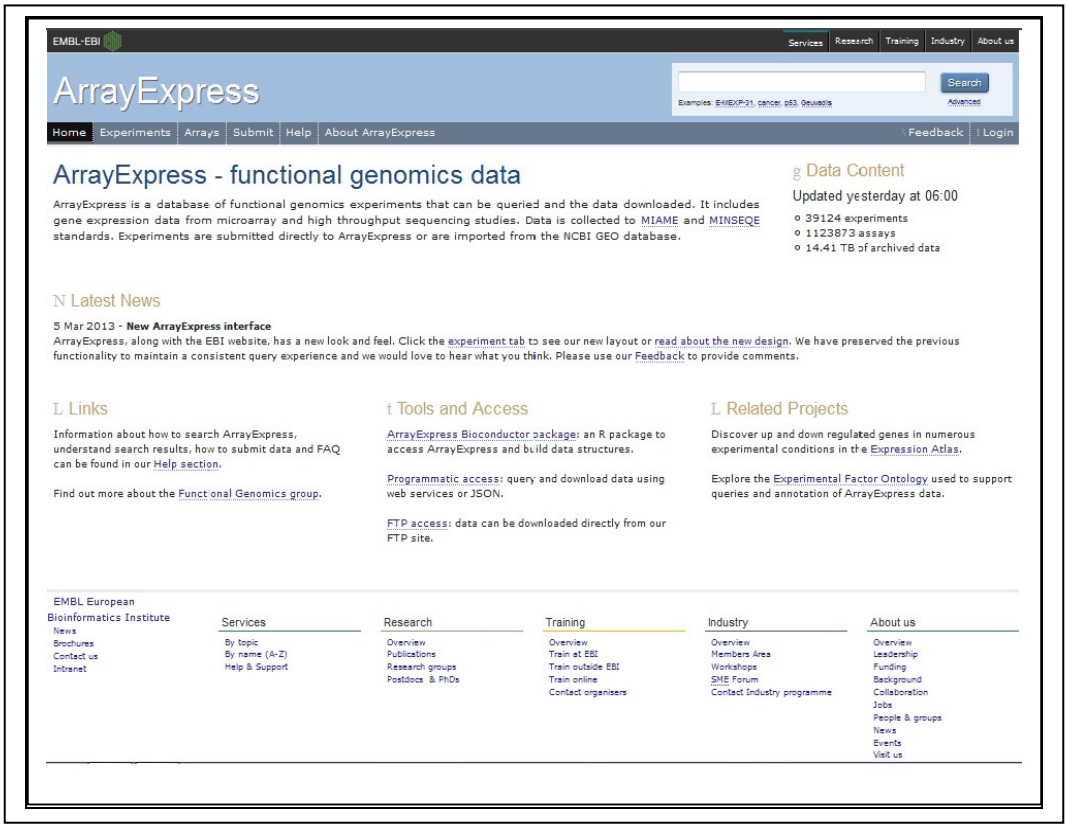


**Fig. 2.2:** Home page of ArrayExpress in the World Wide Web to understand the genome function.

### 2.2.2 Cytoscape:

Cytoscape software is an open source platform for visualization and understanding the protein-protein complex networking and integrating that with any type of attribute data. Plenty of applications are available for understanding various kinds of problem domains, including bioinformatics, social network analysis, and semantic web. Cytoscape software Fig. 2.3 was applied in this study to find out a regulatory hub of the targeted gene differentially exposed to different abiotic stresses in *Arabidopsis thaliana*.

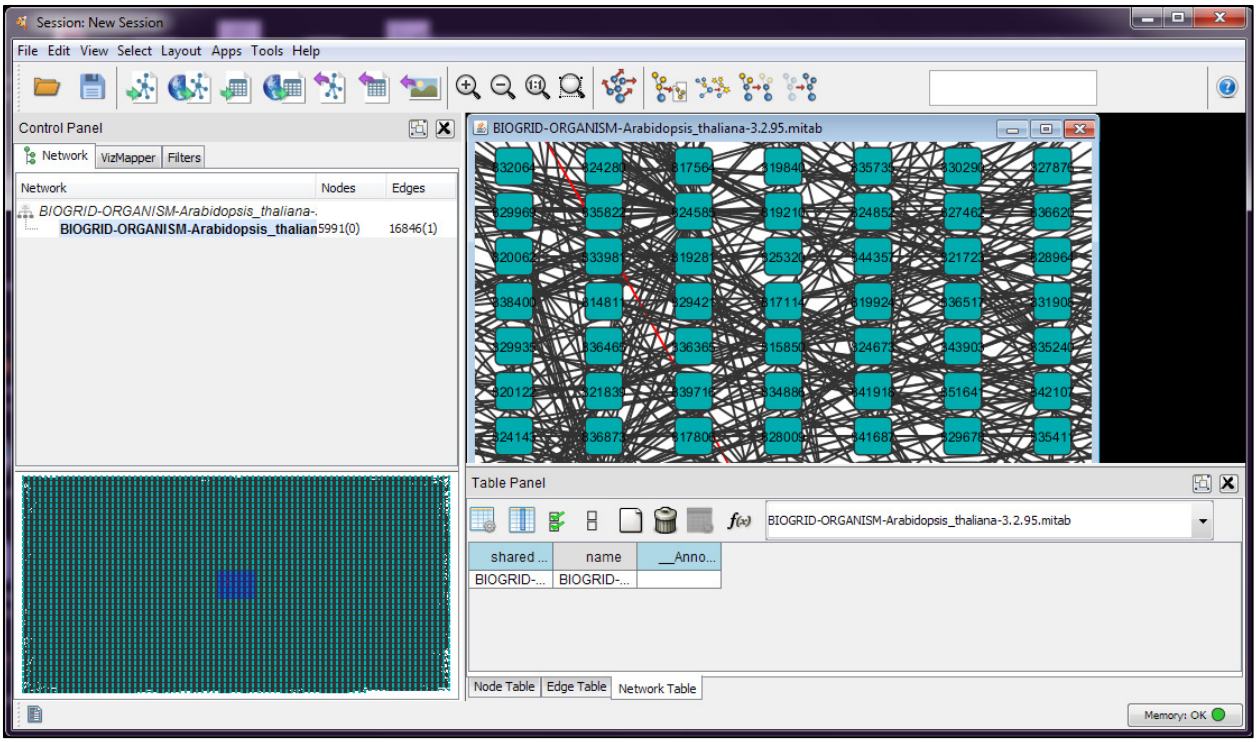


**Fig. 2.3:** Cytoscape software

### 2.2.3 Venny:

A **Venn diagram** or **set diagram** is a diagram that shows all possible logical relations between finite collections of sets. They are used to understand elementary set of theory, as well as illustrate simple set relationships in probability, logic, statistics, linguistics and computer science. Web link for visit the web site is:

http://bioinfogp.cnb.csic.es/tools/venny/ shown in Fig. 2.4.

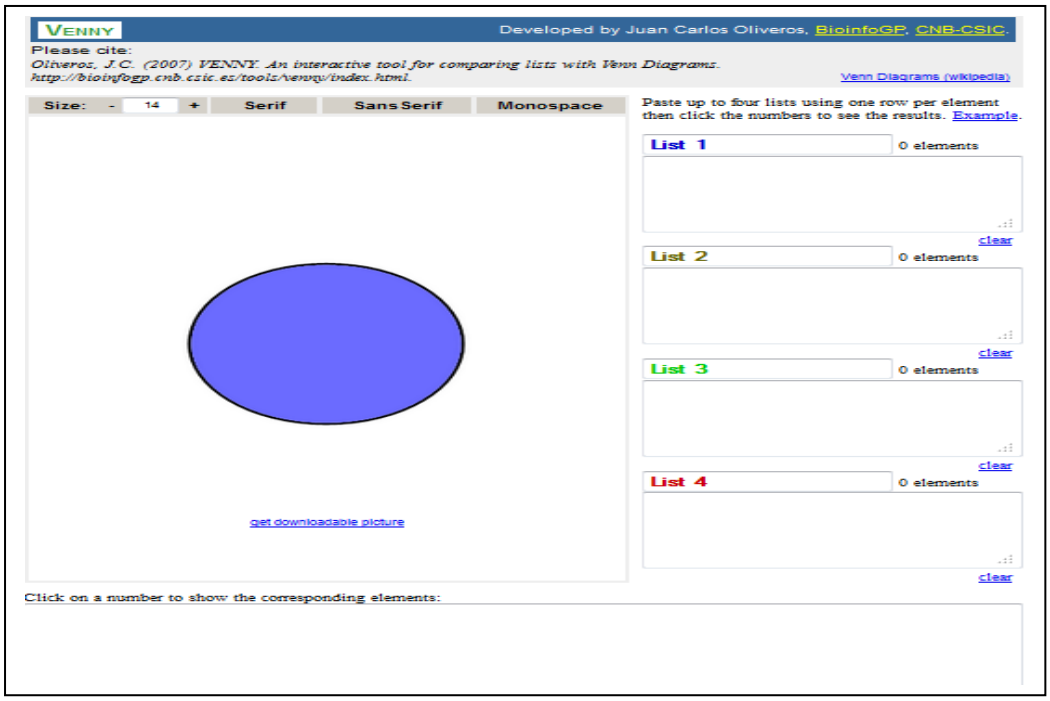


**Fig. 2.4:** Home page of Venny in the World Wide Web

### 2.2.4 European Bioinformatics Institute:

European Bioinformatics Institute (EBI) provides freely available resources for life science experiments which lead to basic research in computational biology. It is based on online database of EMBL Nucleotide Sequence Data Library also known as EMBL-Bank. The original goal is to establish and enrich the central computer database of DNA sequences of EBI. The web address for this site is http://www.ebi.ac.uk/ Fig. 2.5.

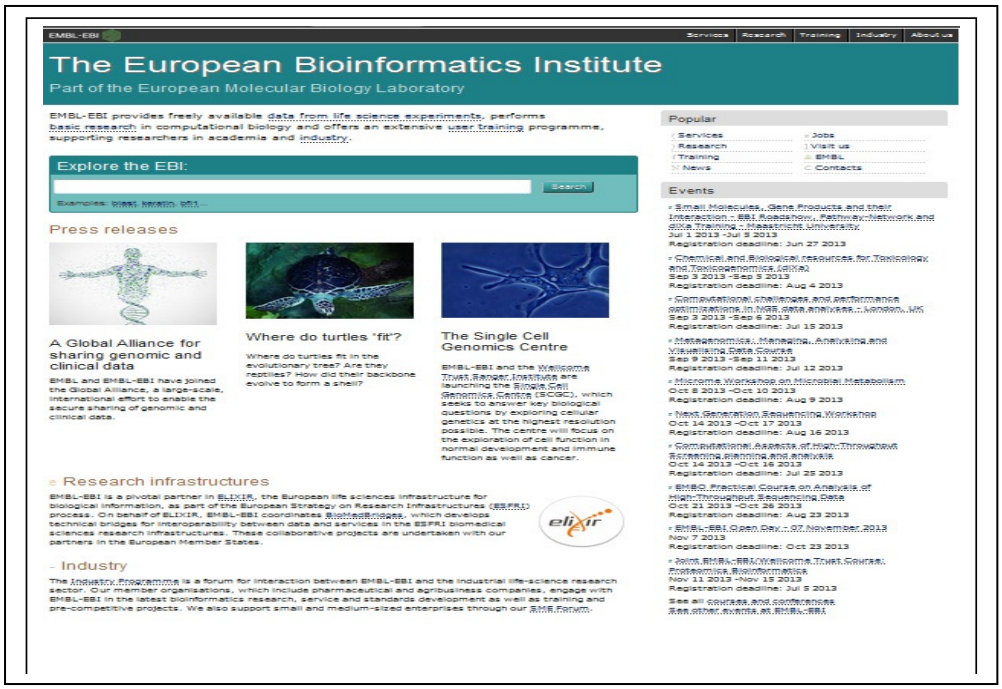


**Fig. 2.5:** Home page browsing view of EMBL-EBI

### 2.2.5 String Database:

STRING is a database of known and predicted protein interactions. The interactions include direct (physical) and indirect (functional) associations; they are derived from four sources: 1) Genomic Context, 2) High-throughput Experiments, 3) Co-expression (Conserved) and 4) Previous Knowledge. STRING quantitatively integrates interaction data from these sources for a large number of organisms, and transfers information between these organisms where applicable. The database currently covers 5,214,234 proteins from 1133 organisms. Web link to browse this database is: http://string-db.org/ shown in Fig. 2.6.

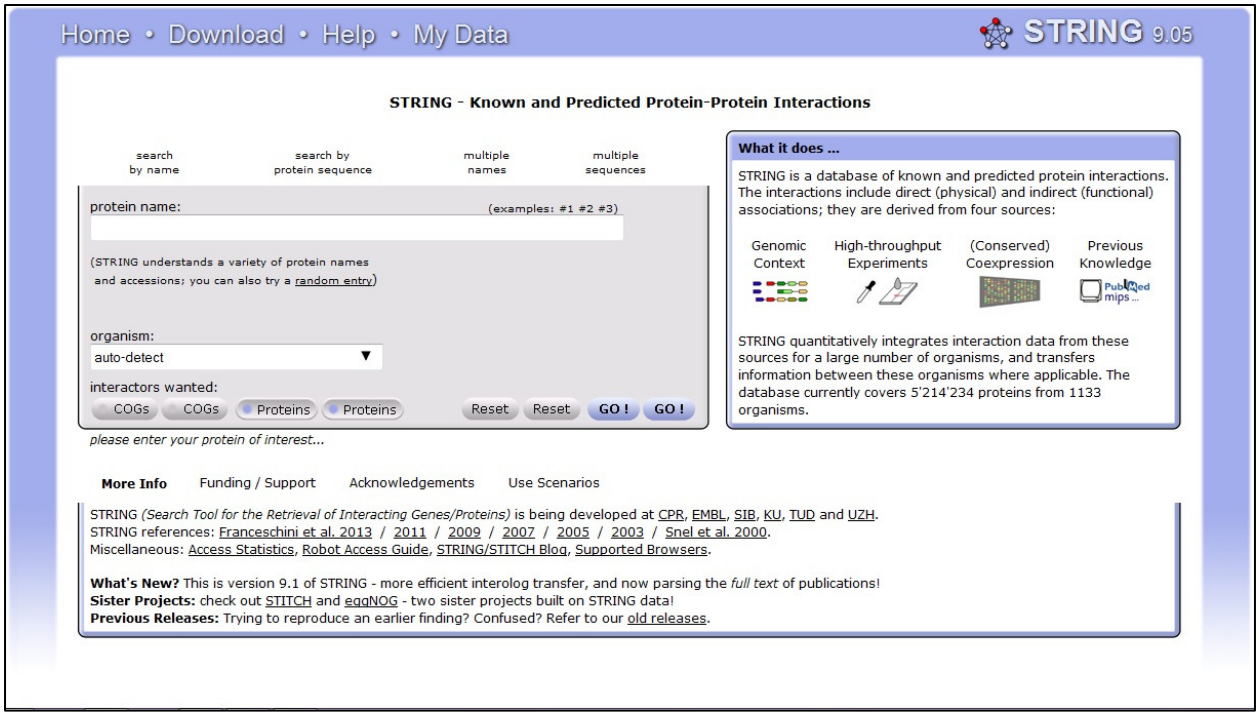


**Fig. 2.6:** Home page browsing view of String database

### 2.2.6 NCBI:

The National Center for Biotechnology Information (NCBI) is part of the United States National Library of Medicine (NLM), a branch of the National Institutes of Health. The NCBI is a bank of databases relevant to biotechnology and biomedicine sector. Major databases include GenBank for DNA sequences and PubMed, a bibliographic database for the biomedical literature. All these databases are available online through the Entrez search engine. The web link to browse this site is: http://www.ncbi.nlm.nih.gov/ shown in Fig. 2.7.

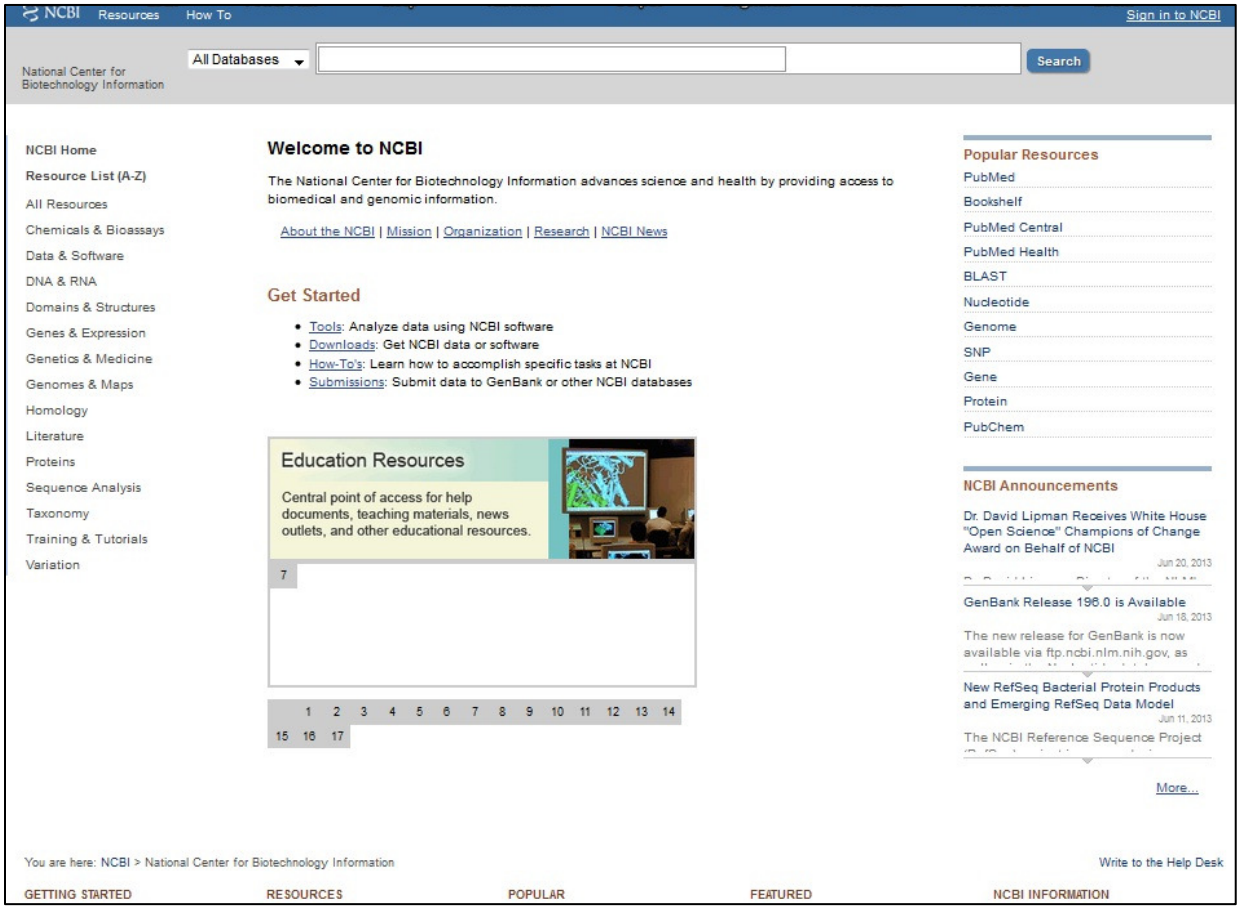


**Fig. 2.7:** Home page browsing view of NCBI

### 2.2.7 NCBI-BLAST:

The Basic Local Alignment Search Tool (BLAST) finds out the regions of local similarity between sequences. The online based program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches of different species. BLAST can be used to query for functional and evolutionary relationships between sequences as well as helps to identify members of gene families included. The web link to browse this site is: http://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch&PROG_DEF=blastn&BLAST_PROG_DEF=megaBlast&SHOW_DEFAULTS=on&BLAST_SPEC=OGP__3702 shown in Fig. 2.8.

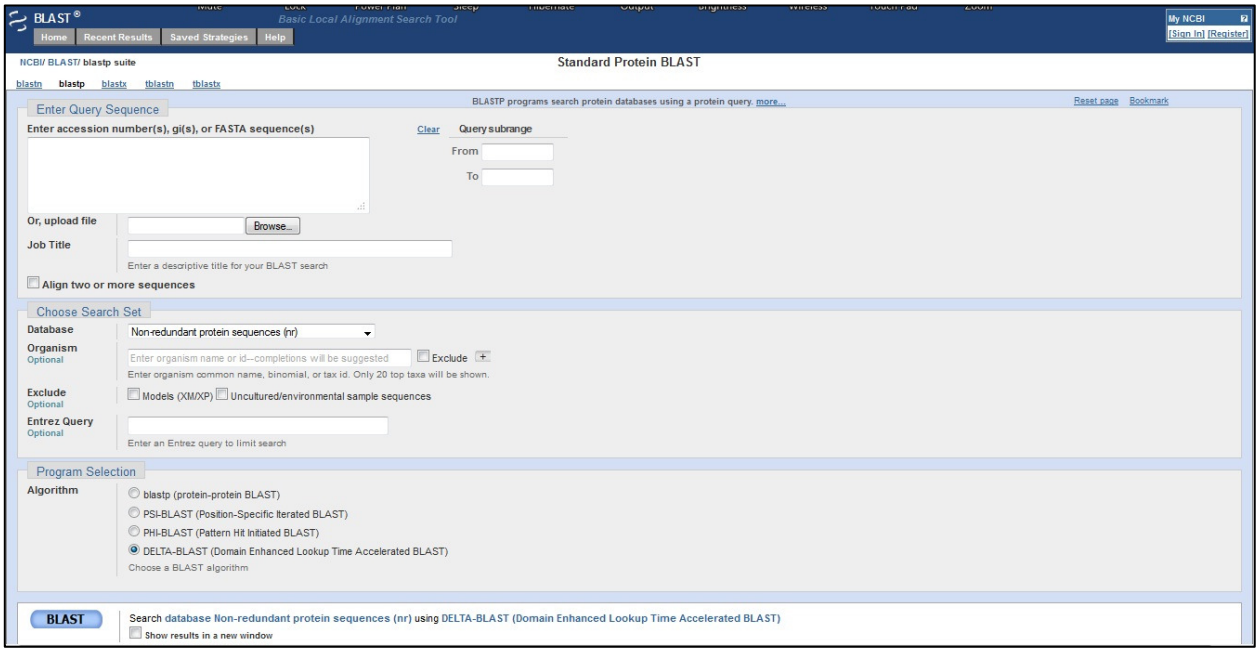


**Fig. 2.8:** Browser view of NCBI-BLAST

### 2.2.8 Interproscan:

This is the software for protein sequence analysis and classification. InterPro provides functional analysis of proteins by classifying them into families and predicting domains and important sites within them. It combines protein signatures from a number of member databases into a single searchable resource, capitalizing on their individual strengths to produce a powerful integrated database and diagnostic tool. The web link of browsing this site is: http://www.ebi.ac.uk/Tools/pfa/iprscan/ shown in Fig. 2.9.

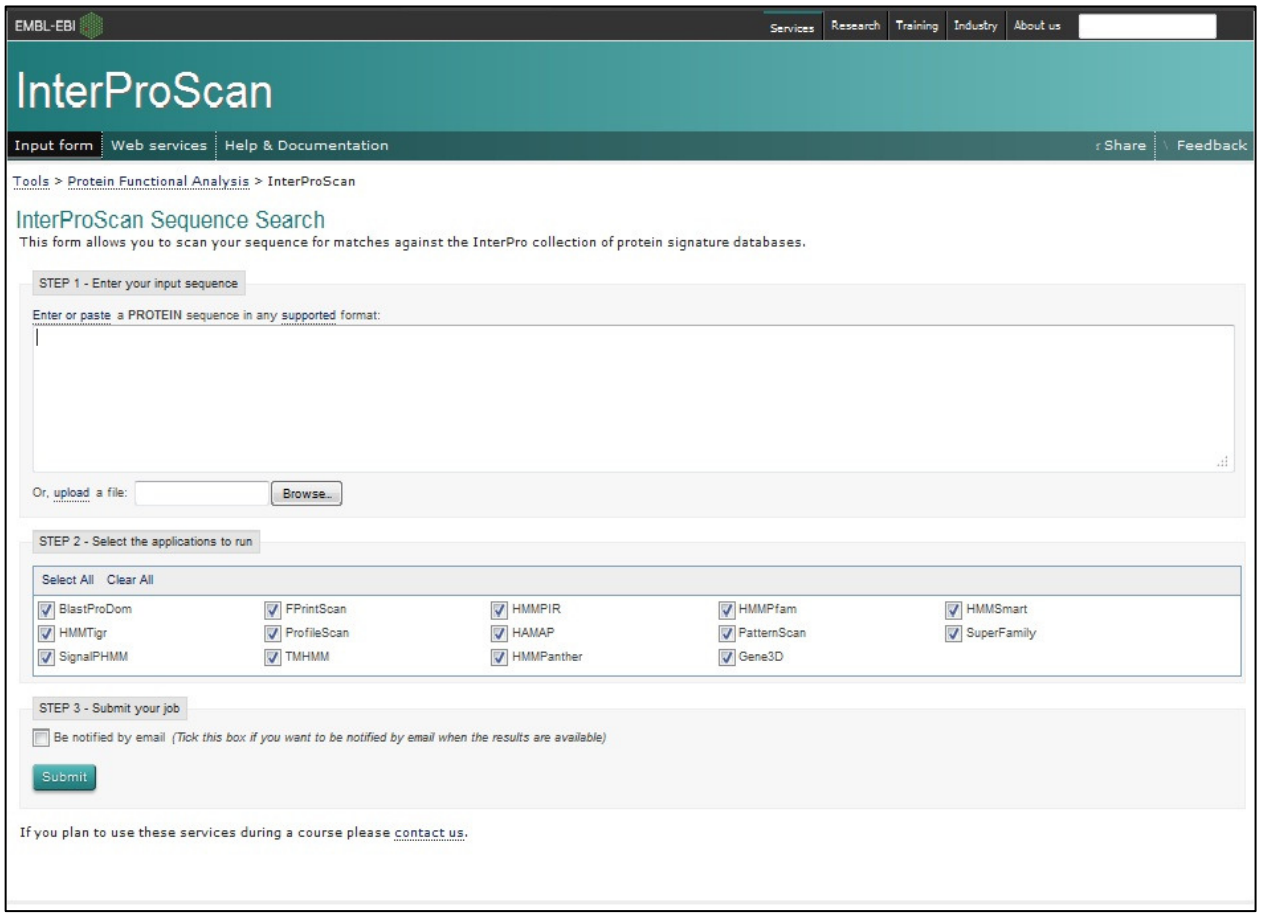


**Fig. 2.9:** Browsing view of InterProScan

### 2.2.9 Gene Ontology:

The Gene Ontology is a tool to standardizing the representation of gene and gene product attributes across species and databases. This datasets provides a controlled vocabulary of terms for describing gene characteristics and gene product annotation data from GO Consortium members, as well as tools to access and process this data. There are three separate aspects to this database: first, the development and maintenance of the ontologies themselves; second, the annotation of gene products, which entails making associations between the ontologies and the genes and gene products in the collaborating databases; and third, development of tools that facilitate the creation, maintenance and use of ontologies. The web link of this site is: http://www.geneontology.org/ Shown in Fig. 2.10.

