Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

STIF: Identification of stress-upregulated transcription factor binding sites in *Arabidopsis thaliana*

Ambika Shyam Sundar[1], Susan Mary Varghese[2], Khader Shameer[1], Nataraja Karaba[2], Makarla Udayakumar[2] and Ramanathan Sowdhamini[1,*]

[1]National Centre for Biological Sciences, Tata Institute of Fundamental Research, UAS-GKVK Campus, Bellary Road, Bangalore 560 065, India;
[2]Department of Crop Physiology, UAS-GKVK Campus, Bellary Road, Bangalore 560 065, India;
Ramanathan Sowdhamini* - E-mail: mini@ncbs.res.in; * Corresponding Author

received May 22, 2008; revised July 12, 2008; accepted July 14, 2008; published July 30, 2008

Abstract:
The expressions of proteins in the cell are carefully regulated by transcription factors that interact with their downstream targets in specific signal transduction cascades. Our understanding of the regulation of functional genes responsive to stress signals is still nascent. Plants like *Arabidopsis thaliana*, are convenient model systems to study fundamental questions related to regulation of the stress transcriptome in response to stress challenges. Microarray results of the *Arabidopsis* transcriptome indicate that several genes could be upregulated during multiple stresses, such as cold, salinity, drought etc. Experimental biochemical validations have proved the involvement of several transcription factors could be involved in the upregulation of these stress responsive genes. In order to follow the intricate and complicated networks of transcription factors and genes that respond to stress situations in plants, we have developed a computer algorithm that can identify key transcription factor binding sites upstream of a gene of interest. Hidden Markov models of the transcription factor binding sites enable the identification of predicted sites upstream of plant stress genes. The search algorithm, STIF, performs very well, with more than 90% sensitivity, when tested on experimentally validated positions of transcription factor binding sites on a dataset of 60 stress upregulated genes.

Availability: Supplementary data is available at http://caps.ncbs.res.in/download/stif

Keywords: transcription factor binding site prediction; gene regulation; stress genes; *Arabidopsis thaliana*; HMM based algorithm

Background:
The interactions between regulatory proteins and DNA control many important processes and responses to environmental stresses, and defects in these interactions can contribute to inefficient stress responses. Plants, like *Arabidopsis thaliana*, are very simple and good model systems to understand fundamental processes such as protein-DNA interactions that happen in response to environmental stresses **[1, 2]**. Numerous studies have shown that transcription factors are important in regulating plant responses to stress. One important step in the control of stress responses is the transcriptional activation or repression of genes. Databases, such as ATHAMAP **[3]**, offer information about the chromosomal positions of genes of interest and possible location of their transcription factors and binding sites. Multiple signalling pathways regulate the stress responses of plants and there is significant overlap between the patterns of gene expression that are induced in plants in response to different stresses **[4]**. Many genes induced by stress challenges, including those encoding transcription factors, have been identified and some of them have been shown to be essential for stress tolerance. Many studies have also revealed some of the complexity and overlap in the responses to different stresses, and are likely to lead to new ways to enhance crop tolerance to disease and environmental stress. The binding specificities of only a small number of transcription factors (TFs) are well characterized. Transcription-factor binding sites (TFBSs) are usually short (around 5-15 base-pairs (bp)) and they frequently contain degenerate sequence motifs. The sequence degeneracy of TFBSs has been selected through evolution and is beneficial, because it confers different levels of activity upon different promoters. Much of the information on TF binding specificity has been determined using traditional methodologies, such as foot-printing methods, (that identify the region of DNA protected by a bound protein), nitrocellulose binding assays, gel-shift analysis (that monitors the change in mobility when DNA and protein bind), South-western blotting (of both DNA and protein) or reporter constructs. These methods are generally quite time-consuming and are not readily scalable to a whole genome **[5]**. One of the interesting problems is to identify the *cis*-acting elements by computational techniques at a whole genome level so as to choose promising targets for detailed experimental investigation. Well-known eukaryotic transcription factors and their binding sites are recorded in

TRANSFAC database **[6]**. There are computational tools to facilitate the retrieval of information from TRANSFAC database, but for the human genome **[7]**. There have also been algorithms that employ position-specific profiles and scoring schemes to recognize putative TFBS **[8, 9]** or probabilistic models **[10]**. These servers and algorithms are largely for eukaryotic general-purpose transcription factors and not specific for plant stress induced genes. There are other computational algorithms to search for possible genes that are downstream of classical TFBS, where the binding site data are encoded as HMMs and searched all around the genome of interest. These methods are called as 'targeted gene finding' since they begin from known TFBS **[11]**. However, this approach is complicated for plant stress genes since stress TF-binding site signatures could potentially be upstream of constitutive genes as well and there could also be overlap in various TFBS. We have collected data of well-known stress specific transcription factors and generated Hidden Markov Method (HMM) of known TFBS. This knowledge-based approach, by building HMM models through well-known abiotic stress *cis*-elements, has been tested extensively to standardize thresholds for scores.

Methodology:
In Arabidopsis thaliana, we have examined 10 families of transcription factors known to be involved in responses to abiotic stress (Table 1 under supplementary material).

Dataset for validation
We have identified 60 stress responsive genes from six different microarray databases and these were collected on the basis of their consistent upregulation in response to abiotic stress signals in most of these microarray databases and across different microarray experiments. To compare the *cis*-elements both in stress on-off conditions, we also identified 60 constitutive genes from six different databases. Genes that get consistently upregulated under abiotic conditions were identified from these databases and used for the validation study.

RARGE
RARGE **[12]** presents Arabidopsis resource data (cDNAs, transposon mutants and microarray experiments) for all biology researchers. RARGE has data from 6 different abiotic stress experiments (i.e. cold, drought, salt, ABA, Light, dehydration stress) with expression levels at different time courses.

DRASTIC
DRASTIC **[13]** is a database of plant genes regulated in response to biotic and abiotic stress, developed and maintained by the Scottish Crop Research