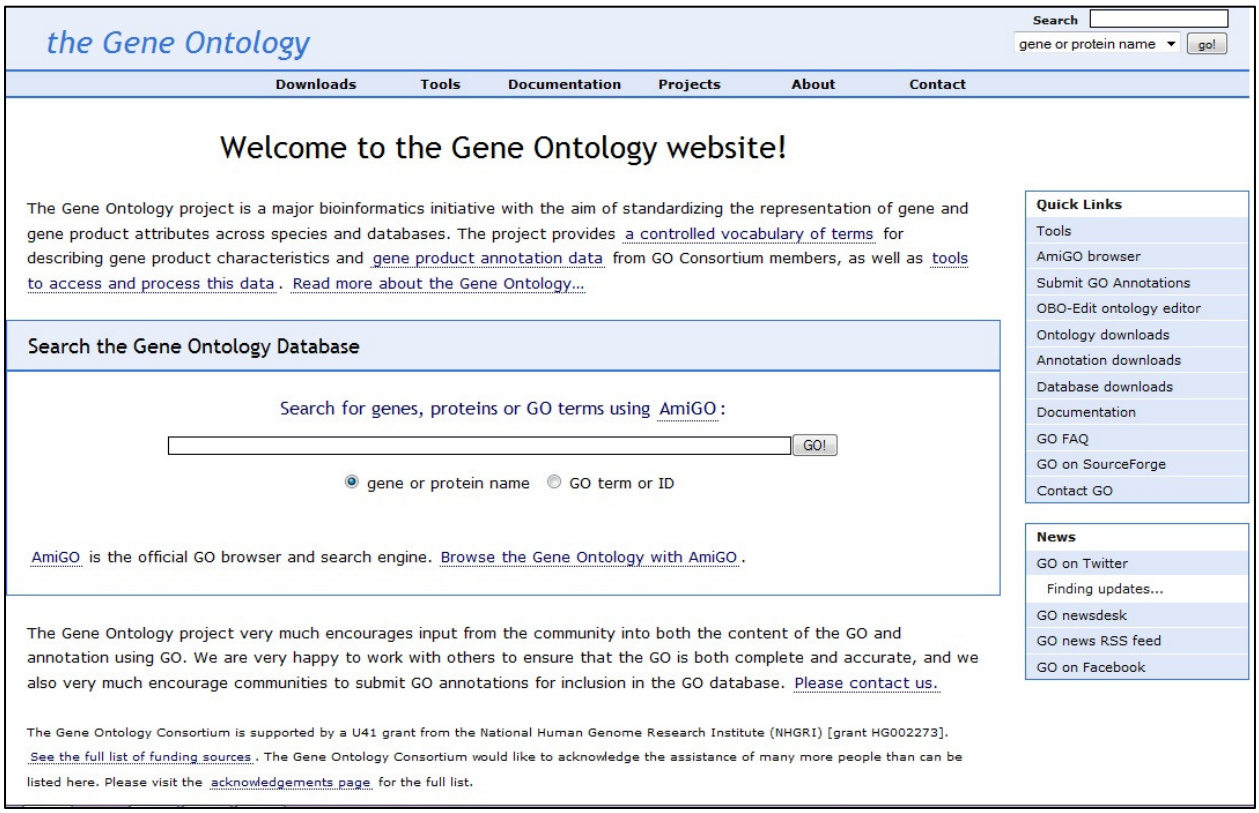


**Fig. 2.10:** Home page browsing view of Gene Ontology

### 2.2.10 GeneMANIA:

GeneMANIA finds out the relation between genes through analysis of functional association like protein and genetic interactions, pathways, co-expression, co-localization and protein domain similarity. It can be used to find new members in a pathway or complex; or find additional genes that may have been missed in any experimental screen; or find new genes with a specific function, such as, protein kinases. Here any types of gene related to any query can be defined by the set of genes which is inputted. If members of any set of genes make up a protein complex, then this database shows more potential members of the protein complex which are not directly connected to each other but somehow they are connected with the concerned gene set. If browser enters a gene list then GeneMANIA will display connections between that specific gene, within selected datasets. GeneMANIA is also accessible via a Cytoscape plugin, designed for power users. The web link for this site is: http://www.genemania.org/ shown in Fig. 2.11.


**Fig. 2.11:** Home page browsing view of GeneMANIA

# Results

## Results

### 3.1 Microarray Data:

Primarily four microarray datasets from ArrayExpress Database were taken. Each of the sets are associated with gene expression data under abiotic stresses like, Salt (Accession: *E-GEOD-33642*), Cold (Accession: *E-MEXP-3714*), Drought (Accession: *E-GEOD-42290*) and transcriptional responses to abscisic acid (Accession: *E-GEOD-45543*). During analyzing the whole dataset, it was found that a very good number of genes, enzymes and transcription factors and microRNA materials were up-regulated. Those up-regulated genes were taken into a single file to create an expression hub using Cytoscape software shown in (Fig. 3.1).

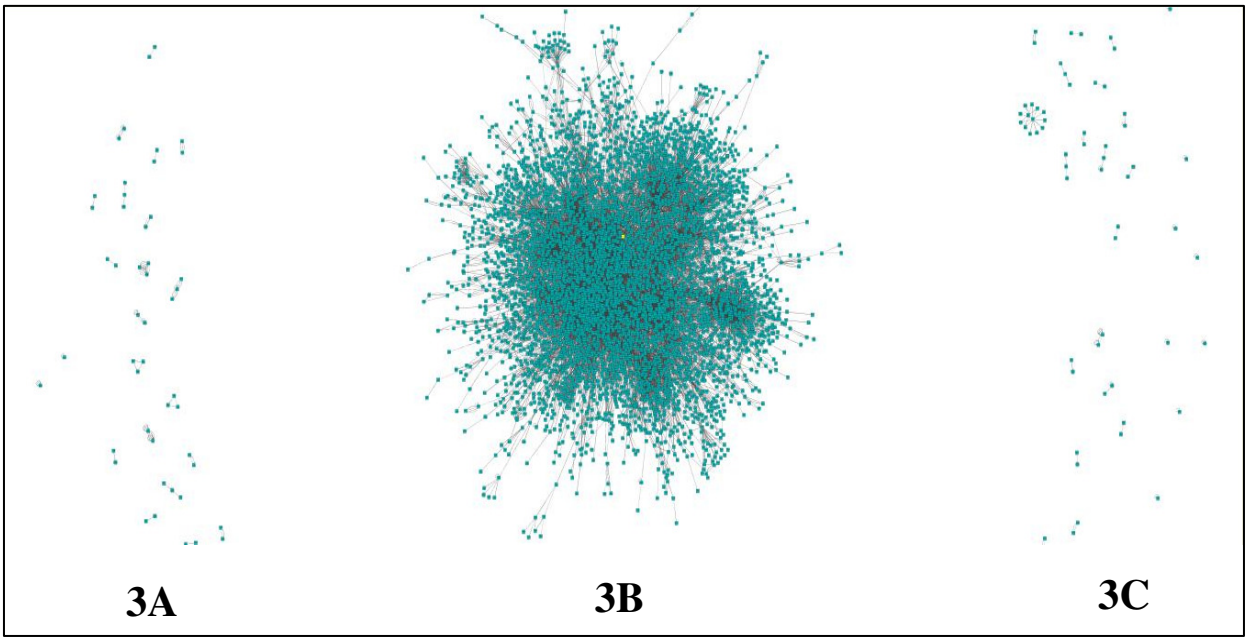


**Fig. 3.1:** A gene expression hub created using Cytoscape software after analyzing the microarray data. 3A and 3C represents distinctly related to the main expression hub 3B

Analysis of this expression hub (Fig. 3B) shows that there are several genes that get up-regulated in the four different stress signals. A number of genes from those were selected and they are as follows:

- Abscisic Acid: 643 Genes have been selected form the hub based on expression comparing with control
- Drought: 526 Genes have been selected
- Cold: 1023 Genes have been selected
- NaCl: 977 Genes have been selected

### 3.2 Common genes found in different stress signals:

The commonly expressing i.e. up-regulated genes found in different stress signals were then sorted by using the Van Diagram technique and using the software VENNY for a good graphical representation of the analyzed data (Fig. 3.2).

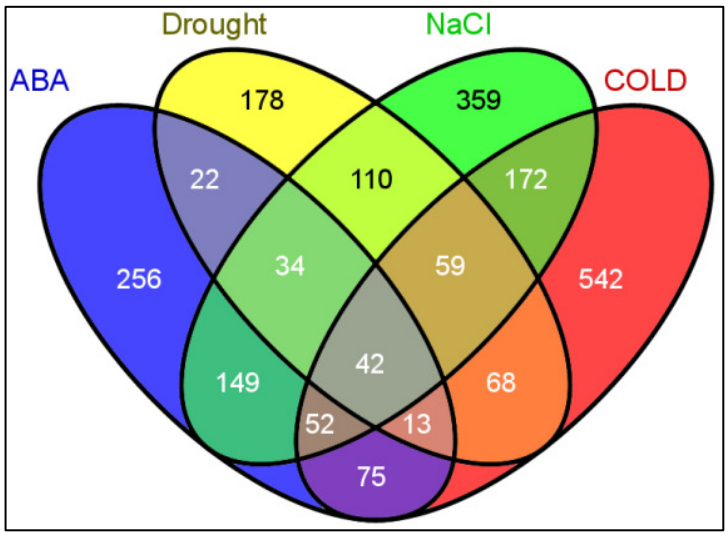


**Fig. 3.2:** A Van diagram to find the expressed genes in common in four different stress signals

From the Van Diagram, a total number of 42 genes were found that showed to have physiological characters to maintain the abiotic stress relating effect in plant system in *Arabidopsis thaliana*.

### 3.3 Commonly up-regulated (Salt, Abscisic Acid, Cold and Drought) forty two (42) Genes:

Descriptions of the forty two genes that found to be commonly up-regulated in the previous section are given bellow in Table 3.1.

**Table 3.1:** Commonly up-regulated genes, with their function and TAIR ID

| Sl. No | Gene Name | TAIR ID | Description of the Gene |
|---|---|---|---|
| 1. | DREB2A | AT5G05410.1 | Belongs to the DREB subfamily A-2 of ERF/AP2 transcription factor family (DREB2A) |
| 2. | ALA1 | AT5G04930.1 | Aminophospholipid translocase (p-type ATPase) involved in chilling response. |
| 3. | CIPK15 | AT5G01810.1 | CBL-interacting serine/threonine protein kinase, also has similarities to SOS2 kinase. |
| 4. | SYP61 | AT1G28490.1 | Encodes one of 24 Arabidopsis syntaxins. Its mRNA has been shown to be expressed. |
| 5. | P5CS1 | AT2G39800.1 | Encodes a delta1-pyrroline-5-carboxylate synthase that catalyzes the rate-limiting enzyme in the biosynthesis of proline. |
| 6. | NHX1 | AT5G27150.1 | Encodes a vacuolar sodium/proton antiporter involved in salt tolerance, ion homeostasis, and leaf development. |
| 7. | ATMPK6 | AT2G43790.1 | Encodes a MAP kinase induced by pathogens, ethylene biosynthesis, oxidative stress and osmotic stress. Also involved in ovule development. |
| 8. | LOS2 | AT2G36530.1 | Involved in light-dependent cold tolerance and encodes an enolase. |
| 9. | RAB18 | AT5G66400.1 | Belongs to the dehydrin protein family, which contains highly conserved stretches of 7-17 residues that are repetitively scattered in their sequences, the K-, S-, Y- and lysine rich segments. |
| 10. | ERD10 | AT1G20450.1 | Encodes a gene induced by low temperature and |

| Sl. No | Gene Name | TAIR ID | Description of the Gene |
|---|---|---|---|
| | | | dehydration. |
| 11. | FAD6 | AT4G30950.1 | Chloroplastic enzyme responsible for the synthesis of 16:2 and 18:2 fatty acids from galactolipids, sulpholipids and phosphatidylglycerol. |
| 12. | LOS1 | AT1G56070.1 | Encodes a translation elongation factor 2-like protein that is involved in cold-induced translation. Mutations in this gene specifically blocks low temperature-induced transcription of cold-responsive genes |
| 13. | HOS1 | AT2G39810.1 | A novel protein with a RING finger motif near the amino terminus. Negative regulator of cold responses. |
| 14. | CBF1 | AT4G25490.1 | Transcriptional activator that binds to the DRE/CRT regulatory element and induces COR (cold-regulated) gene expression increasing plant freezing tolerance. |
| 15. | FRO1 | AT5G67590.1 | Mutant leaves have a reduced capacity for cold acclimation, appear water-soaked, leak electrolytes, and accumulates reactive oxygen species constitutively. |
| 16. | FAD2 | AT3G12120.1 | Major enzyme responsible for the synthesis of 18:2 fatty acids in the endoplasmic reticulum. |
| 17. | SAL1 | AT5G63980.1 | Encodes a bifunctional protein that has 3'(2'),5'-bisphosphate nucleotidase and inositol polyphosphate 1-phosphatase activities and rescues sulfur assimilation mutants in yeast. It is involved in the response to cold, drought, and ABA. |
| 18. | CBF2 | AT4G25470.1 | Encodes a member of the DREB subfamily A-1 of ERF/AP2 transcription factor family (CBF2). The protein contains one AP2 domain. |
| 19. | CBL1 | AT4G17615.1 | Member of AtCBLs (Calcineurin B-like Calcium Sensor Proteins). Protein level is increased upon high salt, mannitol, and cold stresses. |

| Sl. No | Gene Name | TAIR ID | Description of the Gene |
|---|---|---|---|
| 20. | ICE1 | AT3G26744.1 | Encodes a MYC-like bHLH transcriptional activator that binds specifically to the MYC recognition sequences in the CBF3 promoter. |
| 21. | RD28 | AT2G37180.1 | A member of the plasma membrane intrinsic protein PIP2. Functions as aquaporin and is involved in dessication. |
| 22. | LTI30 | AT3G50970.1 | Belongs to the dehydrin protein family, which contains highly conserved stretches of 7-17 residues that are repetitively scattered in their sequences, the K-, S-, Y- and lysine rich segments. LTI29 and LTI30 double overexpressors confer freeze tolerance. |
| 23. | P5CS2 | AT3G55610.1 | Encodes delta 1-pyrroline-5-carboxylate synthetase B. Gene expression is induced by dehydration, high salt and ABA. |
| 24. | DREB2B | AT3G11020.1 | Encodes a member of the DREB subfamily A-2 of ERF/AP2 transcription factor family (DREB2B). |
| 25. | CBF4 | AT5G51990.1 | Encodes a member of the DREB subfamily A-1 of ERF/AP2 transcription factor family (CBF4). |
| 26. | COR15B | AT2G42530.1 | Similar to COR15A (COLD-REGULATED 15A) similar to cold response protein [Thellungiella salsuginea] |
| 27. | CIPK3 | AT2G26980.4 | Encodes a serine-threonine protein kinase whose expression increases in response to abscisic acid, cold, drought, high salt, and wounding conditions. |
| 28. | ERD14 | AT1G76180.1 | Encodes a dehydrin protein whose expression is induced early on in response to dehydration stress. |
| 29. | LOS4 | AT3G53110.1 | Encodes a putative DEAD-Box RNA Helicase and has RNA-dependent ATPase activity. Mutant is Sensitive to chilling stress and heat stress. |
| 30. | SOS3 | AT5G24270.1 | Encodes a calcium sensor that is essential for K+ nutrition, K+/Na+ selectivity, and salt tolerance. |

| Sl. No | Gene Name | TAIR ID | Description of the Gene |
|---|---|---|---|
| 31. | SFR2 | AT3G06510.2 | Encodes a protein with beta-glucosidase activity, mutants show increased sensitivity to freezing |
| 32. | SOS1 | AT2G01980.1 | Encodes a plasma membrane-localized Na+/H+ antiporter SOS1. Functions in the extrusion of toxic Na+ from cells and is essential for plant salt tolerance. |
| 33. | MEK1 | AT4G26070.2 | Member of MAP Kinase Kinase. Likely functions in a stress-activated MAPK pathway. Can phosphorylate the MAPK AtMPK4, in response to stress. |
| 34. | SOS2 | AT5G35410.1 | Encodes a member of the CBL-interacting protein kinase family, is a regulatory component controlling plant potassium nutrition |
| 35. | COR47 | AT1G20440.1 | Cold regulated gene, amino acid sequence homology with Group II LEA (late embryogenesis abundant) proteins. Also responds to osmotic stress, ABA, dehydration and inhibits E.coli growth while overexpressed. |
| 36. | COR15A | AT2G42540.2 | A cold-regulated gene whose product is targeted to the chloroplast and constitutive expression increases freezing tolerance in protoplasts in vitro and chloroplasts in vivo. |
| 37. | SOS5 | AT3G46550.1 | Isolated in a screen for salt hypersensitive mutants. Mutants have thinner cell walls, abnormal siliques and root growth is inhibited under salt stress. The gene has similarity to arabinogalactan proteins and domains associated with cell adhesion. |
| 38. | ATMPK4 | AT4G01370.1 | Encodes a nuclear and cytoplasmically localized MAP kinase involved in mediating responses to pathogens. |
| 39. | PFC1 | AT1G01860.1 | Dimethyladenosine transferase |
| 40. | CPL1 | AT4G21670.1 | Encodes a novel transcriptional repressor harboring two double-stranded RNA-binding domains and a region homologous to the catalytic domain of RNA polymerase II |

| Sl. No | Gene Name | TAIR ID | Description of the Gene |
|---|---|---|---|
| | | | C-terminal domain phosphatases found in yeast and in animals that regulate gene transcription. |
| 41. | ERD5 | AT3G30775.1 | Encodes a proline oxidase that is predicted to localize to the inner mitochondrial membrane, its mRNA expression induced by high levels of Al and by osmotic stress. |
| 42. | CPL2 | AT5G01270.2 | Encodes CPL2, a carboxyl-terminal domain (CTD) phosphatase that dephosphorylates CTD Ser5-PO4 of the RNA polymerase II complex. Regulates plant growth, stress and auxin responses. |

Through literature mining, from these 42 genes in Table: 3.1 it was found six genes as transcription factors, twenty two genes as regulatory genes and eight genes were found as enzymes and rests of the genes were showed as protein.

### 3.4 Protein-protein interaction study of the 42 up-regulated genes:

The 42 commonly up-regulated genes (Table 3.1) were then taken to further studies to see their interaction among themselves in terms of physical interaction, co-expression, literature mining etc. The interaction was visualized in String Database (version 9.02) and found that almost every up-regulated gene came in contact with each other and showed a strong co-relation. About 30 genes were directly connected while others remain distant in connection (Fig. 3.3). Directly connecting proteins were then brought together to see the interaction (Fig. 3.4) and observed a strong co-relation between transcription factors and antiporter genes and enzymes.

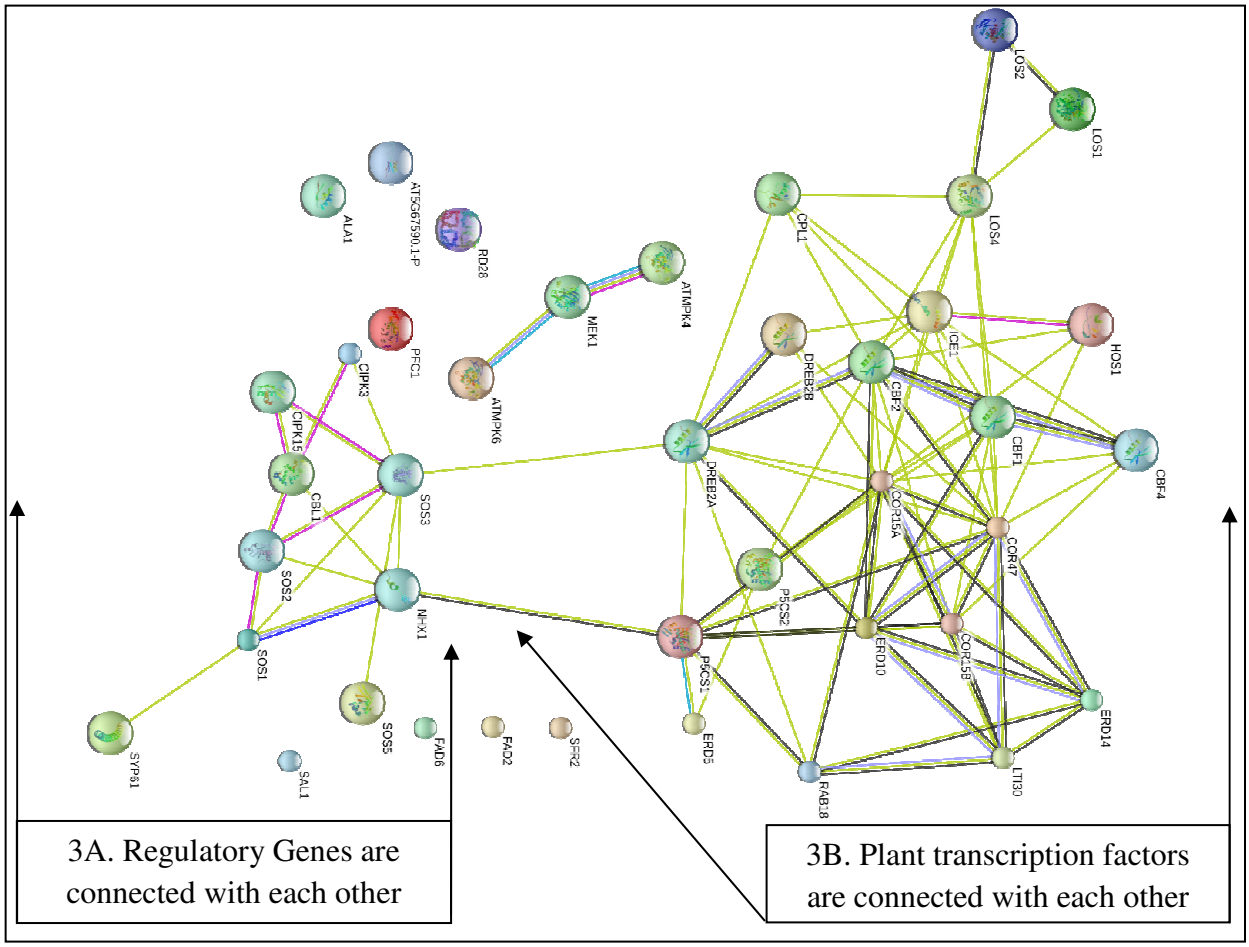


**Fig. 3.3:** Gene regulatory networks of plant transcription factors (TFs), enzymes and regulatory genes in plant abiotic stress responses and abscisic acid-dependent gene expression. (**3A**) Drought, salt, osmotic stress, temperature, and ABA stress factors modulating the level and activity of the Regulatory Genes and their target genes. (**3B**) The boxes represent an indication of TF proteins from the model plant *Arabidopsis* that are connecting with the major modulator of stress responsive genes like SOS1, NHX1 and their targets.

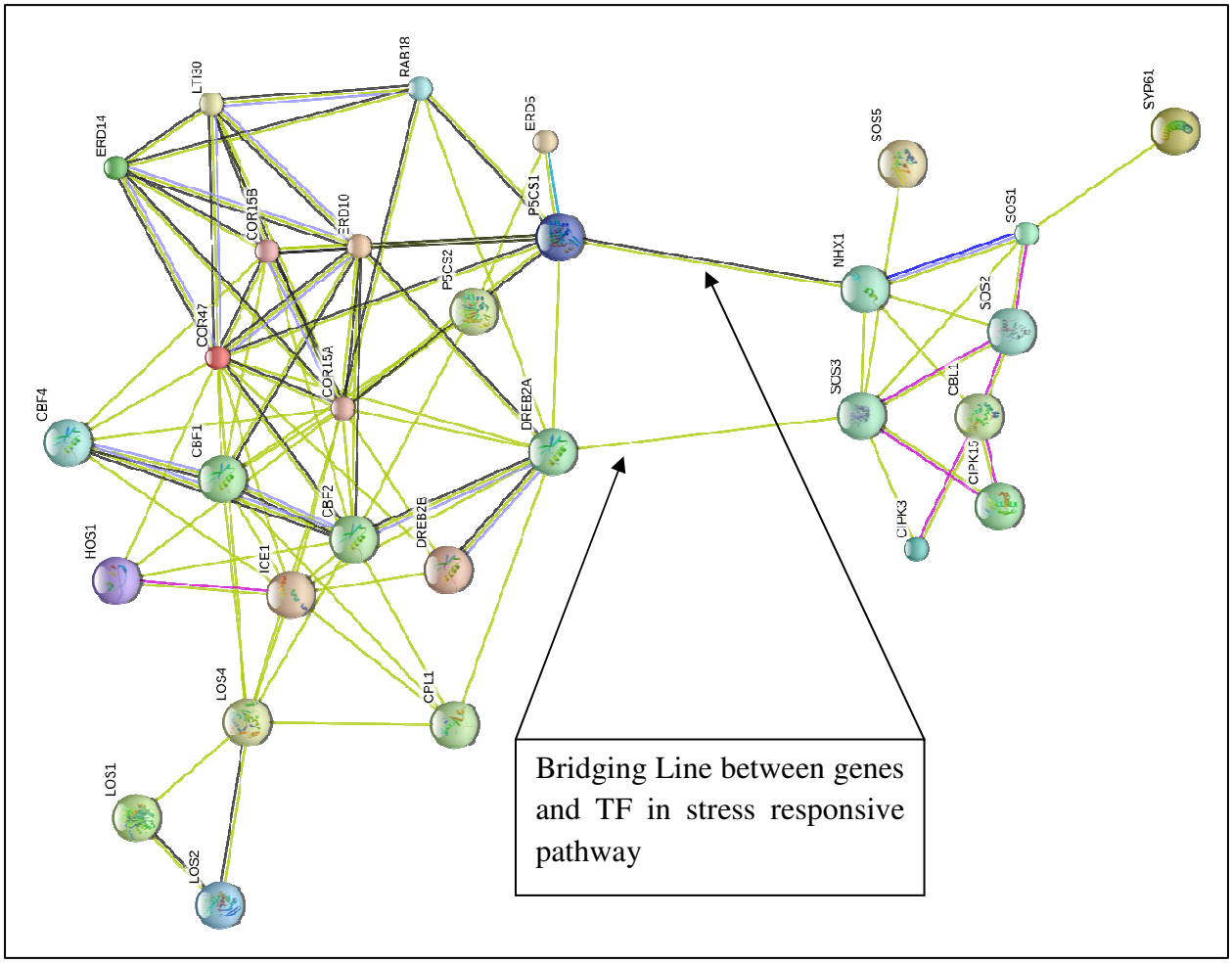


**Fig. 3.4:** A gene regulatory network of thirty commonly up-regulated genes in ABA dependent pathway, salinity stress, cold and drought stress. The bridging between transcription factor and stress responsive genes is clearly indicating their co-relation in this figure.

### 3.5 Connectome genes and transcription factor:

In the light of above result eight genes having strongly co-related connectomes and reported to have crucial role in abiotic stress tolerance were short listed (Table 3.2) for further studies.

**Table 3.2**: Selected genes that have co-related connectomes and their crucial role in abiotic stress tolerance:

| Sl. No. | Gene Name | Characters |
|---|---|---|
| 01. | DREB2A (Dehydration-Responsive Element-Binding Protein 2A) | Transcription Factor |
| 02. | P5CS1 (Delta1-Pyrroline-5-Carboxylate Synthase 1) | Enzyme |
| 03. | CPL1 (C-Terminal Domain Phosphatase-Like 1) | Transcription Factor |
| 04. | ERD5 (Early Responsive to Dehydration 5) | Transcription Factor |
| 05. | NHX1 (Na+/H+ Exchanger) | Vacuolar Antiporter |
| 06. | SOS1 (Salt Overly Sensitive 1) | Plasma Membrane Antiporter |
| 07. | SOS2 (Salt Overly Sensitive 2) | Protein Kinase |
| 08. | SOS3 (Salt Overly Sensitive 3) | Calcium-dependent Protein Serine |

In the next session of the result, all possible characters of the targeted eight molecules was revealed depending on their amino acid, protein domains, individual interactomes and gene ontology to get the whole pictorial view of the genes in three different sectors of life system Biological, Molecular, and Cellular, respectively. Available free tools mentioned in Material and Method section have been extensively applied to get the results to make individual interpretation.

### 3.5.1 DREB2A (Dehydration-Responsive Element-Binding Protein 2A):

#### 3.5.1.1 Amino acid sequence of DREB2A:

DREB2A transcription factor consists of 335 amino acid chain and the sequence has been downloaded from NCBI database for further analysis. The sequence is given below:

MAVYDQSGDRNRTQIDTSRKRKSRSRGDGTTVAERLKRWKEYNETVEEVSTKKRKVPAKGSKKGCMKGKGGPENSR
CSFRGVRQRIWGKWVAEIREPNRGSRLWLGTFPTAQEAASAYDEAAKAMYGPLARLNFPRSDASEVTSTSSQSEVCTV
ETPGCVHVKTEDPDCESKPFSGGVEPMYCLENGAEEMKRGVKADKHWLSEFEHNYWSDILKEKEKQKEQGIVETCQQ
QQQDSLSVADYGWPNDVDQSHLDSSDMFDVDELLRDLNGDDVFAGLNQDRYPGNSVANGSYRPESQQSGFDPLQSLN
YGIPPFQLEGKDGNGFFDDLSYLDLEN

#### 3.5.1.2 Blast hit of DREB2A:

Amino acid sequences were blasted in NCBI Blast P-suite to check conserved domains and sequence similarities among other plant species. From this analysis, it was revealed that DREB2A does not share common sequence except with the AP2 superfamily domain showed in (Fig. 3.5). The closest homology with DREB2A was found in *Arachis hypogaea* (Query coverage: 99% and sequence similarities: 100%), *Eutrema salsugineum* (Query coverage: 71% and sequence similarities: 99%) (Fig. 3.6).

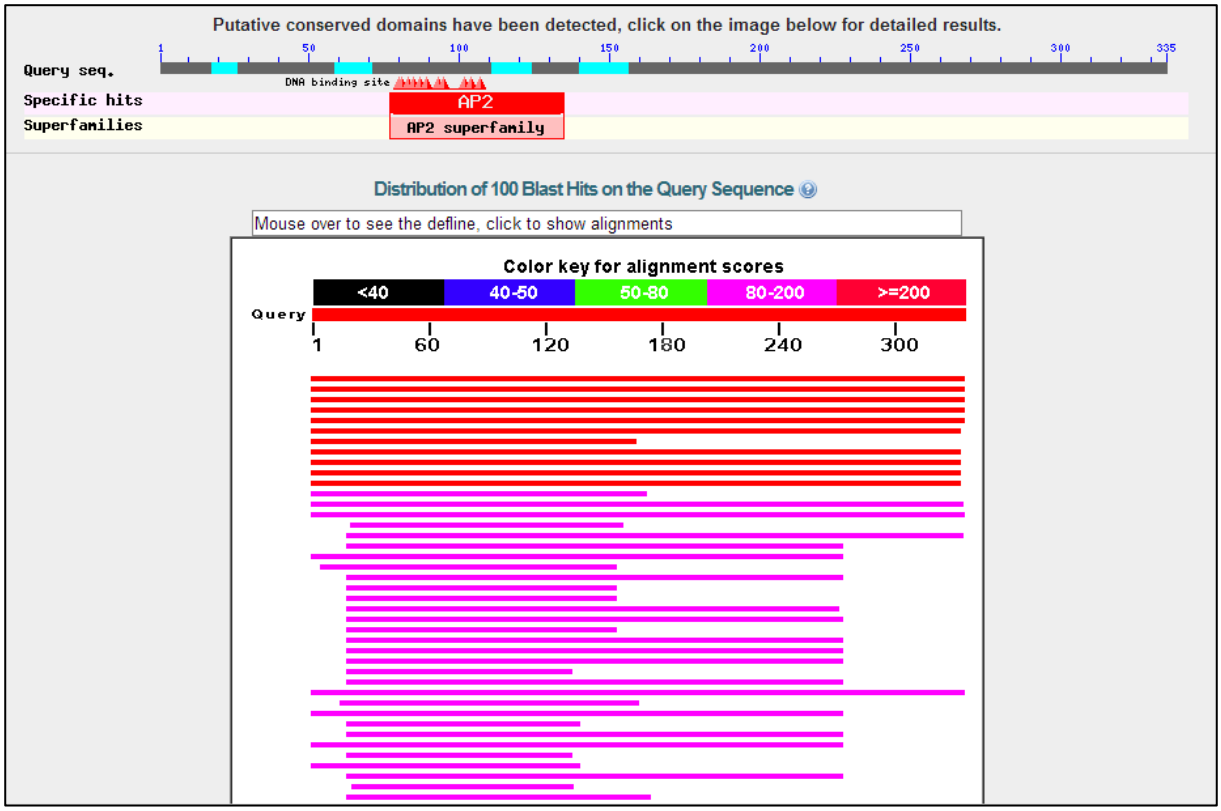


**Fig. 3.5:** Conserved domain sequence similarities of DREB2A with *Arachis hypogaea* plant species from NCBI Blast P-suite.


**Fig. 3.6:** Closest homology of DREB2A through NCBI P-Suite

### 3.5.1.3 Search protein domain of DREB2A by Interproscan:

Interproscan tool from European Bioinformatics Institute (EBI) was used to find out characterized domains present in DREB2A. It was revealed that, three conserved domains were available in the sequence of DREB2A, namely,

- AP2/ ERF domain
- DNA binding domain, integrase type and
- Another domain found as unintegrated (Fig. 3.7)

These domains are conserved and play a crucial role during stress to up-regulate the target genes.

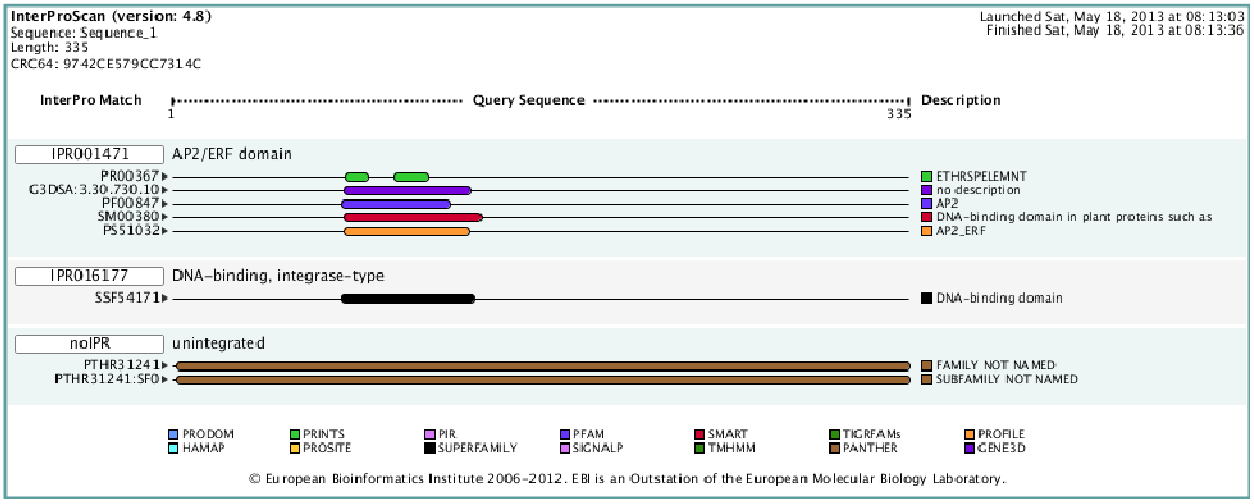


**Fig. 3.7:** Characterized domains present in DREB2A from Interproscan tool

### 3.5.1.4 GO function annotation of DREB2A:

For function annotation GO file was generated. DREB2A transcription factor has major role in response to water deprivation, temperature stimulus, and various stresses. GO file denotes its all type of biological processes and the comprised diagram showed in (Fig. 3.8) which depicts all major functions of DREB2A. It mainly works on oxidative stress response and helps to up-regulate some major genes relating abiotic stress. It also acts as an abiotic stress stimuli.

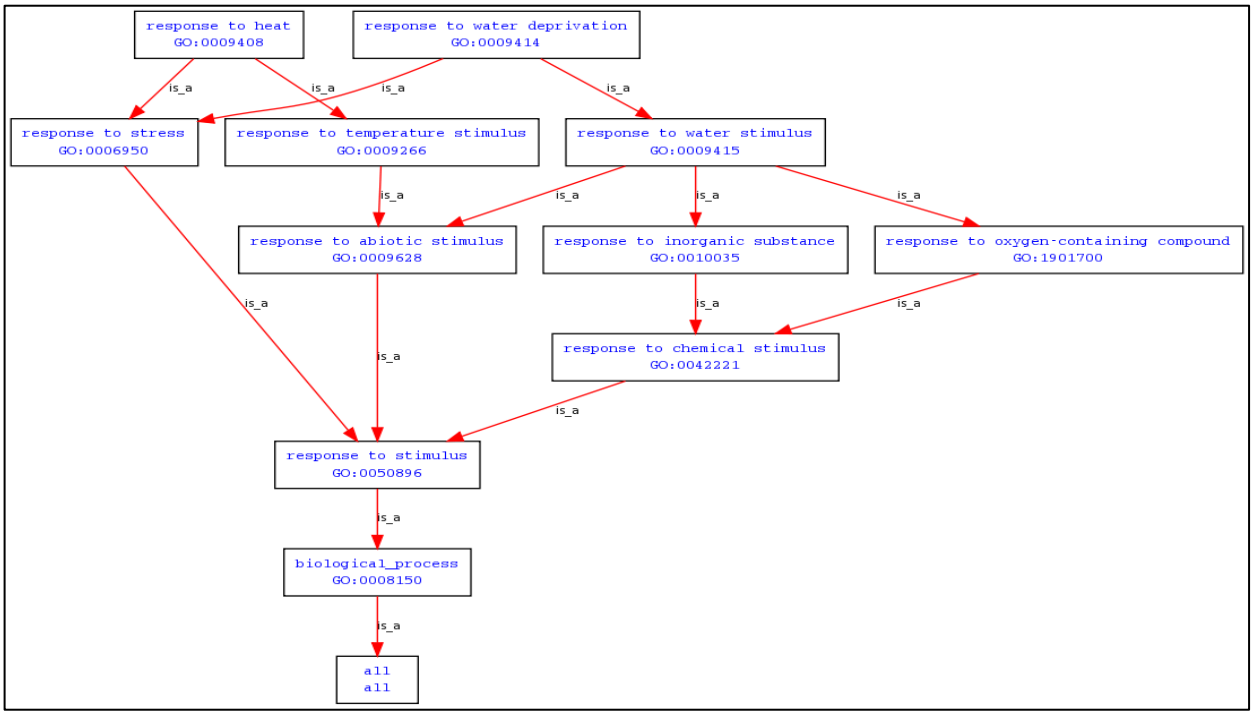


**Fig. 3.8:** Diverse functions of DREB2A gene reveled from Gene Ontology annotation

### 3.5.1.5 DREB2A Interactome:

GeneMania database was used to show the molecular interaction partner of DREB2A TF in *Arabidopsis* and found a strong binding affinity with DRIP1, DRIP2, RPL15, NFD3, KUP6 etc. which are also engaged in stress responsive mechanism showed in (Fig. 3.9) depicted the major connection of DREB2A TF.

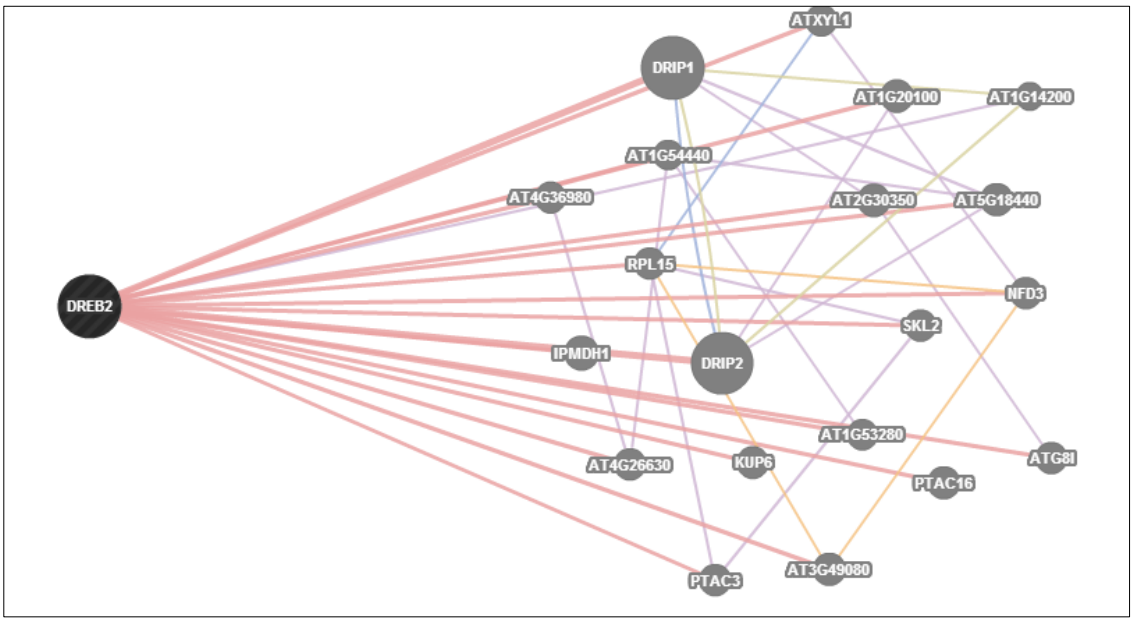


**Fig. 3.9:** *Arabidopsis* interaction networks with different abiotic stress-related TFs as hub. The interactomes are visualized from the GENEmania-Browse viewer

### 3.5.2 P5CS1 (Delta1-Pyrroline-5-Carboxylate Synthase 1):

#### 3.5.2.1 Amino acid sequence of P5CS1:

P5CS1 amino acid sequences were retrieved from NCBI database and the sequence length is 717 amino acid chain long, which encodes number of stress responsive domain. The retrieved sequence chain is given below:

MEELDRSRAFARDVKRIVVKVGTAVVTGKGGRLALGRLGALCEQLAELNSDGFEVILVSSGAVGLGRQRLRYRQLVNS
SFADLQKPQTELDGKACAGVGQSSLMAYYETMFDQLDVTAAQLLVNDSSFRDKDFRKQLNETVKSMLDLRVIPIFNEN
DAISTRRAPYQDSSGIFWDNDSLAALLALELKADLLILLSDVEGLYTGPPSDPNSKLIHTFVKEKHQDEITFGDKSRLGRG
GMTAKVKAAVNAAYAGIPVIITSGYSAENIDKVLRGLRVGTLFHQDARLWAPITDSNARDMAVAARESSRKLQALSSE
DRKKILLDIADALEANVTTIKAENELDVASAQEAGLEESMVARLVMTPGKISSLAASVRKLADMEDPIGRVLKKTEVAD
GLVLEKTSSPLGVLLIVFESRPDALVQIASLAIRSGNGLLLKGGKEARRSNAILHKVITDAIPETVGGKLIGLVTSREEIPDL
LKLDDVIDLVIPRGSNKLVTQIKNTTKIPVLGHADGICHVYVDKACDTDMAKRIVSDAKLDYPAACNAMETLLVHKDL
EQNAVLNELIFALQSNGVTLYGGPRASKILNIPEARSFNHEYCAKACTVEVVEDVYGAIDHIHRHGSAHTDCIVTEDHEV
AELFLRQVDSAAVFHNASTRFSDGFRFGLGAEVGVSTGRIHARGPVGVEGLLTTRWIMRGKGQVVDGDNGIVYTHQDI
PIQA

### 3.5.2.2 Blast hit of P5CS1:

Amino acid sequences were then blasted in NCBI Blast P-suite to check conserved domains and sequence similarities among other species. From the blast analysis, it was revealed that P5CS1 found highly conserved sequence similarities between plant species showed in (Fig. 3.10). The closest homology with P5CS1 was found in *Brassica napus* (Query coverage: 100% and Sequence similarity: 97%) and *Arabis stellen* (Query coverage 85% and sequence similarity 95%) (Fig. 3.11).

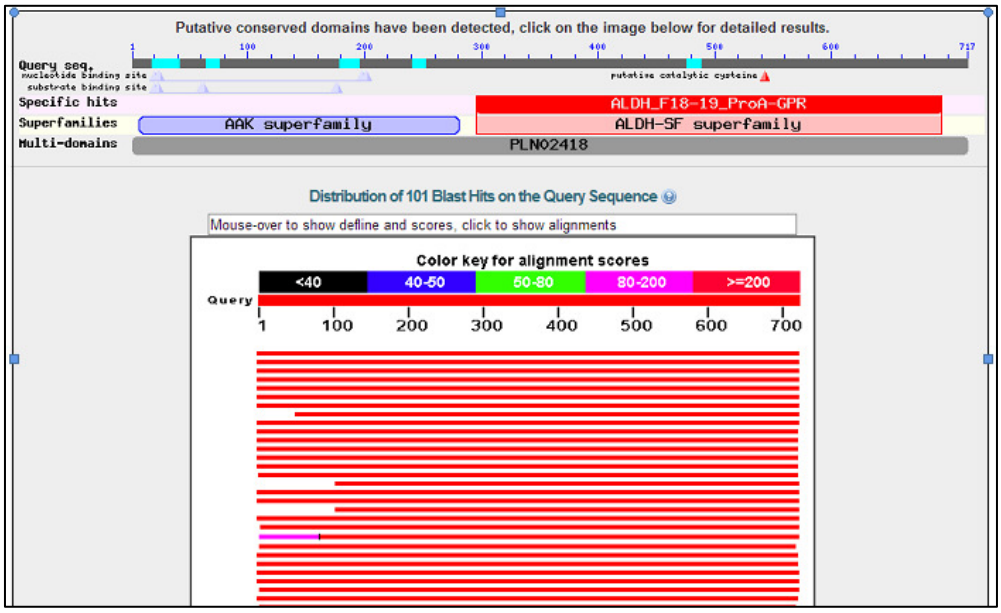


**Fig. 3.10:** Conserved domain sequence similarities of P5CS1 with *Brassica napus* from NCBI Blast P-suite.

| | Description | Max score | Total score | Query cover | E value | Ident | Accession |
|---|---|---|---|---|---|---|---|
| ☐ | delta1-pyrroline-5-carboxylate synthase 1 [Arabidopsis thaliana] >ref\|NP_001189714.1\| delta1-pyrroline-5-carboxylate synthase 1 [Arabidopsis thaliana] >: | 1459 | 1459 | 100% | 0.0 | 100% | NP_181510.1 |
| ☐ | AT2G39800 [Arabidopsis thaliana] | 1363 | 1363 | 93% | 0.0 | 100% | BAH20244.1 |
| ☐ | delta1-pyrroline-5-carboxylate synthase 1 [Arabidopsis thaliana] >dbj\|BAH19477.1\| AT2G39800 [Arabidopsis thaliana] >gb\|AEC09728.1\| delta1-pyrroline-5 | 1256 | 1256 | 85% | 0.0 | 100% | NP_973641.1 |
| ☐ | putative delta-1-pyrroline 5-carboxylase synthetase P5C1 [Arabidopsis thaliana] | 1458 | 1458 | 100% | 0.0 | 99% | AAL87255.1 |
| ☐ | delta1-pyrroline-5-carboxylate synthase 1 [Arabidopsis thaliana] >gb\|AEC09731.1\| delta1-pyrroline-5-carboxylate synthase 1 [Arabidopsis thaliana] | 1395 | 1395 | 100% | 0.0 | 99% | NP_001189715.1 |
| ☐ | delta1-pyrroline-5-carboxylate synthetase [Arabidopsis thaliana] | 1450 | 1450 | 100% | 0.0 | 99% | BAA06864.1 |
| ☐ | delta1-pyrroline-5-carboxylate synthase 1 [Arabidopsis lyrata subsp. lyrata] >gb\|EFH57938.1\| delta1-pyrroline-5-carboxylate synthase 1 [Arabidopsis lyrata | 1433 | 1433 | 100% | 0.0 | 98% | XP_002881679.1 |
| ☐ | hypothetical protein CARUB_v10022685mg, partial [Capsella rubella] >gb\|EOA26262.1\| hypothetical protein CARUB_v10022685mg, partial [Capsella rub | 1378 | 1378 | 100% | 0.0 | 98% | XP_006293364.1 |
| ☑ | delta 1-pyrroline-5-carboxylate synthetase A [Brassica napus] | 1368 | 1368 | 100% | 0.0 | 97% | AAK01360.1 |
| ☑ | delta-1-pyrroline 5-carboxylase synthetase [Arabis stelleri] | 1158 | 1158 | 85% | 0.0 | 95% | ADG08111.1 |

**Fig. 3.11:** Closest homology of P5CS1 through NCBI P-Suite

### 3.5.2.3 Search of protein domain of P5CS1 by Interproscan:

Interproscan tool from European Bioinformatics Institute (EBI) was used to find out characterized domains present in P5CS1. It was revealed that, twelve conserved domains are available in the sequence of P5CS1, namely, (Fig. 3.12).

- Gamma-glutamyl phosphate reductase GPR
- Aspartate/glutamate/uridylate kinase
- Glutamate/acetylglutamate kinase
- Glutamate 5-kinase/delta-1-pyroline-5-carboxylate synthase
- Delta I-pyroline-5- carboxylate synthase
- Aldehyde dehyrogenase domain
- Aldehyde/histidinol dehydrogenase
- Aldehyde dehydrogenase, N-terminal
- Aldehyde dehydrogenase, C-terminal
- Glutamte 5-kinase, conserved site
- Gamma-glutamyl phosphate reductase GPR, conserved site, and
- Unintregated

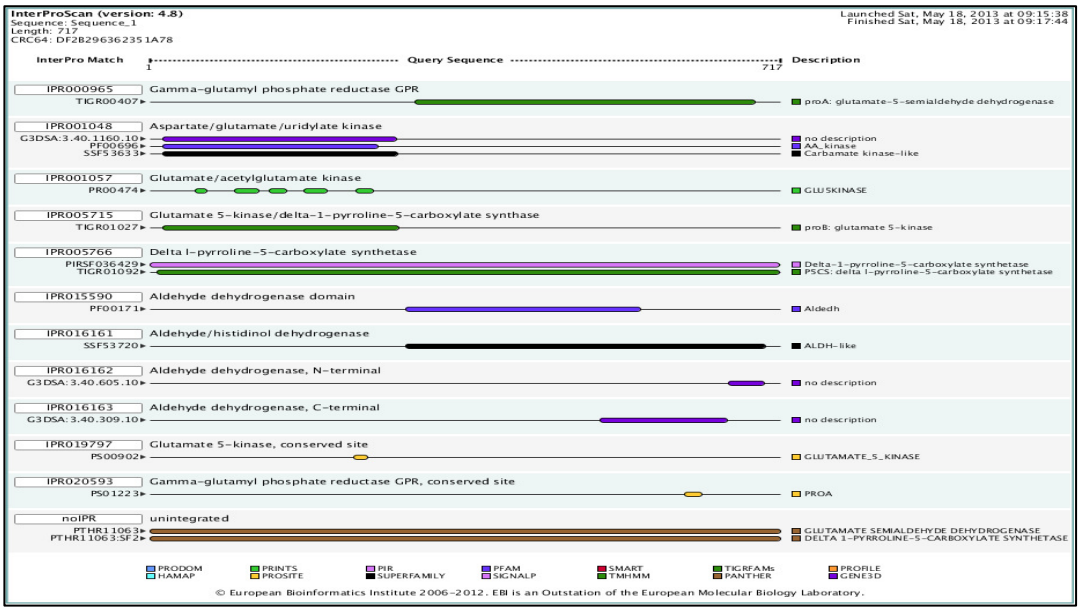


**Fig. 3.12:** Characterized domains present in P5CS1 through Interproscan tool

### 3.5.2.4 GO function annotation of P5CS1:

GO function of P5CS1 showed that it has strong biological functions that play a significant role during salt stress, water deprivation, osmotic stress, chemical stimulus etc. The GO annoted graphics (Fig. 3.13) is provided. It functions in drought response, salt stress by altering its target genes expressions.

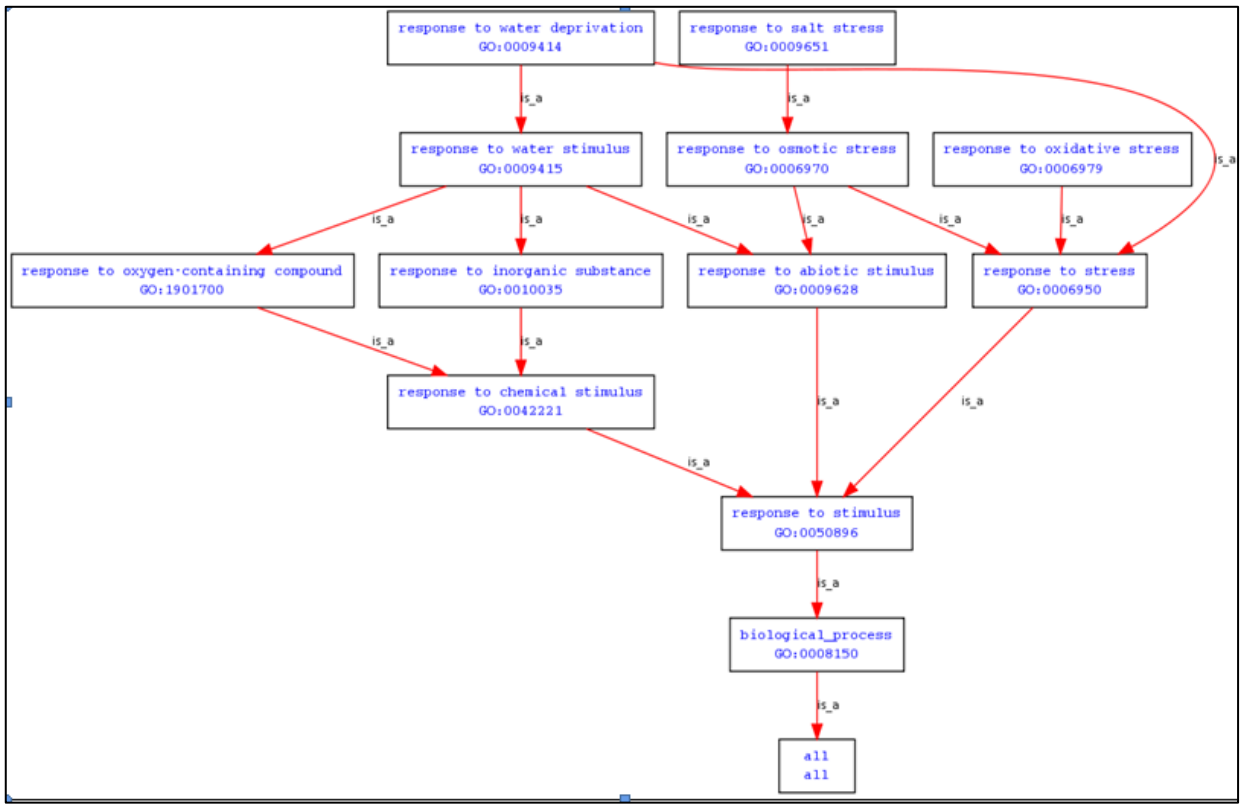


**Fig. 3.13:** Diverse functions of P5CS1 gene reveled from Gene Ontology annotation

### 3.5.2.5 P5CS1 Interactome:

The interaction partner P5CS1 interacts with NAGK, EMB2772, ALDH2C4, ALDH2, ALDH3 etc. which are the very common stress responsive factor in Arabidopsis. The (Fig. 3.14) below shows the pictorial view of the interaction.

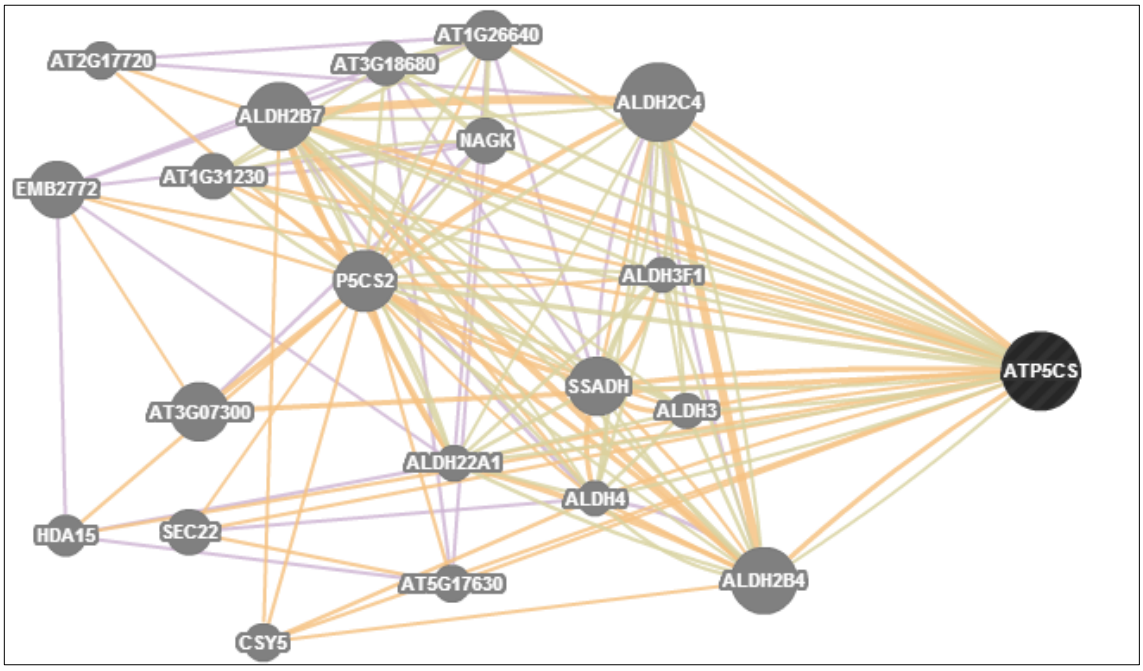


**Fig. 3.14:** Interacting molecule with P5CS1 revealed from GeneMania Database

### 3.5.3 CPL1 (C-Terminal Domain Phosphatase-Like 1):

#### 3.5.3.1 Amino acid sequence of CPL1:

CPL1 amino acid sequence was downloaded from NCBI database which is about 967 amino acid chain long. The sequence has been used to find out protein domain structures. The amino acid sequence is presented below:

MYSNNRVEVFHGDGRLGELEIYPSRELNQQQDDVMKQRKKKQREVMELAKMGIRISHFSQSGERCPPLAILTTISSCGL
CFKLEASPSPAQESLSLFYSSCLRDNKTAVMLLGGEELHLVAMYSENIKNDRPCFWAFSVAPGIYDSCLVMLNLRCLGI
VFDLDETLVVANTMRSFEDKIDGFQRRINNEMDPQRLAVIVAEMKRYQDDKNLLKQYIESDQVVENGEVIKVQSEIVPA
LSDNHQPLVRPLIRLQEKNIILTRINPMIRDTSVLVRMRPSWEELRSYLTAKGRKRFEVYVCTMAERDYALEMWRLLDP
EGNLINTNDLLARIVCVKSGFKKSLFNVFLDGTCHPKMALVIDDRLKVWDEKDQPRVHVVPAFAPYYSPQAEAAATPV
LCVARNVACGVRGGFFRDFDDSLLPRIAEISYENDAEDIPSPPDVSHYLVSEDDTSGLNGNKDPLSFDGMADTEVERRL
KEAISASSAVLPAANIDPRIAAPVQFPMASASSVSVPVPVQVVQQAIQPSAMAFPSIPFQQPQQPTSIAKHLVPSEPSLQSS
PAREEGEVPESELDPDTRRRLLILQHGQDTRDPAPSEPSFPQRPPVQAPPSHVQSRNGWFPVEEEMDPAQIRRAVSKEYP
LDSEMIHMEKHRPRHPSFFSKIDNSTQSDRMLHENRRPPKESLRRDEQLRSNNNLPDSHPFYGEDASWNQSSSRNSDLDF
LPERSVSATETSADVLHGIAIKCGAKVEYKPSLVSSTDLRFSVEAWLSNQKIGEGIGKSRREALHKAAEASIQNLADGYM
RANGDPGPSHRDATPFTNENISMGNANALNNQPFARDETALPVSSRPTDPRLEGSMRHTGSITALRELCASEGLEMAFQ
SQRQLPSDMVHRDELHAQVEIDGRVVGEGVGSTWDEARMQAAERALSSVRSMLGQPLHKRQGSPRSFGGMSNKRLKP
DFQRSLQRMPSSGRYS

### 3.5.3.2 Blast hit of CPL1:

Amino acid sequences were then blasted in NCBI Blast P-suit to check conserved domains and sequence similarities with other plant species. From the blast analysis, it was revealed that the CPL1 is moderately common in sequence among plant species. The blasted graphical view is represented below (Fig.15). The closest homology with CPL1 was found in *Theobroma cacao* (Query coverage: 99% and Sequence similarity: 63%) (Fig. 3.16).

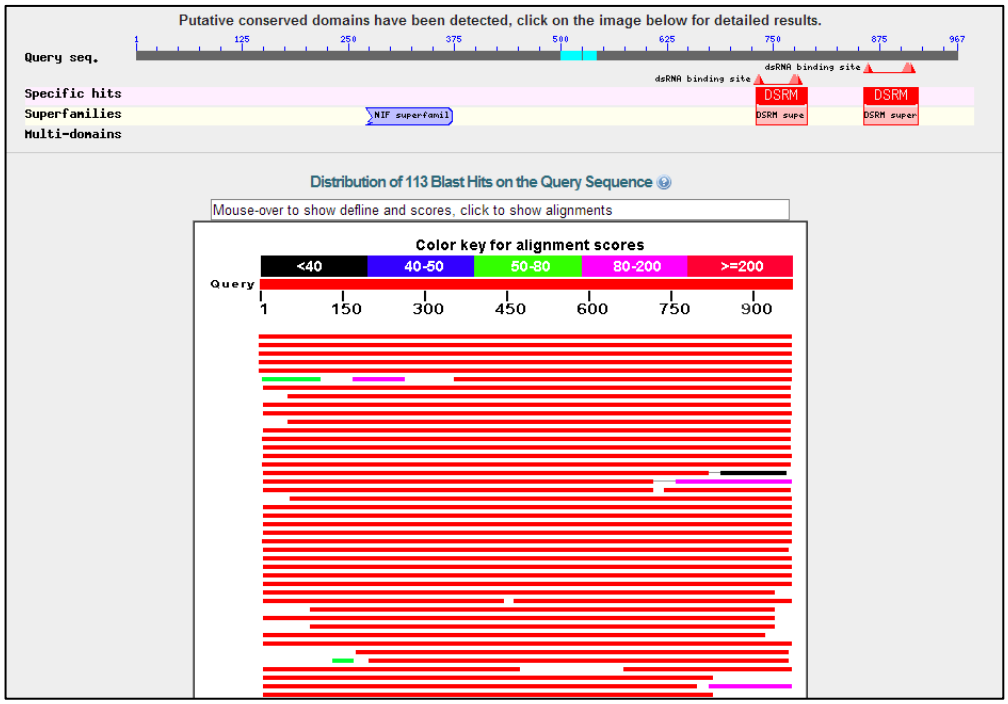


**Fig. 3.15:** Conserved domain sequence similarities of CPL1 with *Theobroma cacao* from NCBI Blast P-suite

| Description | Max score | Total score | Query cover | E value | Ident | Accession |
|---|---|---|---|---|---|---|
| RNA polymerase II C-terminal domain phosphatase-like 1 [Arabidopsis thaliana] >sp\|Q5YDB6.1\|CPL1_ARATH RecName: Full=RNA polymerase II C-te | 1997 | 1997 | 100% | 0.0 | 100% | NP_193898.3 |
| hypothetical protein [Arabidopsis thaliana] | 1994 | 1994 | 100% | 0.0 | 99% | BAF01152.1 |
| putative protein [Arabidopsis thaliana] >emb\|CAB81274.1\| putative protein [Arabidopsis thaliana] | 1931 | 1931 | 100% | 0.0 | 97% | CAB36811.1 |
| hypothetical protein ARALYDRAFT_492708 [Arabidopsis lyrata subsp. lyrata] >gb\|EFH46132.1\| hypothetical protein ARALYDRAFT_492708 [Arabidopsis | 1844 | 1844 | 100% | 0.0 | 93% | XP_002869873.1 |
| hypothetical protein CARUB_v10004071mg [Capsella rubella] >gb\|EOA15976.1\| hypothetical protein CARUB_v10004071mg [Capsella rubella] | 1737 | 1737 | 100% | 0.0 | 89% | XP_006283078.1 |
| hypothetical protein EUTSA_v10024324mg [Eutrema salsugineum] >gb\|ESQ55202.1\| hypothetical protein EUTSA_v10024324mg [Eutrema salsugineu | 1716 | 1716 | 100% | 0.0 | 88% | XP_006413749.1 |
| hypothetical protein EUTSA_v10024324mg [Eutrema salsugineum] >gb\|ESQ55201.1\| hypothetical protein EUTSA_v10024324mg [Eutrema salsugineu | 1696 | 1696 | 100% | 0.0 | 87% | XP_006413748.1 |
| putative protein [Arabidopsis thaliana] | 1255 | 1255 | 63% | 0.0 | 99% | BAD94401.1 |
| hypothetical protein CICLE_v10014168mg [Citrus clementina] | 1196 | 1196 | 98% | 0.0 | 64% | ESR62542.1 |
| C-terminal domain phosphatase-like 1 isoform 1 [Theobroma cacao] | 1194 | 1194 | 99% | 0.0 | 63% | EOY28302.1 |

**Fig. 3.16:** Closest homology of CPL1 through NCBI P-Suite

### 3.5.3.3 Intreproscan domain search for CPL1:

Interproscan tool from European Bioinformatics Institute (EBI) was used to find out characterized domains present in CPL1. It was revealed that five conserved domains (Fig. 3.16) were available in the sequence of CPL1, namely,

- Double-stranded RNA-binding
- NLI interacting factor
- Double stranded RNA-binding-like domain
- HAD-like domain and
- Unintegrated

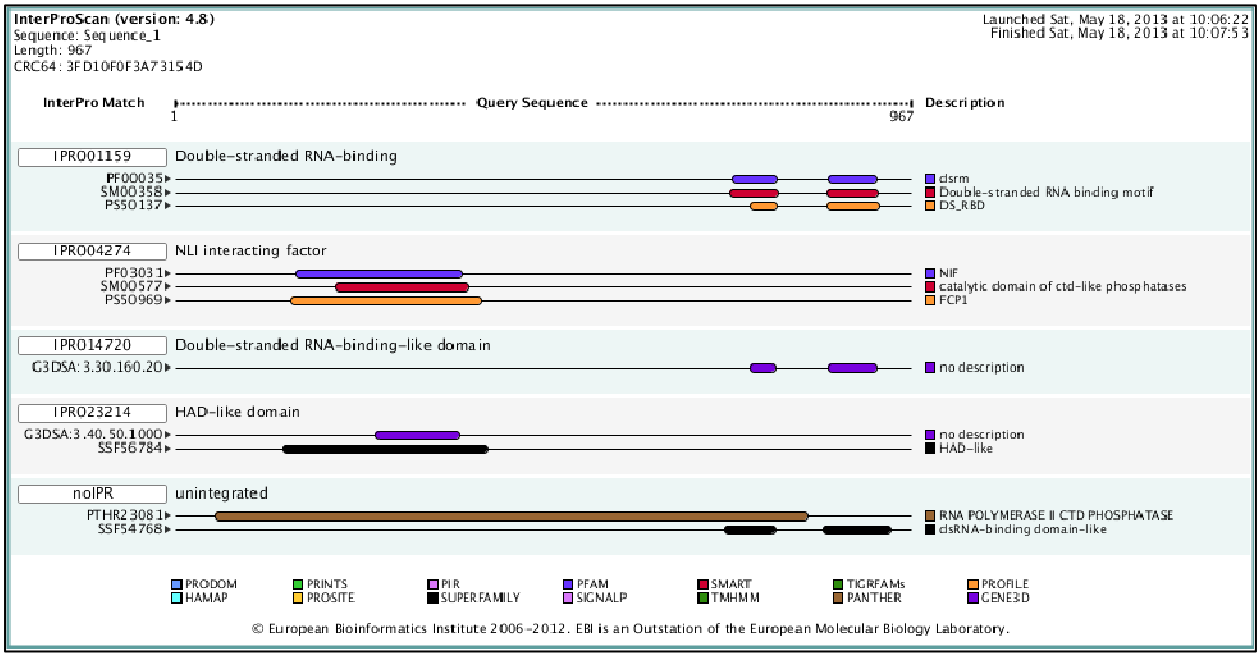


**Fig. 3.17:** InterProScan result depicted different protein domain in CPL1

### 3.5.3.4 GO function annotation of CPL1:

CPL1 function was revealed by doing Gene Ontology search and the below there is VizMap presentation of the query output (Fig. 3.18). The output reveals that CPL1 carries significant importance in phosphatase activity, response to abiotic stress and hydrolase activity. The annotation shows the different mechanisms of the CPL1 gene (Fig. 3.18). CPL1 has diverse

functions including salinity stress tolerance, phosphates activity and even in biotic stress like wounding etc.

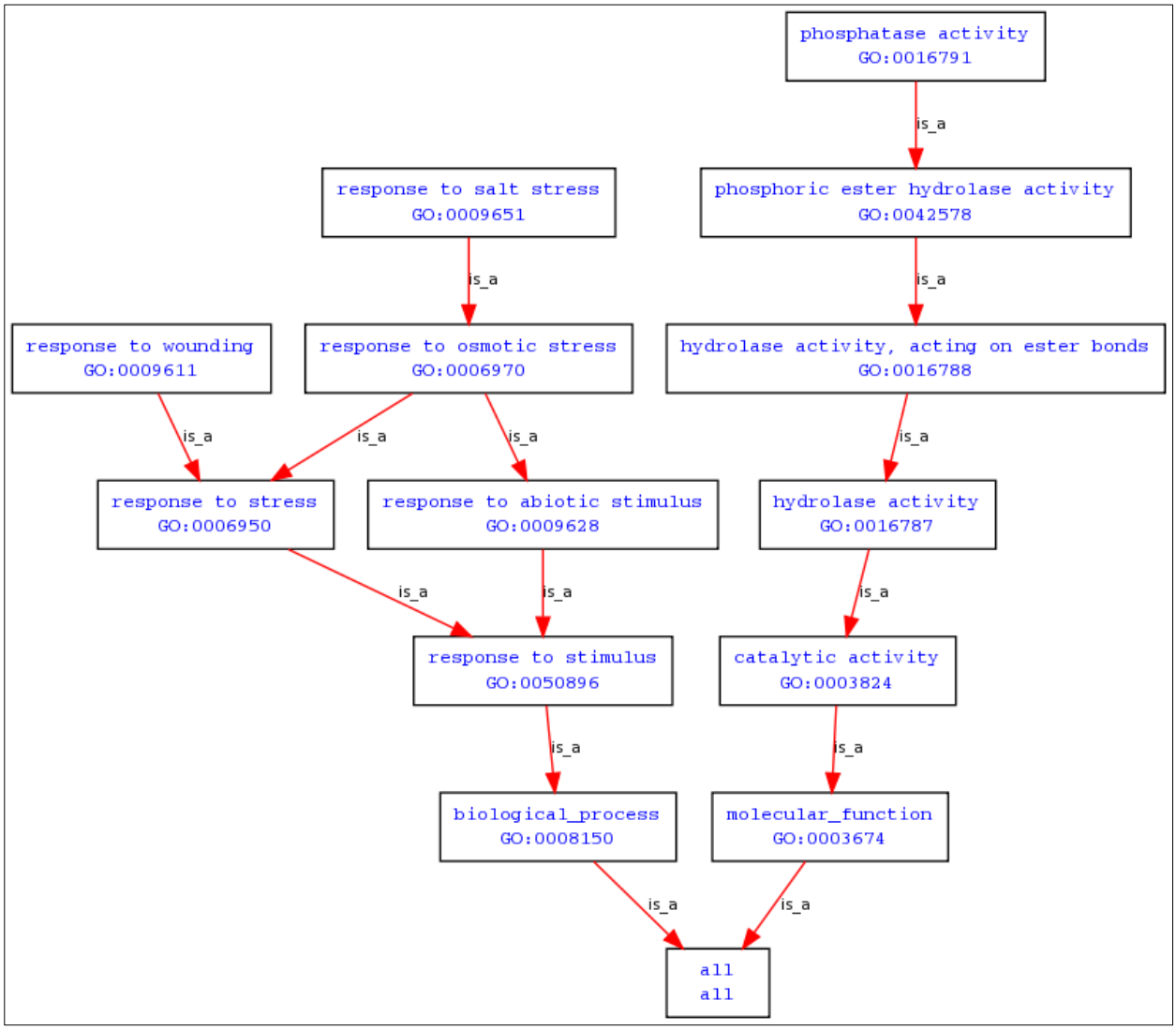


**Fig. 3.18:** Diverse functions of CPL1 gene reveled from Gene Ontology annotation

### 3.5.3.5 CPL1 Interactome:

CPL1 interaction was revealed from GeneMania which shows that DMS3, RTL1, MYB3, DRB2, DRB4 are closely interacting with CPL1. The interaction is depicted below (Fig. 3.19):

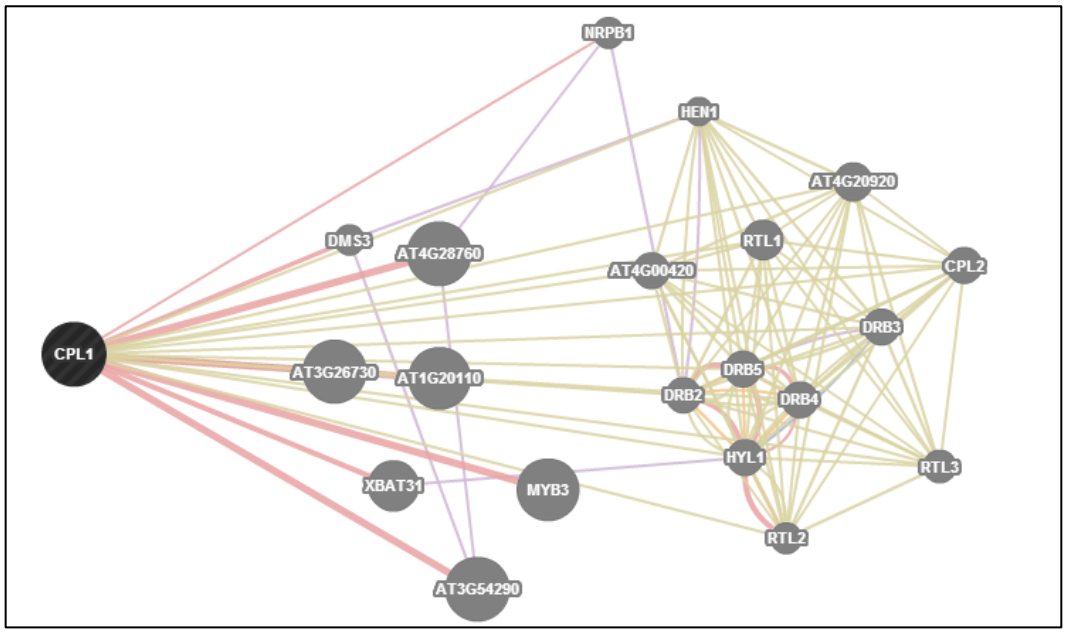


**Fig. 3.19:** CPL1 interaction with important stress responsive molecules

### 3.5.4 ERD5 (Early Responsive to Dehydration 5):

#### 3.5.4.1 Amino acid sequence of ERD5:

Amino acid sequence was retrieved from NCBI database which is about 499 amino acid chain long. The sequence is given below:

MATRLLRTNFIRRSYRLPAFSPVGPPTVTASTAVVPEILSFGQQAPEPPLHHPKPTEQSHDGLDLSDQARLFSSIPTSDLLR
STAVLHAAAIGPMVDLGTWVMSSKLMDASVTRGMVLGLVKSTFYDHFCAGEDADAAAERVRSVYEATGLKGMLVY
GVEHADDAVSCDDNMQQFIRTIEAAKSLPTSHFSSVVVKITAICPISLLKRVSDLLRWEYKSPNFKLSWKLKSFPVFSESS
PLYHTNSEPEPLTAEEERELEAAHGRIQEICRKCQESNVPLLIDAEDTILQPAIDYMAYSSAIMFNADKDRPIVYNTIQAYL
RDAGERLHLAVQNAEKENVPMGFKLVRGAYMSSEASLADSLGCKSPVHDTIQDTHSCYNDCMTFLMEKASNGSGFGV
VLATHNADSGRLASRKASDLGIDKQNGKIEFAQLYGMSDALSFGLKRAGFNVSKYMPFGPVATAIPYLLRRAYENRGM
MATGAHDRQLMRMELKRRLIAGIA

#### 3.5.4.2 Blast hit of ERD5:

Amino acid sequences were then blasted in NCBI Blast P-suit to check conserved domains and sequence similarities among other plant species. From the blast analysis, it was revealed that ERD5 is highly common in sequence with different plant species (Fig. 3.20).

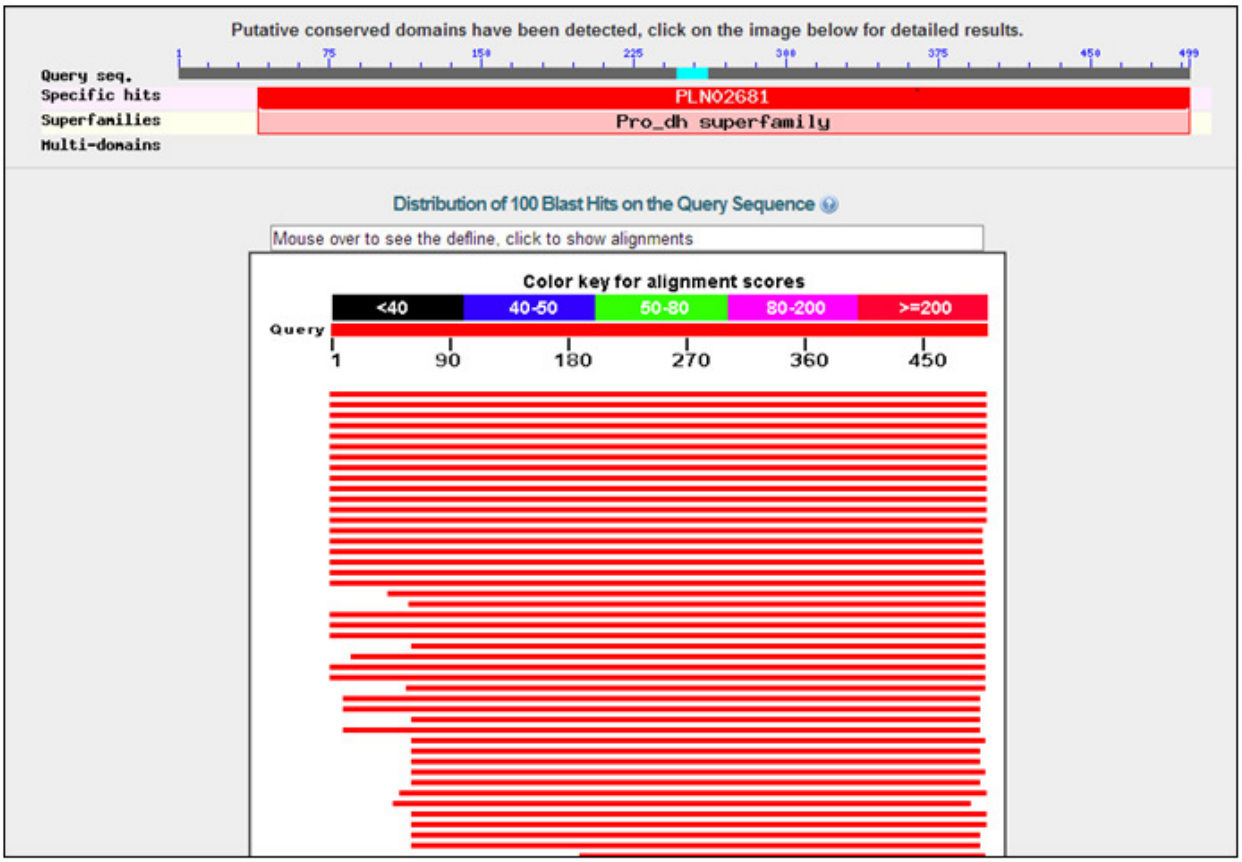


**Fig. 3.20:** Conserved domain sequence similarities of ERD5 from NCBI Blast P-suite

### 3.5.4.3 Interproscan domain search in ERD5:

Interproscan tool from European Bioinformatics Institute (EBI) was used to find out characterized domains present in ERD5. It was revealed that, three conserved domains (Fig. 3.21) were available in the sequence of ERD5, namely,

- Proline dehydrogenase
- Proline oxidase, and
- unintegrated

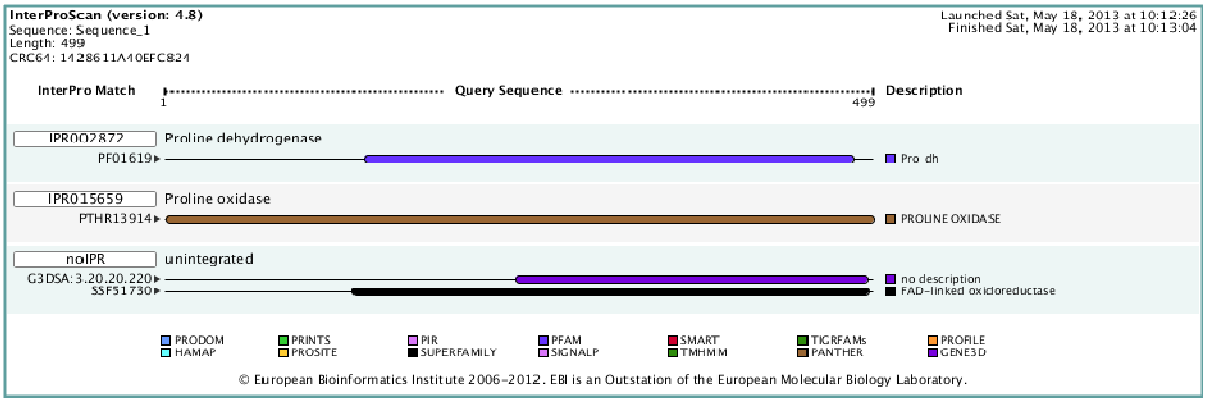


**Fig. 3.21:** Domains found through InterProScan tool in ERD5

### 3.5.4.4 GO function annotation for ERD5:

GO annotation for ERD5 revealed its biological functions in abiotic stress which is presented below (Fig. 3.22). ERD5 mainly acts as a drought responsive elements and also responsible for oxidative stress response.

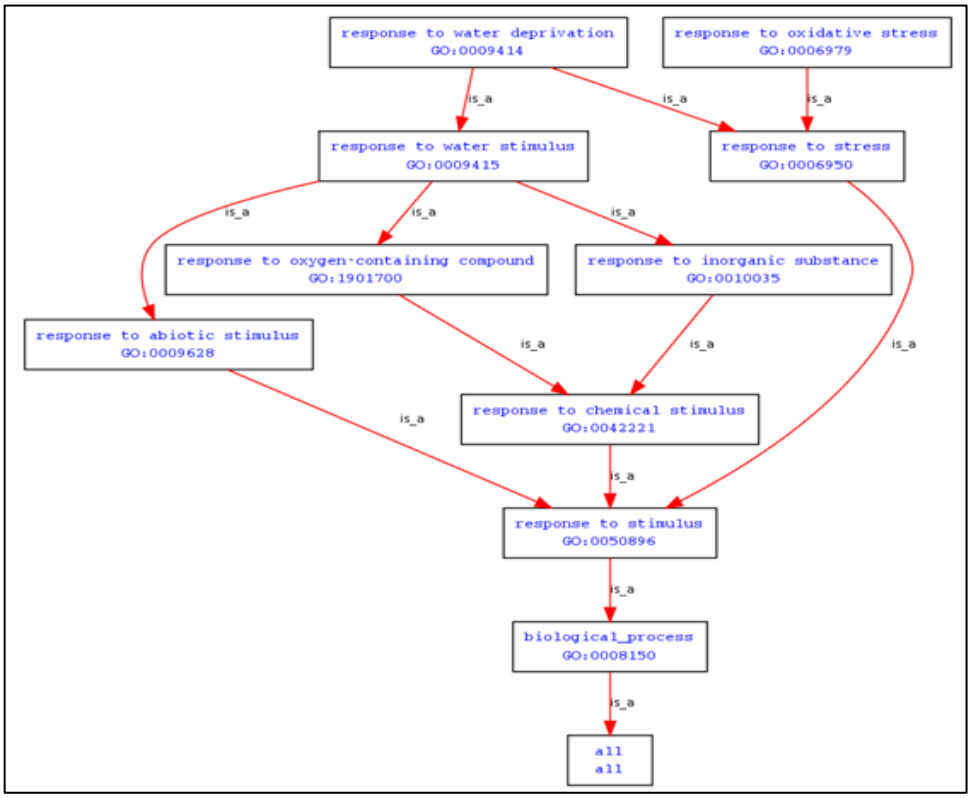


**Fig. 3.22:** Biological functions of ERD5

### 3.5.4.5 ERD5 Interactome:

Further studies with ERD5 showed that it interacts with ATFH, EXO, GDH2, CPK4 etc. and the complex network was revealed from GeneMania database (Fig. 3.23).

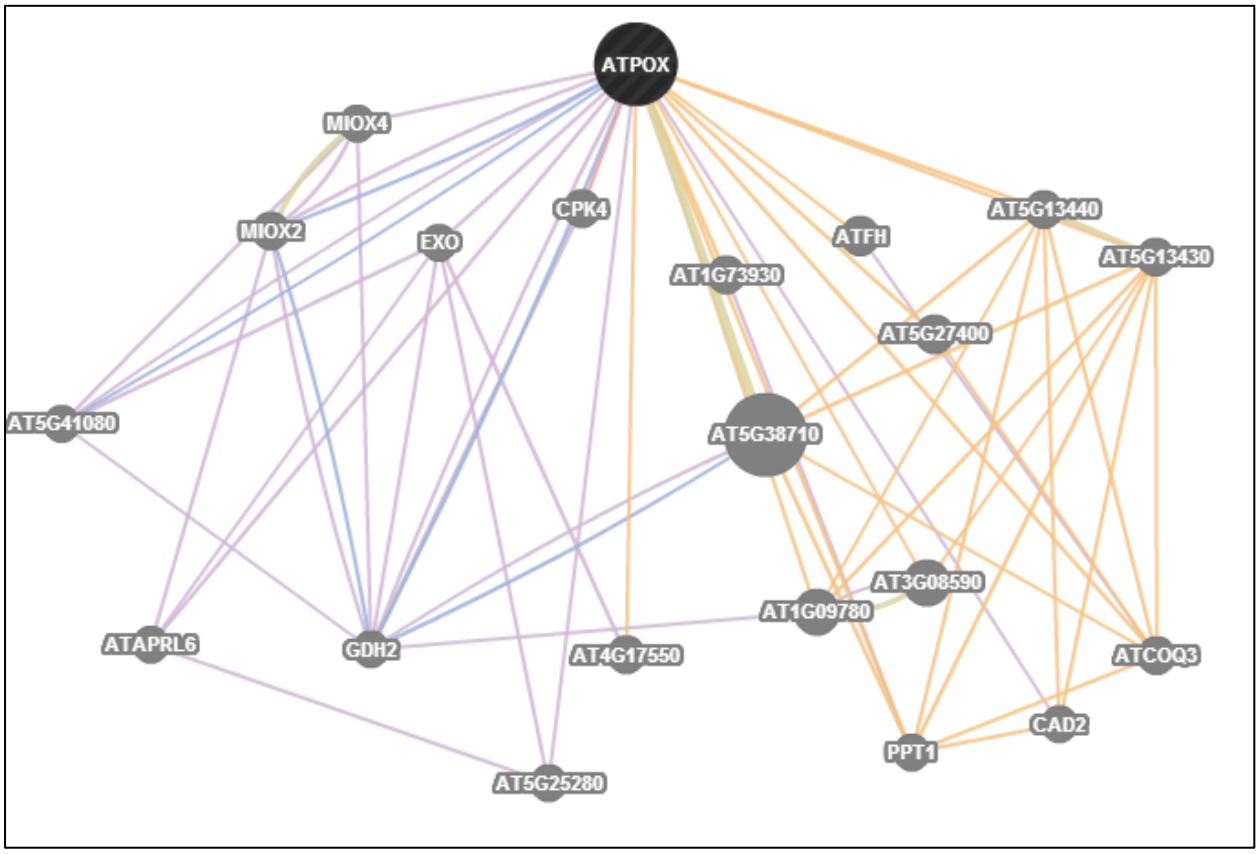


**Fig. 3.23:** ERD5 interacts with abiotic stress responsive molecule

### 3.5.5 NHX1 (Na+/H+ Exchanger):

#### 3.5.5.1 Amino acid sequence of NHX1:

The amino acid sequence of the Na+/H+ antiporter has been revealed from NCBI dataset for further analysis. The amino acid sequence is about 538 amino acid chain long which is as follows:

MLDSLVSKLPSLSTSDHASVVALNLFVALLCACIVLGHLLEENRWMNESITALLIGLGTGVTILLISKGKSSHLLVFSEDL
FFIYLLPPIIFNAGFQVKKKQFFRNFVTIMLFGAVGTIISCTIISLGVTQFFKKLDIGTFDLGDYLAIGAIFAATDSVCTLQVL
NQDETPLLYSLVFGEGVVNDATSVVVFNAIQSFDLTHLNHEAAFHLLGNFLYLFLLSTLLGAATGLISAYVIKKLYFGRH
STDREVALMMLMAYLSYMLAELFDLSGILTVFFCGIVMSHYTWHNVTESSRITTKHTFATLSFLAETFIFLYVGMDALDI
DKWRSVSDTPGTSIAVSSILMGLVMVGRAAFVFPLSFLSNLAKKNQSEKINFNMQVVIWWSGLMRGAVSMALAYNKF
TRAGHTDVRGNAIMITSTITVCLFSTVVFGMLTKPLISYLLPHQNATTSMLSDDNTPKSIHIPLLDQDSFIEPSGNHNVPRP
DSIRGFLTRPTRTVHYYWRQFDDSFMRPVFGGRGFVPFVPGSPTERNPPDLSKA

### 3.5.5.2 Blast hit of NHX1:

Amino acid sequences were then blasted in NCBI Blast P-suit to check conserved domains and sequence similarities among other plant species. From the blast analysis, it was revealed that NHX1 is highly common in sequence similarities in different plant species (Fig. 3.24).

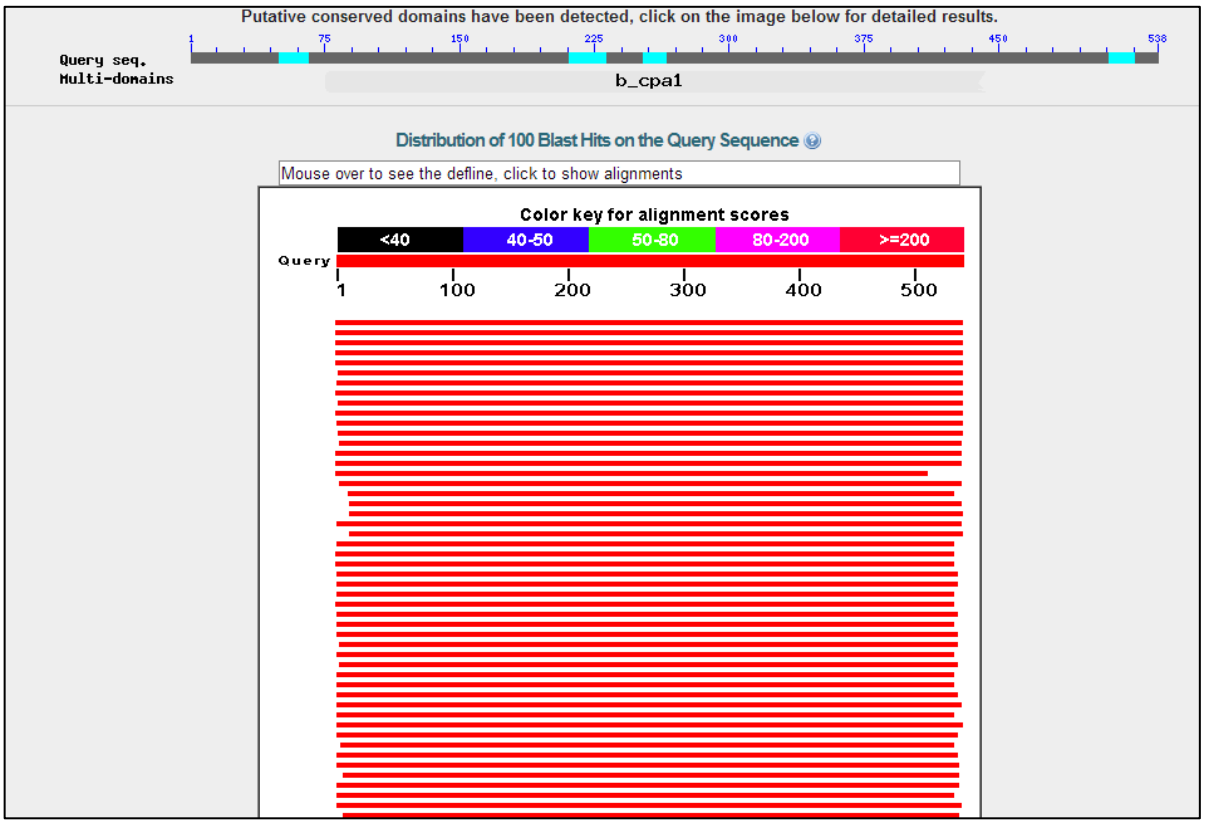


**Fig. 3.24:** Blast Hit with NHX1 gene and found the match with maximum conserved region

### 3.5.5.3 Interproscan domain search within NHX1:

Interproscan tool from European Bioinformatics Institute (EBI) was used to find out characterized domains present in NHX1. It was revealed that, four conserved domains (Fig. 3.25) were available in the sequence of NHX1, namely,

- Na+/H+ exchanger
- Cation/H+ exchanger
- Cation/H+ exchanger, CPA1 family and
- unintegrated

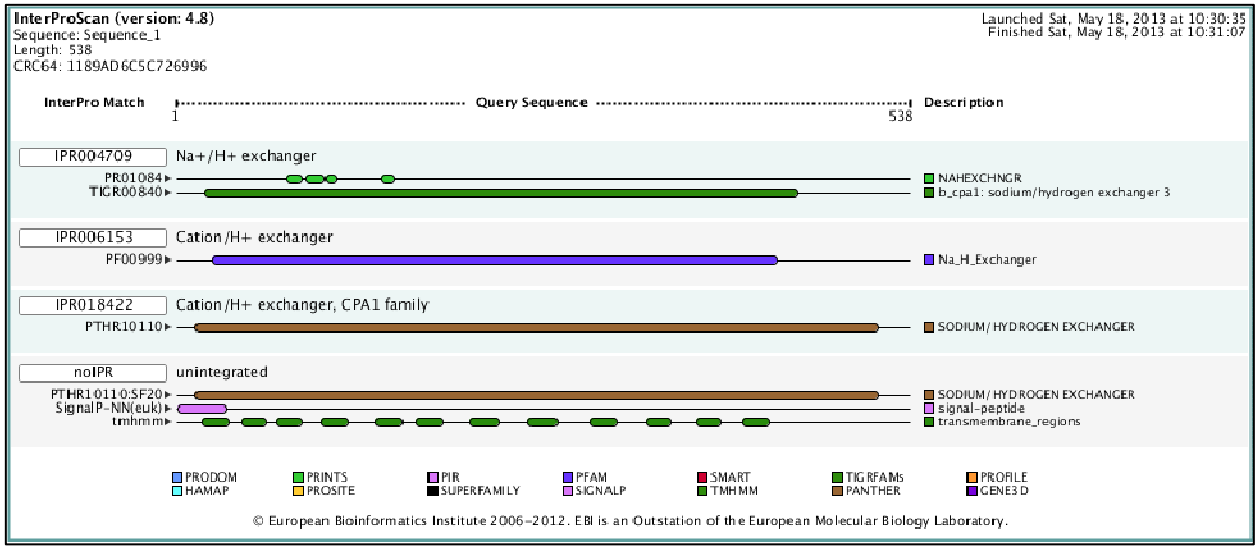


**Fig. 3.25:** Domains predicted by InterProScan of NHX1 amino acid sequence

### 3.5.5.4 GO function annotation of NHX1:

NHX1 was characterized as a regulator molecule and found to play a significant role in transporting excess Na+ into vacuole and also involved in K+ sequestration (Fig. 3.26). It was also found to be involved in osmotic stress response and other mechanism related to abiotic stress tolerance mechanism which entirely revealed in GO analysis. Figure 3.26 presents the functions of NHX1.

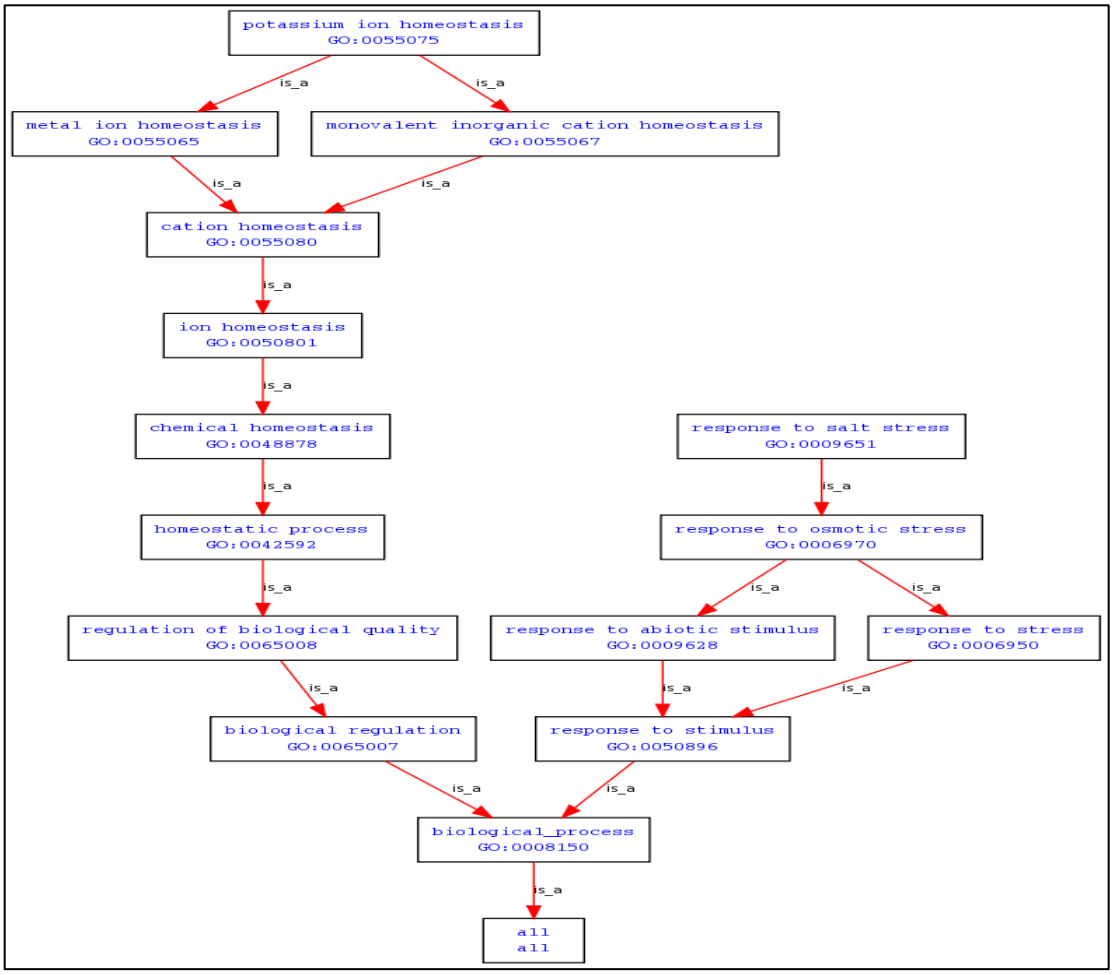


**Fig. 3.26:** GO profiles of NHX1 gene which depicted all biological functions of NHX1

### 3.5.5.5 NHX1 Interactome:

The individual interaction of NHX1 using GeneMania database revealed that it has a strong physical binding affinity with other antiporters, like, NHX2, NHX3, CHX2, SOS1 etc. which makes it more important molecule in stress response mechanism. The whole interaction with NHX1 has been shown below (Fig. 3.27).

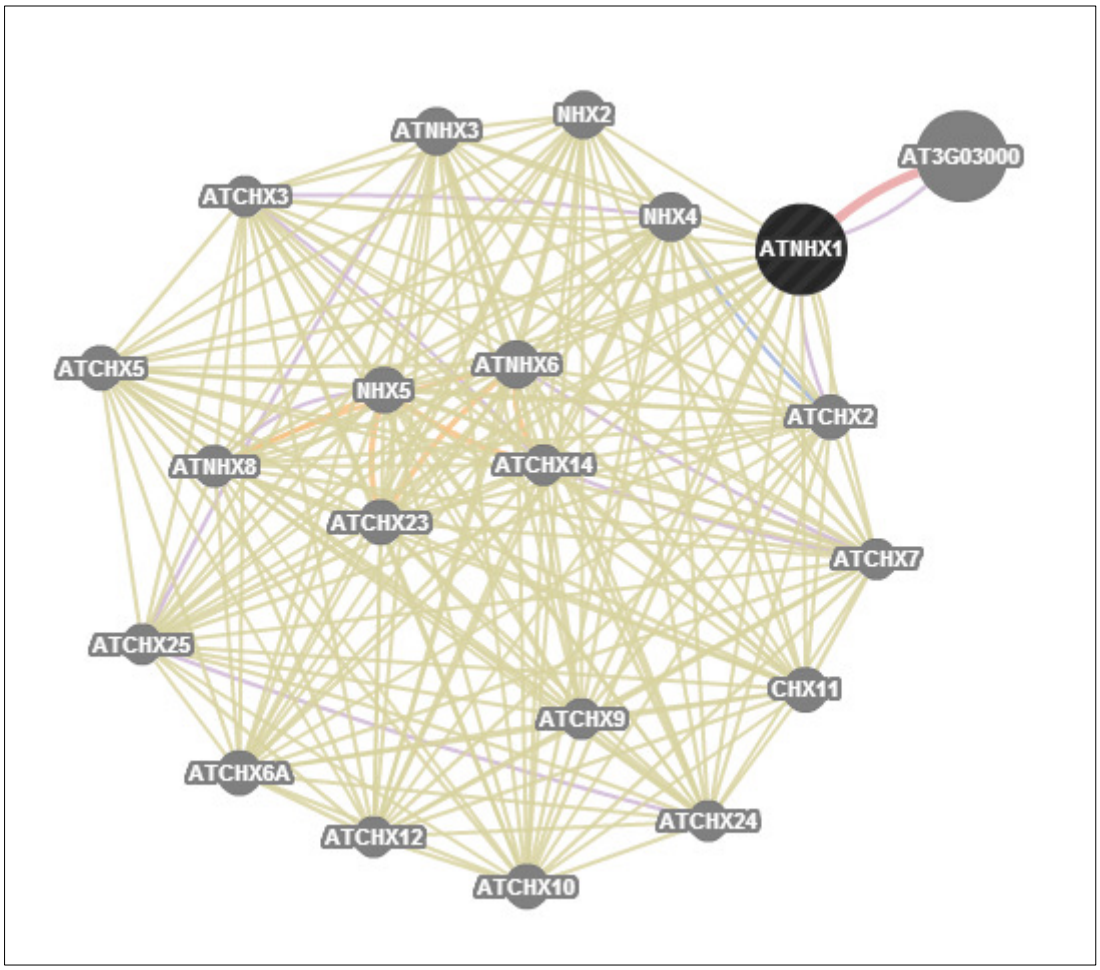


**Fig. 3.27:** NHX1 interacting partner revealed from GeneMania

### 3.5.6 SOS1 (Salt Overly Sensitive 1):

#### 3.5.6.1 Amino Acid Sequence of SOS1:

SOS1 amino acid sequence has been retrieved from NCBI database which is 1146 amino acid chain long. The sequence is as follows:

MTTVIDATMAYRFLEEATDSSSSSSSSKLESSPVDAVLFVGMSLVLGIASRHLLRGTRVPYTVALLVIGIALGSLEYGAK
HNLGKIGHGIRIWNEIDPELLLAVFLPALLFESSFSMEVHQIKRCLGQMVLLAVPGVLISTACLGSLVKVTFPYEWDWKT
SLLLGGLLSATDPVAVVALLKELGASKKLSTIIEGESLMNDGTAIVVFQLFLKMAMGQNSDWSSIIKFLLKVALGAVGIG
LAFGIASVIWLKFIFNDTVIEITLTIAVSYFAYYTAQEWAGASGVLTVMTLGMFYAAFARTAFKGDSQKSLHHFWEMVA
YIANTLIFILSGVVIAEGILDSDKIAYQGNSWRFLFLLYVYIQLSRVVVVGVLYPLLCRFGYGLDWKESIILVWSGLRGAV
ALALSLSVKQSSGNSHISKETGTLFLFFTGGIVFLTLIVNGSTTQFVLRLLRMDILPAPKKRILEYTKYEMLNKALRAFQD
LGDDEELGPADWPTVESYISSLKGSEGELVHHPHNGSKIGSLDPKSLKDIRMRFLNGVQATYWEMLDEGRISEVTANIL
MQSVDEALDQVSTTLCDWRGLKPHVNFPNYYNFLHSKVVPRKLVTYFAVERLESACYISAAFLRAHTIARQQLYDFLG
ESNIGSIVINESEKEGEEAKKFLEKVRSSFPQVLRVVKTKQVTYSVLNHLLGYIENLEKVGLLEEKEIAHLHDAVQTGLK

KLLRNPPIVKLPKLSDMITSHPLSVALPPAFCEPLKHSKKEPMKLRGVTLYKEGSKPTGVWLIFDGIVKWKSKILSNNHS
LHPTFSHGSTLGLYEVLTGKPYLCDLITDSMVLCFFIDSEKILSLQSDSTIDDFLWQESALVLLKLLRPQIFESVAMQELRA
LVSTESSKLTTYVTGESIEIDCNSIGLLLEGFVKPVGIKEELISSPAALSPSNGNQSFHNSSEASGIMRVSFSQQATQYIVET
RARAIIFNIGAFGADRTLHRRPSSLTPPRSSSSDQLQRSFRKEHRGLMSWPENIYAKQQQEINKTTLSLSERAMQLSIFGS
MVNVYRRSVSFGGIYNNKLQDNLLYKKLPLNPAQGLVSAKSESSIVTKKQLETRKHACQLPLKGESSTRQNTMVESSD
EEDEDEGIVVRIDSPSKIVFRNDL

### 3.5.6.2 Blast hit of SOS1:

Amino acid sequences were blasted in NCBI Blast P-suit to check conserved domains and sequence similarities among other plant species. From the blast analysis, it was revealed that SOS1 is highly common in different plant species sequence (Fig. 3.28).

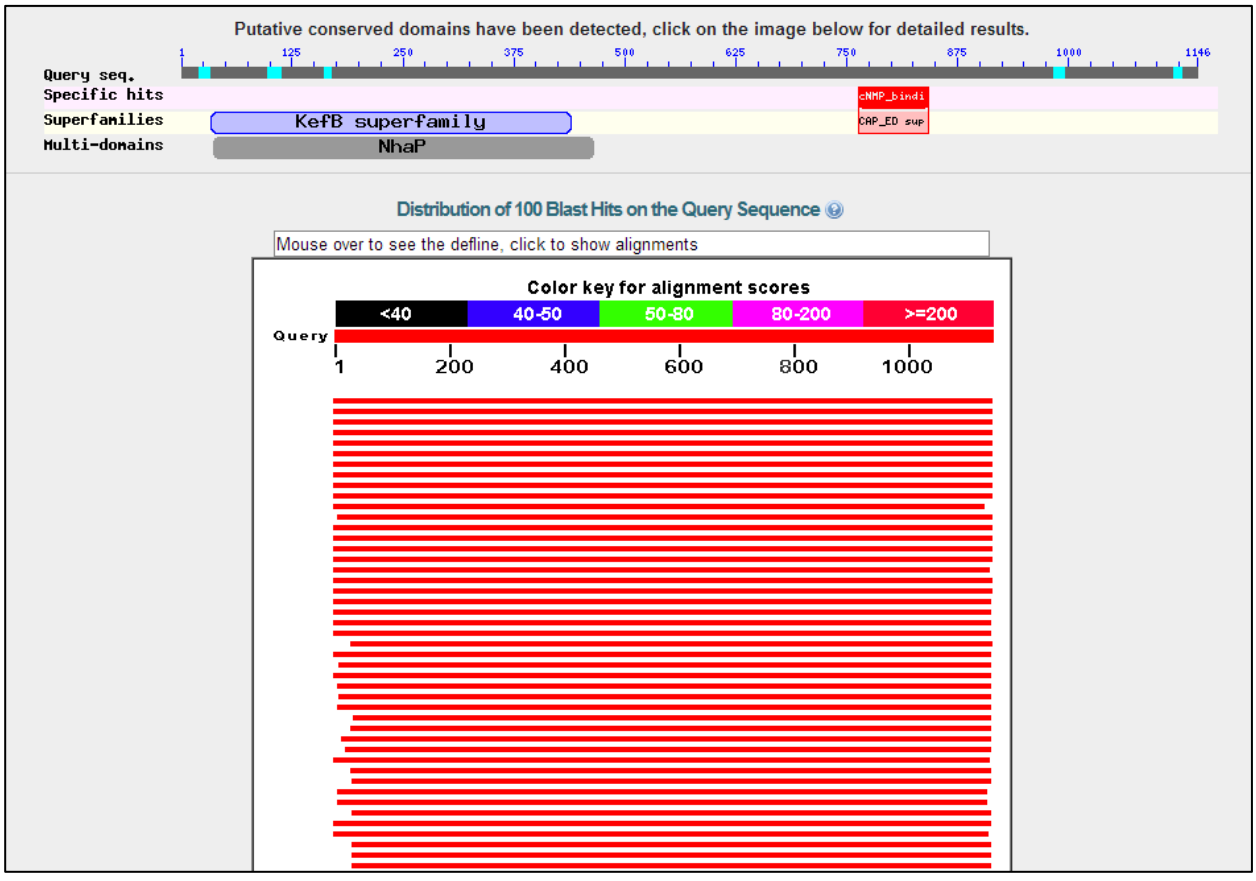


**Fig. 3.28:** Blast hit of SOS1 gene

### 3.5.6.3 SOS1 Interproscan domain search:

Interproscan tool from European Bioinformatics Institute (EBI) was used to find out characterized domains present in SOS1. It was revealed that, six conserved domains were available in the sequence of SOS1 (Fig. 3.29), namely,

- Cyclic nucleotide-binding domain
- Cation/H+ exchanger
- Na+/H+ exchanger, isoforms 7/8
- Cation/H+ exchanger, CPA1 family
- Cyclic nucleotide-binding-like, and
- Unintegrated.

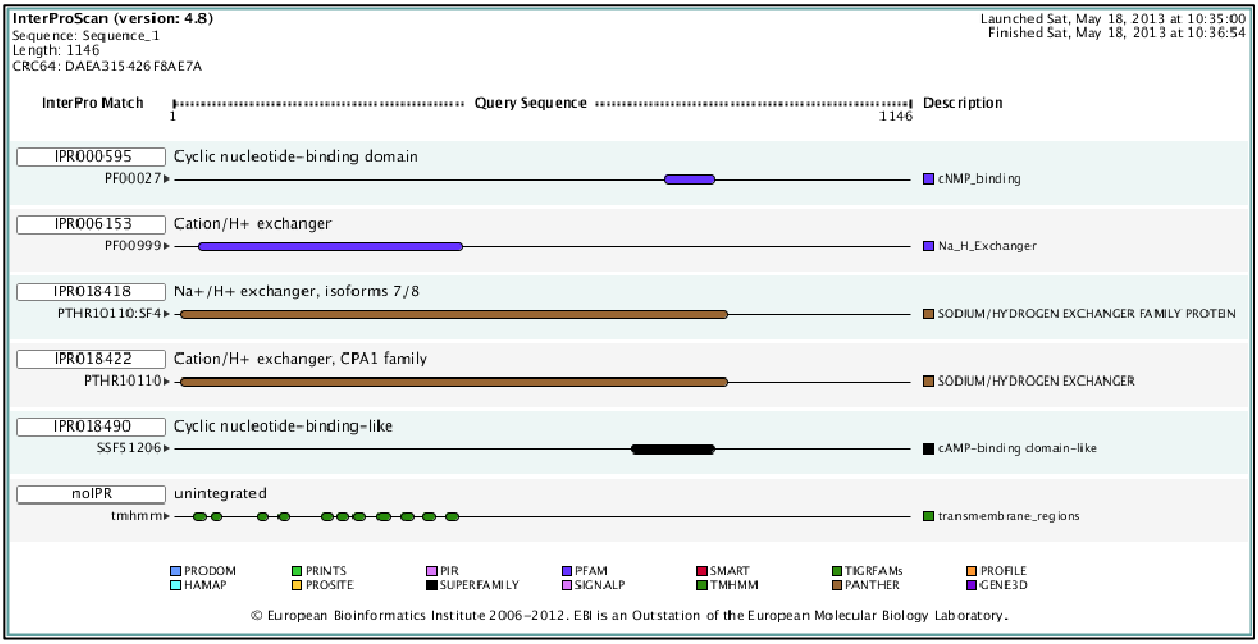


**Fig. 3.29:** Interproscan predicted domain of SOS1 gene

### 3.5.6.4 GO Function Annotation:

GO annotation of SOS1 gene revealed that it has a major role in modulating stress response mechanism like response to salinity, cation transporter, metal ion transporter etc. The functions of SOS1 gene revealed by GO is shown below (Fig. 3.30). It has functions diversity. It mainly works on Na+ pumping during salinity stress and also acts as a regulator to stimulate oxidative stress responsive genes like RCD1, DREB2A, GPX etc.

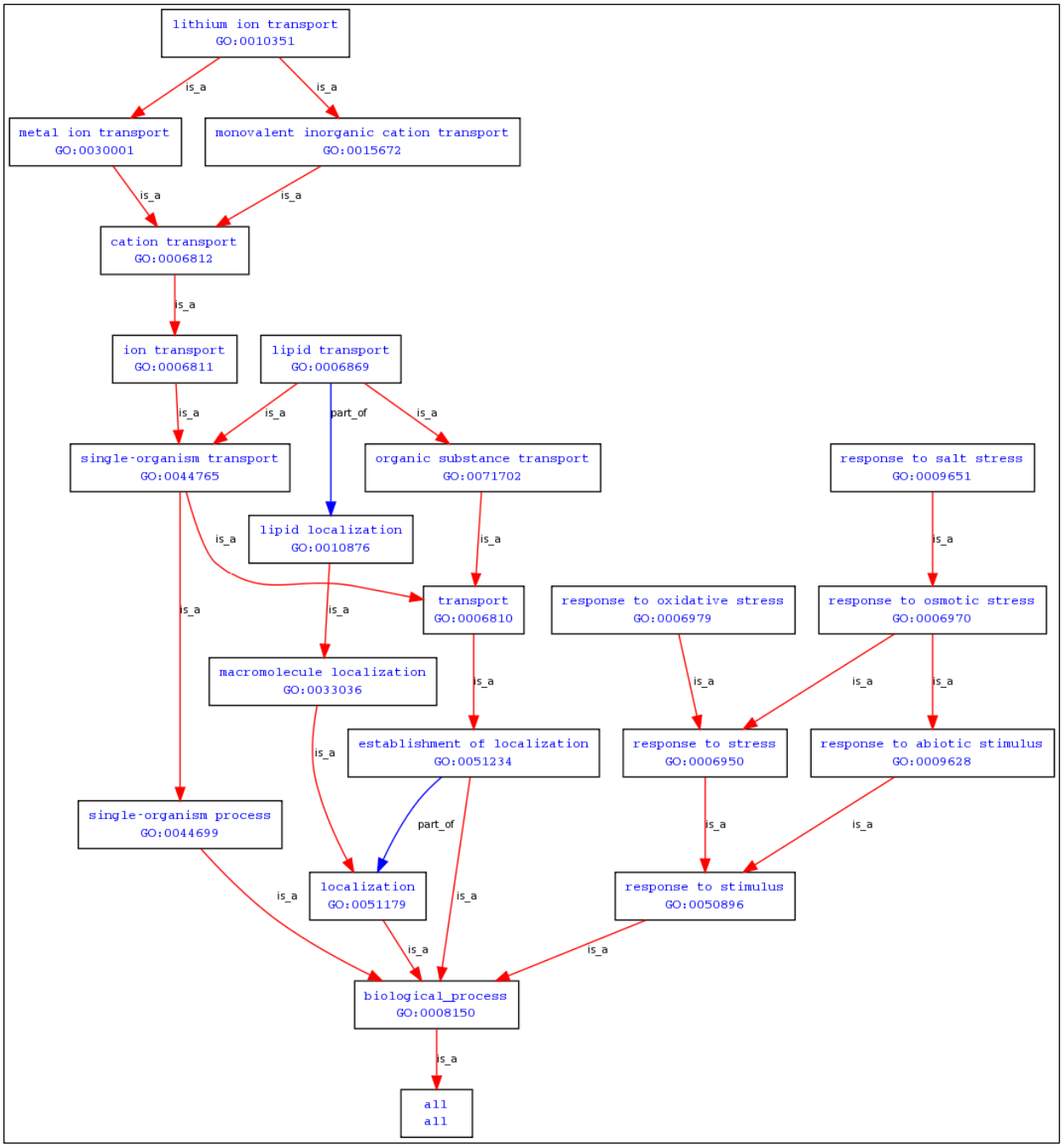


**Fig. 3.30:** GO profiles of SOS1 gene which depicted all biological functions of SOS1

### 3.5.6.5 SOS1 Interactome:

GeneMania database again applied to reveal the interacting mechanism of SOS1 gene, it revealed that it bears the interaction potential with NHX1, NHX2, NHX4, KEA5, CNGC1, SOS2 and SOS3 etc. (Fig. 3.31).

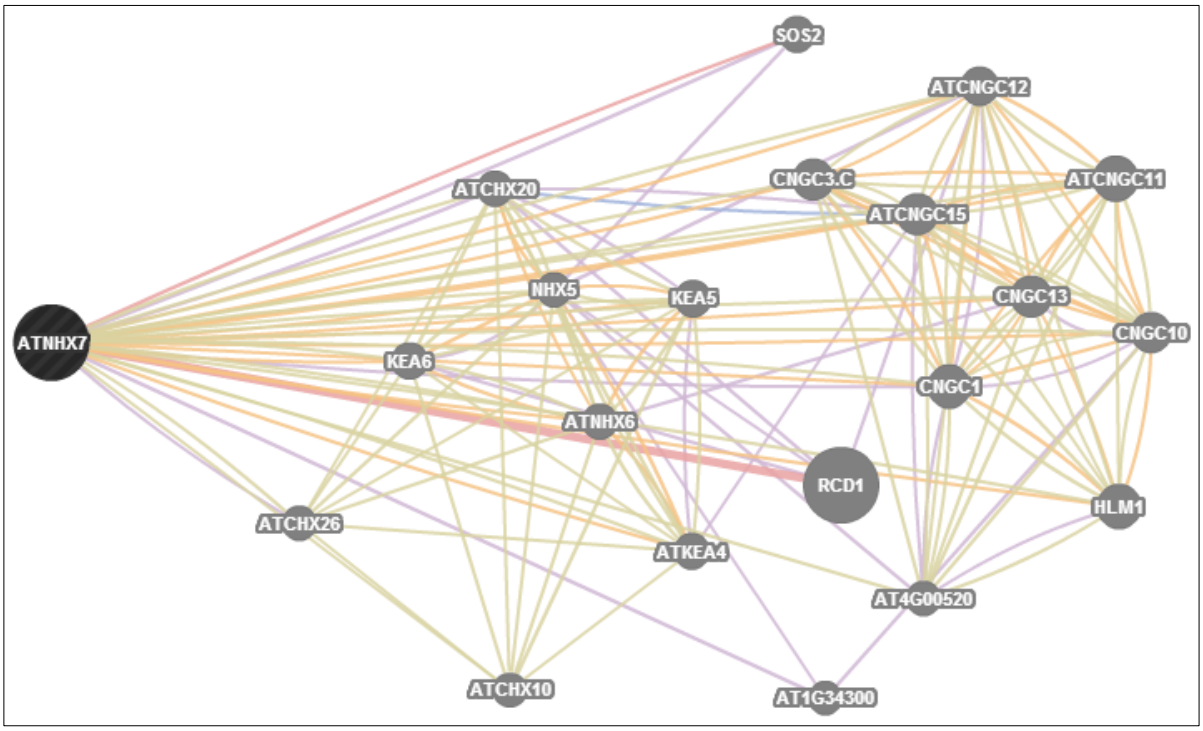


**Fig. 3.31:** Protein-Protein Interaction of SOS1 gene revealed from GeneMania database

### 3.5.7 SOS2 (Salt Overly Sensitive 1):

### 3.5.7.1 Amino Acid Sequence of SOS2:

Amino Acid sequence of SOS2 gene has been retrieved from NCBI database which is 446 amino acid chain long. The amino acid sequence is as follows:

MTKKMRRVGKYEVGRTIGEGTFAKVKFARNTDTGDNVAIKIMAKSTILKNRMVDQIKREISIMKIVRHPNIVRLYEVLA
SPSKIYIVLEFVTGGELFDRIVHKGRLEESESRKYFQQLVDAVAHCHCKGVYHRDLKPENLLLDTNGNLKVSDFGLSAL
PQEGVELLRTTCGTPNYVAPEVLSGQGYDGSAADIWSCGVILFVILAGYLPFSETDLPGLYRKINAAEFSCPPWFSAEVK
FLIHRILDPNPKTRIQIQGIKKDPWFRLNYVPIRAREEEEVNLDDIRAVFDGIEGSYVAENVERNDEGPLMMNAFEMITLS
QGLNLSALFDRRQDFVKRQTRFVSRREPSEIIANIEAVANSMGFKSHTRNFKTRLEGLSSIKAGQLAVVIEIYEVAPSLFM
VDVRKAAGETLEYHKFYKKLCSKLENIIWRATEGIPKSEILRTITF

### 3.5.7.2 Blast hit of SOS2:

Amino acid sequences were then blasted in NCBI Blast P-suit to check conserved domains and sequence similarities among other plant species. From the blast analysis, it was revealed that SOS2 is highly similar in sequence with different plant species (Fig. 3.32).

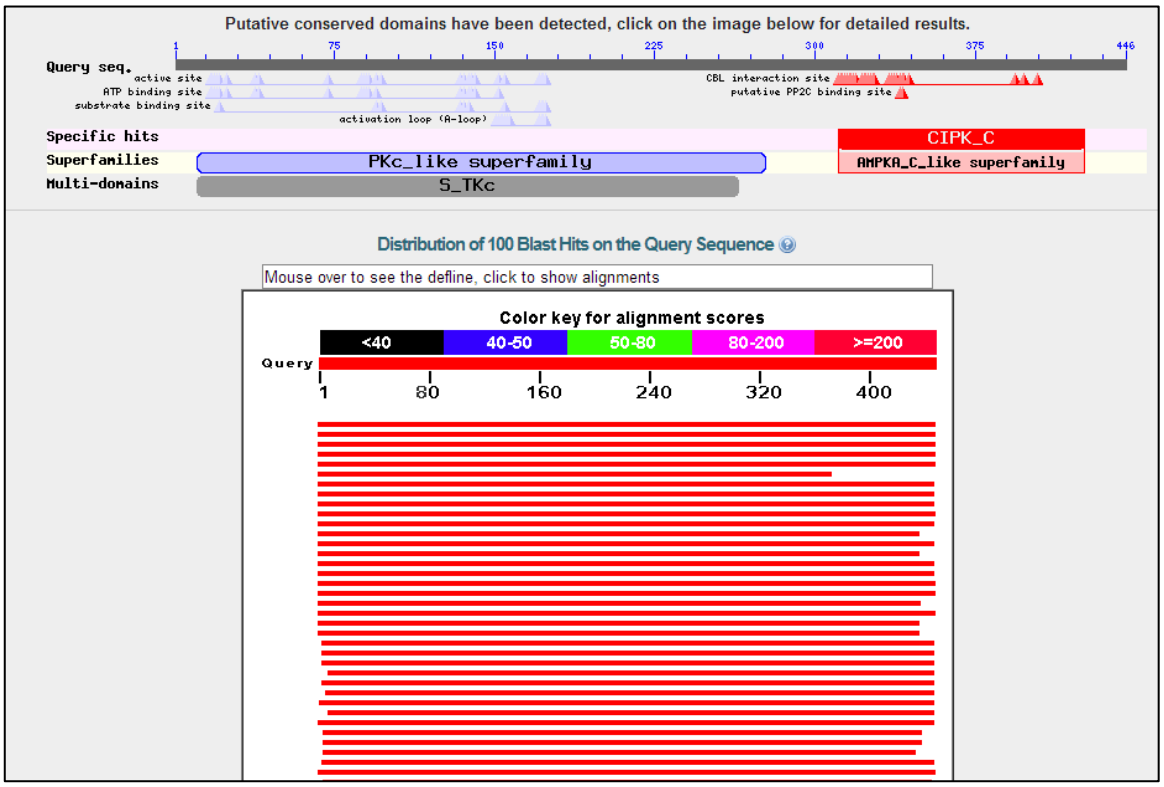


**Fig. 3.32:** Blast hit with SOS2 gene revealed conserved region among species.

### 3.5.7.3 Protein domain search of SOS2 by Interproscan:

Interproscan tool from European Bioinformatics Institute (EBI) was used to find out characterized domains present in SOS2. It was revealed that, nine conserved domains were available in the sequence of SOS2, namely

- Protein kinase, catalytic domain
- Serine/threonine-/dual specificity protein kinase, catalytic domain
- NAF domain
- Serine/threonine-protein kinase, active site
- Protein kinase-like domain
- Protein kinase, ATP binding site
- NAF/FISL domain
- Tyrosine-protein kinase, catalytic domain and
- unintegrated (Fig. 3.33).


**Fig. 3.33:** Predicted domains found in SOS2 gene from Interproscan.

### 3.5.7.4 GO function annotation of SOS2:

SOS2 gene has a strong role in protein kinase activity, salt stress response, phosphotransferase activity etc. GO file denotes its all type of biological processes and the comprised diagram of SOS2 is presented below (Fig. 3.34). It has kinase activity and regulates SOS1 expression during stress.

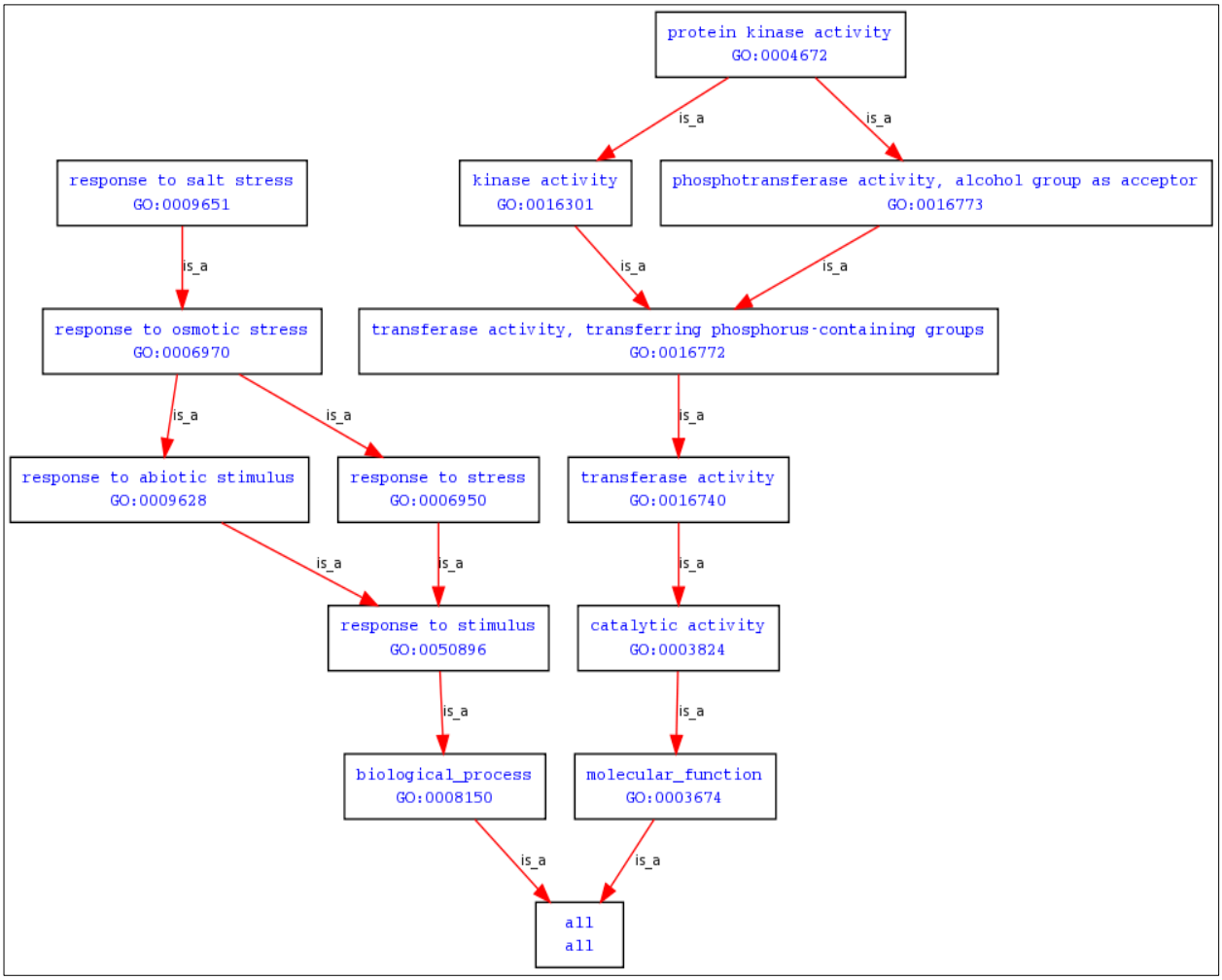


**Fig. 3.34:** GO result of SOS2 gene

### 3.5.7.5 SOS2 Interactome:

The SOS2 gene found to interact with CBL7, SOS1, SOS3 genes and all other major modulator during stress. The interaction was revealed through GeneMania (Fig. 3.35).

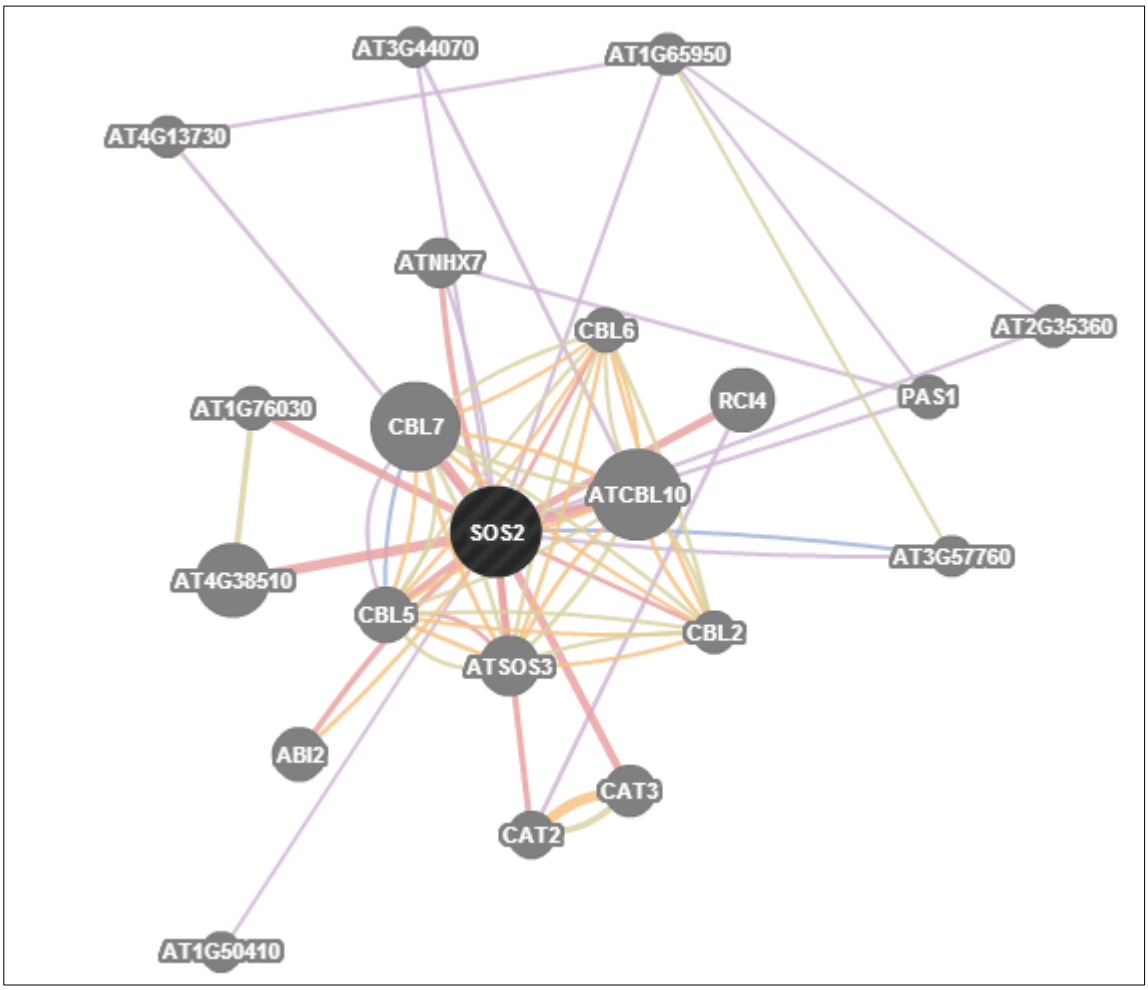


**Fig. 3.35:** SOS2 interaction partner retrieved using GeneMania.

### 3.5.8 SOS3 (SALT OVERLY SENSITIVE 3):

#### 3.5.8.1 Amino acid sequence of SOS3:

SOS3 amino acid sequence retrieved from NCBI database. The amino acid sequence is 222 amino acid chain long which is as below:

MGCSVSKKKKKNAMRPPGYEDPELLASVTPFTVEEVEALYELFKKLSSSIIDDGLIHKEEFQLALFRNRNRRNLFADRIF
DVFDVKRNGVIEFGEFVRSLGVFHPSAPVHEKVKFAFKLYDLRQTGFIEREELKEMVVALLHESELVLSEDMIEVMVDK
AFVQADRKNDGKIDIDEWKDFVSLNPSLIKNMTLPYLKDINRTFPSFVSSCEEEEMELQNVSS

#### 3.5.8.2 Blast hit of SOS3:

Amino acid sequences were then blasted in NCBI Blast P-suit to check conserved domains and sequence similarities among other plant species. From the blast analysis, it was revealed that SOS3 is highly common in sequence in different plant species (Fig. 3.36).

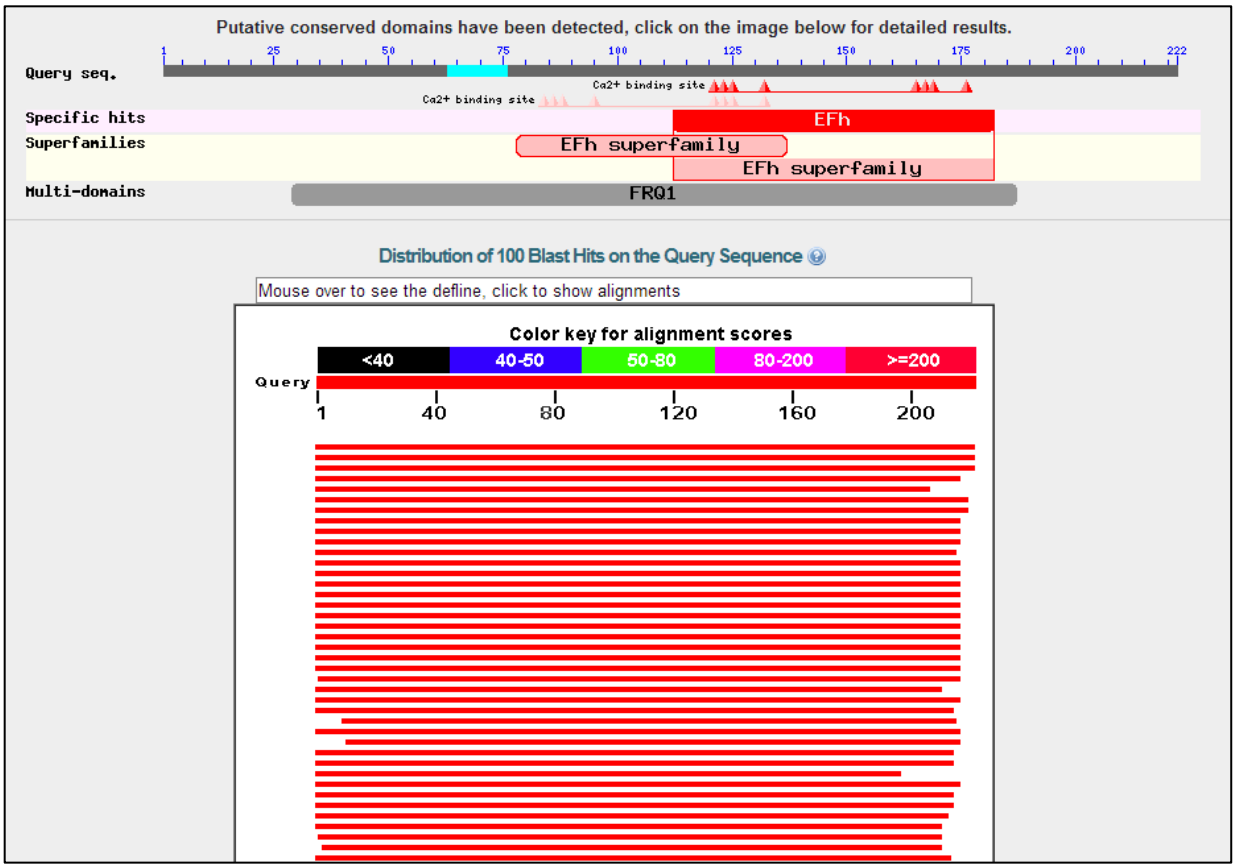


**Fig. 3.36:** Conserved domain sequence similarities of SOS3 gene from NCBI Blast P-suite.

### 3.5.8.3 Protein domain of SOS3 search by Interproscan:

Interproscan tool from European Bioinformatics Institute (EBI) was used to find out characterized domains present in SOS3. It was revealed that, four conserved domains were available in the sequence of SOS3, namely,

- Recoverin
- EF-hand domain
- EF-hand-like domain, and
- unintegrated (Fig. 3.37).

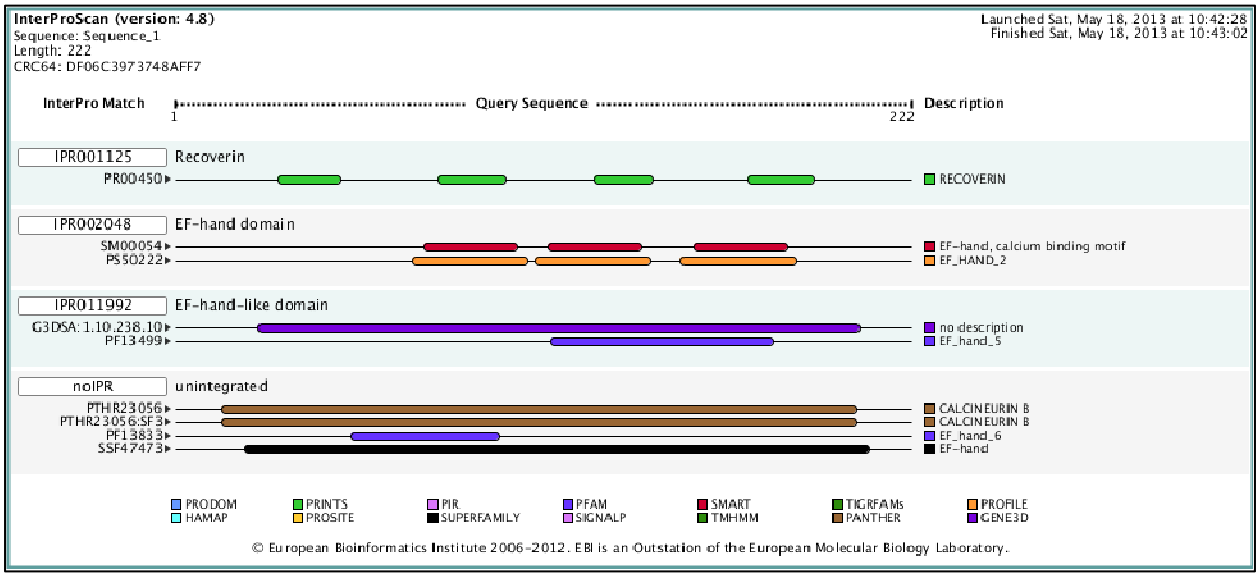


**Fig. 3.37:** Predicted domain of SOS3 gene from Interproscan search tool

### 3.5.8.4 GO function annotation of SOS3:

GO annotation revealed that SOS3 in involved in abiotic stress response, osmotic stress, salinity stress etc. GO file denotes its all type of biological processes and the comprised diagram of SOS3 gene is presented in (Fig. 3.38). It is mainly a Ca+ signal acceptor during salinity stress and binds to FISL motif of SOS2 and activates that. Then SOS2 shows up with kinase activity and phosphorylates SOS1.

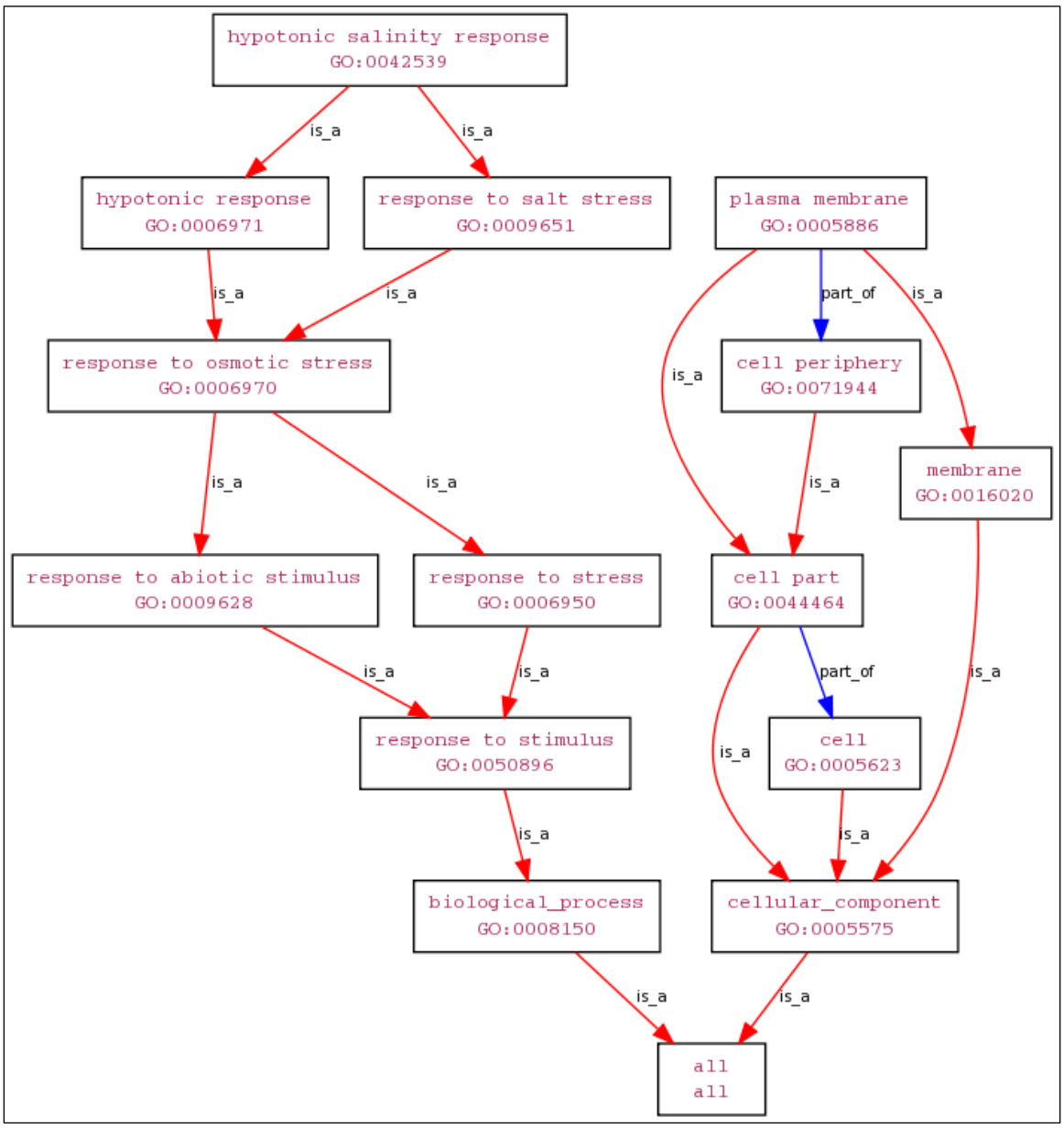


**Fig. 3.38:** GO annotation of SOS3 gene.

### 3.5.8.5 SOS3 Interactome:

The SOS3 gene interaction reveals that it has a strong interaction with protein kinase, CBL9, GPX3 etc. The interaction reveals that SOS3 is a major player of stress response mechanism (Fig. 3.39).

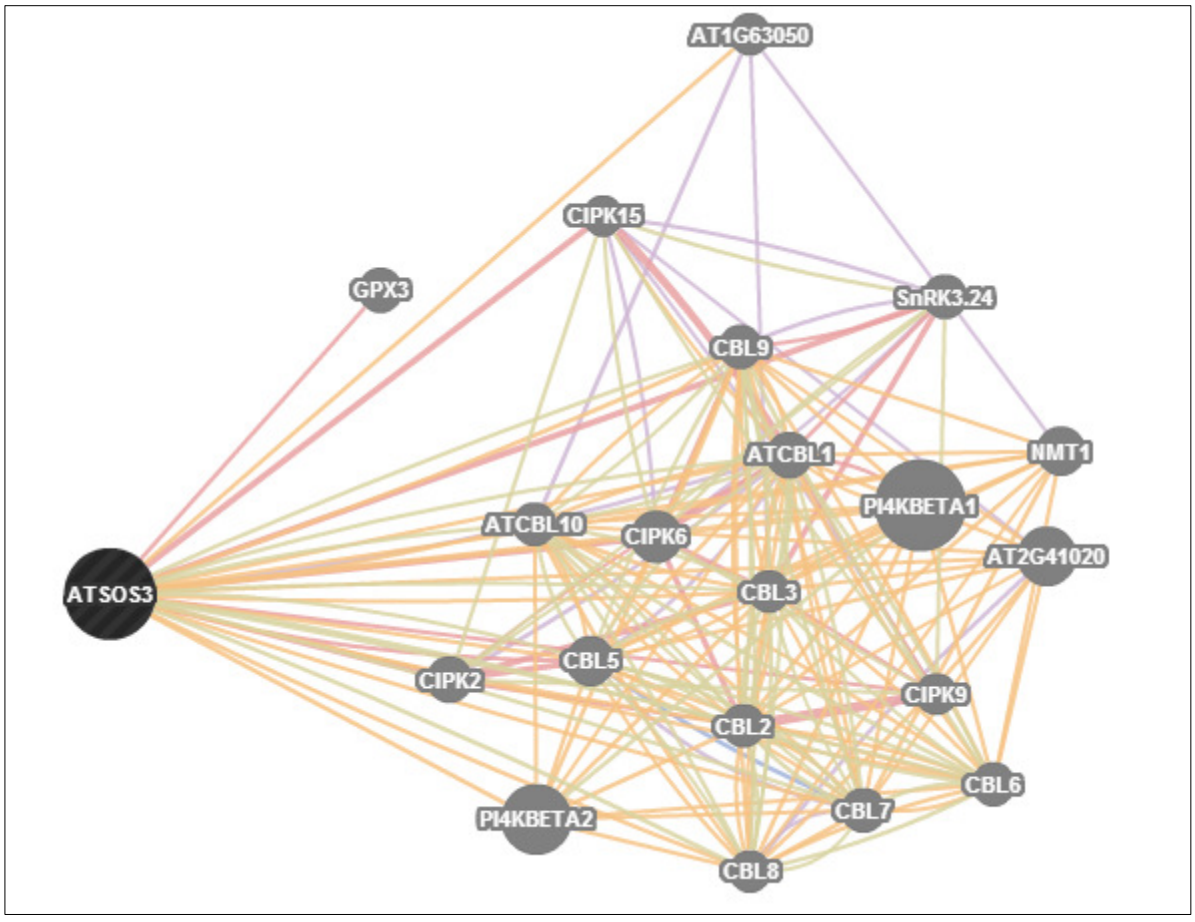


**Fig. 3.39:** SOS3 interaction revealed from GeneMania

### 3.6 Protein-Protein Network of These Targeted Transcription Factors and Regulatory Genes:

The different types of database used in the present study showed that these selected eight molecules are strongly co-related connectomes. Moreover, the experimental study suggests that these eight transcription factors, enzyme and regulatory genes are not only directly connected to each other but also they are connected to some other genes which also gets up-regulated during abiotic stress tolerance. The (Fig. 3.40) of eight transcription factors, enzyme and regulatory genes with the other genes which are indirectly connected to different abiotic stresses are given bellow,

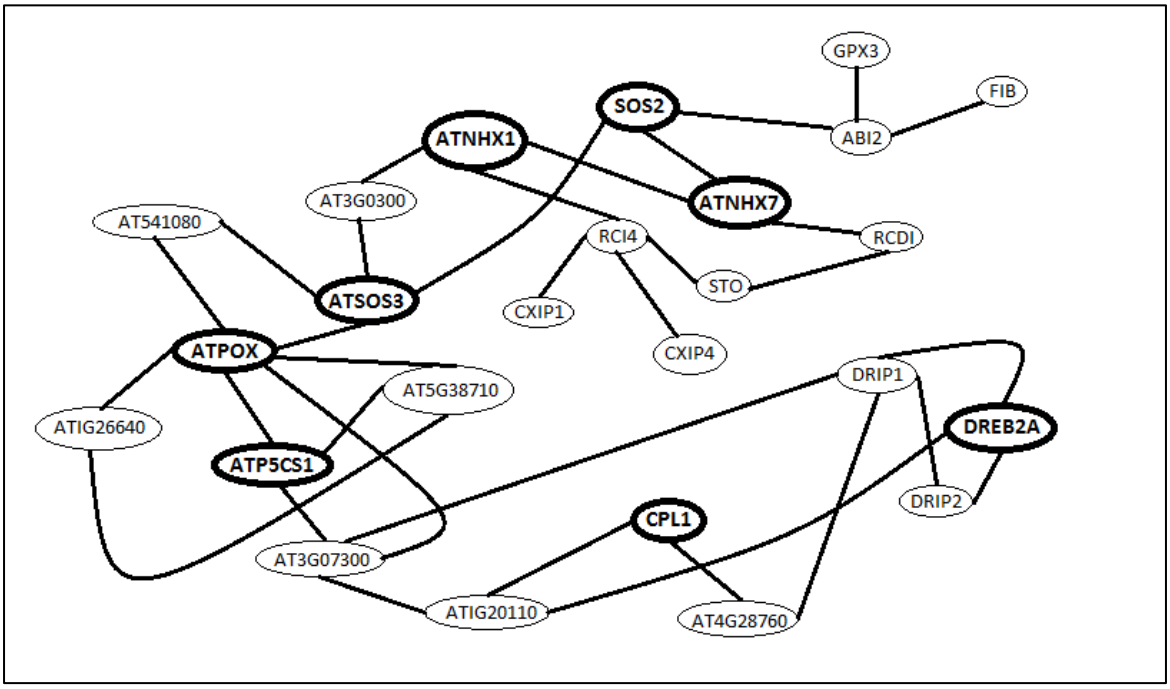


**Fig. 3.40:** Final interactome is shown between targeted transcription factors, enzyme and regulatory genes. **Bold** circles denote the regulatory genes, TFs and Enzyme.

# Discussion

## Discussion

Plants are constantly exposed to a wide range of environmental stresses, such as, drought, high salt, heat and cold. These abiotic stresses lead to growth constraints thus reduction of productivity. In many crop species this yield reduction may reach up-to 50% of the average production (Bray *et al*. 2000).

Tolerance and susceptibility of plants towards these abiotic stresses are in a very intricate network. These stresses are found to affect plants at multiple stages of development and often several stresses have been reported to affect the plant concurrently (Chinnusamy *et al.* 2004). For this reason, the basic mechanisms of abiotic stress tolerance and adaptation have been an interesting area to study.

Computational modeling has an important role in revealing genome-wide regulatory mechanisms. One can understand this intricate network using various softwares. In current study, Bioinformatics approaches have been taken to address the queries regarding abiotic stress responsive pathways in plants. Microarray data were retrieved from the most accessible database, ArrayExpress. This database is a repository for functional genomics data from both microarray and high-throughput sequencing studies, many of which are supported by peer-reviewed publications (Rustici G *et al*. 2013). The retrieved data were integrated and expression analyses were performed using Cytoscape. It can be used to visualize and analyze network graphs of any kind involving nodes and edges. There are so many reports of using Cytoscape as a means of analyzing expression data analysis (Cline M S *et al*. 2007; J. Montojo *et al*. 2010). Venny platform has also been reported as an extensive way of analyzing common patterns which was also used in the current study (Oliveros J C 2007). Pattern identifications and analyses were elaborately expanded using some basic programs like BLAST, Gene Ontology, Gene Mania and String databases. BLAST is one of the most widely used bioinformatics programs, because it addresses a fundamental problem and the heuristic algorithm it uses is much faster than calculating an optimal alignment (David W. Mount 2004). Gene ontology is a major bioinformatics initiative to unify the representation of gene and gene product attributes across all species. More specifically, the project aims to maintain and develop its controlled vocabulary of gene and gene product attributes; Annotate genes and gene products, and

assimilate and disseminate annotation data; provide tools for easy access to all aspects of the data provided by the project. This program receives excellent acceptations in understanding basic molecular interconnections (Barry Smith *et al*. 2007). In addition, GeneMANIA helps to find other genes that are related to a set of input genes, using a very large set of functional association data. Association data include protein and genetic interactions, pathways, co-expression, co-localization and protein domain similarity. It gained popularities right after the cytoscape plug-in and has been used in many peer-reviewed publications (Warde-Farley D *et al*. 2010). On the other hand, the STRING database contains information from numerous sources, including experimental data, computational prediction methods and public text collections. The latest version 9.0 contains information on about 5.2 millions proteins from 1133 species. This database is well referred and is used in many reported works (Franceschini A *et al*. 2013). So the statement from the above discussion really helps to come up with a strong concluding remark regarding the validity of the work that has been initiated in the current study. All tools and databases are extensively used to clarify biological queries and there have been lots more reported works published around the globe.

In model plant, *Arabidopsis thaliana*, thousands of genes have been identified to play different roles in different responses (tolerance and susceptibility). Naika and his colleagues (2013) did a similar study with *Arabidopsis* on stress responses. They compiled a datasets on abiotic stress responses and used functional enrichment analyses like GO (Gene Ontology) and PO (Plant Ontology) annotation to understand plant specific features associated with differentially up-regulation of genes for individual signals. They have also used STRING to derive interactome and to predicate protein–protein interaction like the present study using Stress-responsive transcription Factors DataBase (STIFDB), PubMed and Gene Expression Omnibus (GEO). They found 3091 genes differentially up-regulated during 14 different abiotic stresses, namely abscisic acid, aluminum, cold, cold–drought–salt, dehydration, drought, heat, iron, light, NaCl, osmotic stress, oxidative stress, UV-B and wounding. In the present study, only four out of the 14 abiotic stress signals, Slat, Cold, Drought and ABA were studied. However, Naika and his colleagues (2013) kept the up-regulated gene-group under an ID and didn't go further to gene level to find out the specific gene and their specific interaction. But in the present study such interconnectome was done in *Arabidopsis thaliana.*

In the present study, the eight genes which found to be common during different abiotic stresses in *Arabidopsis thaliana* are DREB2A (Transcription Factors); P5CS1 (Enzyme); CPL1 (Transcription Factors); ERD5 (Transcription Factors); NHX1 (Vacuolar Antiporter); SOS1 (Plasma Membrane Antiporter); SOS2 (Protein Kinase) and SOS3 (Calcium-dependent Protein Serine). All these genes are well documented for their role in abiotic stress tolerance.

Yoh Sakuma *et al.* (2006) reported that DREB2A is one of the most important cis-acting dehydration- responsive elements (DRE) sequence and activates expression of downstream genes involved in drought and salt-stress response in *Arabidopsis thaliana*. Through transgenic *Arabidopsis* it has been revealed that by using microarray data analysis over expressing of transgenic it also showed activity in heat-shock related gene. Gyongyi Szekely *et al.* (2007) described that P5CS1 is an enzyme that catalyse the rate-limiting step of proline biosynthesis, abiotic stress response, are encoded by two closely related P5CS genes in *Arabidopsis*. Transcription of the P5CS genes is differentially regulated by different environmental stresses, like, drought, salinity and abscisic acid, suggesting that these genes play specific roles in the control of proline biosynthesis thus in tolerance towards these stresses (Emre Aksoy *et al.* 2013) proposed that in *Arabidopsis thaliana* enzymes CPL1 regulates transcriptional responses to multiple environmental stresses including osmotic-stress/abscisic acid and iron deficiency stress. Another gene proposed by Murilo Siqueira Alves and Luciano Gomes Fietto (2013) is ERD5. This is defined as one of the genes which are rapidly up-regulated during drought stress. They encoded the protein sequence which shows a great structural and functional diversity and constitutes the defense mechanism against abiotic stress in plant species. Jordan B Sottosanto *et al.* (2007) showed that NHX1 encodes vacuolar sodium or proton antiporter and it is involved in salt tolerance, ion homeostasis, and leaf development and also acts in low affinity electroneutral exchange of protons for cations, such as, Na (+) or K (+) across membranes. It can also exchange Li (+) and Cs (+) with a lower affinity. This gene is found to be engaged in vacuolar ion compartmentalization which is necessary for cell volume regulation and cytoplasmic Na (+) detoxification.

Though individually these genes have been reported to be up-regulated during abiotic stress tolerance, in the present study, further prediction of their interaction was investigated. They were found not only to be up-regulated but also found to create a bridging network along with several genes which are not directly connected but with those eight proteins they produce a network web to give response to different abiotic stress.

Software based investigations show that regulatory network of different type of genes can be explicated indicating their molecular function and biological processes. So, in future attempts need to be taken in the wet bench to analyze their activity in-total to have an in-depth idea of their actual activity under stress condition so that it could bring some answers to the farmers in the crop sector as well as in the nature.

## References


- Adams M D *et al.* (2000)., The Genome Sequence of *Drosophila melanogaster*., Sciencewww.sciencemag.org., Vol. 287 no. 5461 pp. 2185-2195., DOI: 10.1126/science.287.5461.2185.

- Aksoy E, Jeong I S and Koiwa H (2013)., Loss of Function of Arabidopsis C-Terminal Domain Phosphatase-Like1 Activates Iron Deficiency Responses at the Transcriptional Level., Plant Physiology: Vol. 161, pp. 330–345,

- Alon U (2003)., Biological networks: the tinkerer as an engineer., *Science*: 301, 1866–7.

- Bray E A, Bailey-Serres J and Weretilnyk E (2000)., Responses to abiotic stresses., In Biochemistry and Molecular Biology of Plants, Buchanan BB, Gruissem W, and Jones RL, eds (Rockville, MD: American Society of Plant Biologists)., pp. 1158–1203.

- Barry Smith *et al*., (2007)., The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration., Nat Biotechnol., 25(11): 1251. doi: 10.1038/nbt1346.

- Bartels D and Sunkar R (2005)., Drought and salt tolerance in plants., Crit Rev Plant Sci: 24: 23–58.

- Bhalla R, Narasimhan K and Swarup S. (2005)., Metabolomics and its role in understanding cellular responses in plants., Plant Cell Rep: 24(10): 562- 571.

- Boone C *et al.* (2007)., Exploring genetic interactions and networks with yeast., Nat. Rev. Genet., **8**, 437–449.

- Brady S M, Zhang L, Megraw M, Martinez N J, Jiang E, Yi C S, Liu W, Zeng A, Taylor-Teeples M, Kim D, Ahnert S, Ohler U, Ware D, Walhout A J M and Benfey P N (2011)., A stele enriched gene regulatory network in the Arabidopsis root., Mol Syst Biol: 7(459).

- Chen H, Hwang J E, Lim C J, Kim D Y, Lee S Y and Lim C O (2010)., Arabidopsis DREB2C functions as a transcriptional activator of HsfA3 during the heat stress response., Biochem Biophys Res Commun: 401(2), 238-44.

- Chinnusamy V, Schumaker K, Zhu J (2004)., Molecular genetics perspectives on cross-talk and specificity in abiotic stress signaling in plants., Journal of Experimental Botany., 55: 225-236.

- Cline M S *et al*., (2007)., Integration of biological networks and gene expression data using Cytoscape., Nature Protocol., 2(10):2366-82.

- Date S V and Marcotte E M (2003)., Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages., Nat Biotechnol., 21(9): 1055-1062.

- David W. Mount (2004)**.,** Adapted from "Sequence Database Searching for Similar Sequences," Chapter 6, in Bioinformatics: Sequence and Genome Analysis*,* 2nd edition., by David W. Mount. Cold Spring Harbor Laboratory Press., Cold Spring Harbor., NY, USA.

- Djamei A, Pitzschke A, Nakagami H, Rajh I and Hirt H (2007)., Trojan horse strategy in Agrobacterium transformation: abusing MAPK defense signaling., Science: 318(5849), 453-6.

- Dong Q, Kroiss L, Oakley F D, Wang B B and Brendel V (2005)., Comparative EST analyses in plant systems., Methods Enzymol., 395: 400- 418.

- Franceschini A *et al*., (2013)., STRING v9.1: protein-protein interaction networks, with increased coverage and integration., Nucleic Acids Res: 2013 Jan;41 (Database issue): D808-15. doi: 10.1093/nar/gks1094. Epub 2012 Nov 29.

- Godoy M, Franco-Zorrilla J M, Perez-Perez J, Oliveros J C, Lorenzo O and Solano R (2011)., Improved protein-binding microarrays for the identification of DNA-binding specificities of transcription factors., Plant J: 6(4), 700-11.

- Handcock M and Gile K (2010)., Modeling social networks from sampled data., The Annals of Applied Statistics.

- Harbison C T, Gordon D B, Lee T I, Rinaldi N J, Macisaac K D, Danford T W, Hannett N M, Tagne J B, Reynolds D B, Yoo J, Jennings E G, Zeitlinger J, Pokholok D K, Kellis M, Rolfe P A, Takusagawa K T, Lander E S, Gifford D K, Fraenkel E and Young R A (2004)., Transcriptional regulatory code of a eukaryotic genome., Nature: 431(7004), 99-104.

- Hardtke C S, Gohda K, Osterlund M T, Oyama T, Okada K and Deng X W (2000)., HY5 stability and activity in arabidopsis is regulated by phosphorylation in its COP1 binding domain., The EMBO Journal: 19(18), 4997-5006.

- Ho Y, Gruhler A, Heilbut A, Bader G D, Moore L, Adams S L, Millar A, Taylor P, Bennett K, Boutilier K, Yang L, Wolting C, Donaldson I, Schandorff S, Shewnarane J, Vo M, Taggart J, Goudreault M, Muskat B, Alfarano C, Dewar D, Lin Z, Michalickova K, Willems A R, Sassi H, Nielsen P A, Rasmussen K J, Andersen J R, Johansen L E, Hansen L H, Jespersen H, Podtelejnikov A, Nielsen E, Crawford J, Poulsen V, S?rensen B D, Matthiesen J, Hendrickson R C, Gleeson F, Pawson T, Moran M F., Durocher D, Mann M, Hogue C W, Figeys D and M Tyers (2002)., Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry., Nature., 415:180–183.

- Kaufmann K, Muino J M, Osteras M, Farinelli L, Krajewski P and Angenent G C (2010)., Chromatin immunoprecipitation (ChIP) of plant transcription factors followed by sequencing (ChIP-SEQ) or hybridization to whole genome arrays (ChIP-CHIP)., Nat Protoc: 5(3), 457-72.

- Kreps J A, Wu Y, Chang H S, Zhu T, Wang X, Harper J F (2002)., Transcriptome changes for Arabidopsis in response to salt, osmotic, and cold stress., Plant Physiol 130: 2129–2141.

- Li J, Li X, Su H, Chen H, and Galbraith D W (2006)., A Framework of Integrating Gene Relations from Heterogeneous Data Sources: An Experiment on *Arabidopsis thaliana*., *Bioinformatics* (2006) 22 (16): 2037-2043.,doi:10.1093/bioinformatics/btl345.

- Lodish H, Berk A, Zipursky S L, Matsudaira P, Baltimore D and Darnell J., Molecular Cell Biology, 4th edition., New York: W. H. Freeman; 2000. ISBN-10: 0-7167-3136-3.

- Lopez-Molina L, Mongrand S and Chua N H (2001)., A postgermination developmental arrest checkpoint is mediated by abscisic acid and requires the ABI5 transcription factor in Arabidopsis., PNAS: 98(8), 4782-7.

- Mao G, Meng X, Liu Y, Zheng Z, Chen Z and Zhang S (2011)., Phosphorylation of a WRKY transcription factor by two pathogen-responsive MAPKs drives phytoalexin biosynthesis in Arabidopsis., The Plant Cell: 23(4), 1639-53.

- McDermott J, Bumgarner R and Samudrala R (2005)., Functional annotation from predicted protein interaction networks., Bioinformatics., 21(15): 3217- 3226.

- J. Montojo *et al*., (2010)., GeneMANIA Cytoscape plugin: fast gene function predictions on the desktop., Bioinformatics., 26 (22):2927-2928.doi:10.1093/bioinformatics/btq562.

- Morett E, Korbel J O, Rajan E, Saab-Rincon G, Olvera L, Olvera M, Schmidt S, Snel B and Bork P (2003)., Systematic discovery of analogous enzymes in thiamin biosynthesis., Nat Biotechnol., 21(7): 790-795.

- Morohashi K, Xie Z and Grotewold E (2009)., Gene-specific and genome-wide ChIP approaches to study plant transcriptional networks., Methods Mol Biol: 553, 3-12.

- Naika M, Shameerz K and Sowdhamini R (2013)., Comparative analyses of stress-responsive genes in *Arabidopsis thaliana*: insight from genomic data mining, functional enrichment, pathway analysis and phenomics., Mol. BioSyst., 2013, **9**, 1888-1908.

- Noyes M B, Christensen R G, Wakabayashi A, Stormo G D, Brodsky M H and Wolfe S A (2008). Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites., Cell: 133(7), 1277-89.

- Oliphant A R, Brandl C J and Struhl K (1989)., Defining the sequence specificity of DNA-binding proteins by selecting binding sites from random-sequence oligonucleotides: analysis of yeast GCN4 protein. Mol Cell Biol: 9(7), 2944-9.

- Oliveros, J C (2007)., VENNY., An interactive tool for comparing lists with Venn Diagrams., http://bioinfogp.cnb.csic.es/tools/venny/index.html.

- Rensink W A and Buell C R (2005)., Microarray expression profiling resources for plant genomics., Trends Plant Sci., 10(12): 603-609.

- Rustici. G *et al*., (2013)., ArrayExpress update-trends in database growth and links to data analysis tools., Nucleic Acids Res: doi: 10.1093/nar/gks1174. Pubmed ID: 23193272.

- Sakuma Y, Maruyama K, Qin F, Osakabe Y, Shinozaki K and Yamaguchi-Shinozaki K(2006)., Dual function of an Arabidopsis transcription factor DREB2A in water-stress-responsive and heat-stress-responsive gene expression., PNAS: vol. 103, 18822–18827, No 49.

- Schneider T D and Stephens R M (1990)., Sequence logos: a new way to display consensus sequences. Nucleic Acids Res: 18(20), 6097-100.

- Seki M, Narusaka M, Ishida J, Nanjo T, Fujita M, Oono Y, Kamiya A, Nakajima M, Enju A and Sakurai T (2002)., Monitoring the expression profiles of ca 7000 Arabidopsis genes under drought, cold and high salinity stresses using a full-length cDNA microarray., Plant J., 31: 279–292.

- Strong M, Mallick P, Pellegrini M, Thompson M J and Eisenberg D (2003)., Inference of protein function and protein linkages in *Mycobacterium tuberculosis* based on prokaryotic genome organization: a combined computational approach., Genome Biol., 4(9): R59.

- Szekely G, Abraham E, Cseplo A, Rigo G, Zsigmond L, Csiszar J, Ayaydin F, Strizhov N, Jasik J, Schmelzer E, Koncz C and Szabados L (2008)., Duplicated P5CS genes of Arabidopsis play distinct roles in stress regulation and developmental control of proline biosynthesis., The Plant Journal: 53, 11–28., doi: 10.1111/j.1365-313X.2007.03318.x.

- Tran L S P, Nakashima K, Sakuma Y, Osakabe Y, Qin F, Simpson S D, Maruyama K, Fujita Y, Shinozaki K and Yamaguchi-Shinozaki K (2007)., Co-expression of the stress-inducible zinc finger homeodomain ZFHD1 and NAC transcription factors enhances expression of the ERD1 gene in Arabidopsis., Plant J: 49(1), 46-63.

- Uetz P, Giot L, Cagney G, Mansfield T A, Judson R S, Knight J R, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, and Rothberg J M (2010)., A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*., Nature., 403: 623–627.

- Warde-Farley D *et al*. (2010)., The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function., Nucleic Acids Res. 2010 Jul;38(Web Server issue):W214-20. doi: 10.1093/nar/gkq537.

- Walhout A J M (2006)., Unraveling transcription regulatory networks by protein-DNA and protein-protein interaction mapping., Genome Res: 16(12), 1445-54.

- Weltmeier F, Ehlert A, Mayer C S, Dietrich K, Wang X, Schutze K, Alonso R, Harter K, Vicente-Carbajosa J and Droge-Laser W (2006)., Combinatorial control of Arabidopsis proline dehydrogenase transcription by specific heterodimerisation of bZIP transcription factors., The EMBO Journal: 25(13), 3133-43.

- Wichadakul D, McDermott J and Samudrala R (2007)., Prediction and integration of regulatory and protein-protein interactions., Computational Systems Biology. J., Mcdermott R, Ireton K, Montgomery R, Bumgarner and Samudrala R, Humana Press: [in press].

- Xiong L, Schumaker K S, Zhu J K (2002)., Cell signaling during cold, drought, and salt stress., Plant Cell (Suppl)., 14: S165–S183.

- Yamada T and Bork P (2009)., Evolution of biomolecular networks: lessons from metabolic and protein interactions., Nature Reviews Molecular Cell Biology., 10: 791–803.

- Yu H, Luscombe N M, Lu H X, Zhu X, Xia Y, Han J D, Bertin N, Chung S, Vidal M and Gerstein M (2004). Annotation Transfer Between Genomes: Protein-Protein Interologs and Protein-DNA Regulogs., Genome Res., 14(6): 1107-1118.

- Zhu Q, Zhang J, Gao X, Tong J, Xiao L, Li W and Zhang H (2010)., The Arabidopsis AP2/ERF transcription factor RAP 2.6 participates in ABA, salt and osmotic stress responses., Gene: 457(1-2), 1-12.

## Appendices

| Browser Name | Web Address |
|---|---|
| **ArrayExpress** | http://www.ebi.ac.uk/arrayexpress/ |
| **Venny** | http://bioinfogp.cnb.csic.es/tools/venny/ |
| **European Bioinformatics Institute** | http://www.ebi.ac.uk/ |
| **String Database 9.02** | http://string-db.org/ |
| **National Center for Biotechnology Information (NCBI)** | http://www.ncbi.nlm.nih.gov/ |
| **Basic Local Alignment Search Tool (BLAST)** | http://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch&PROG_DEF=blastn&BLAST_PROG_DEF=megaBlast&SHOW_DEFAULTS=on&BLAST_SPEC=OGP__3702 |
| **Interproscan** | http://www.ebi.ac.uk/Tools/pfa/iprscan/ |
| **Gene Ontology** | http://www.geneontology.org/ |
| **GeneMANIA** | http://www.genemania.org/ |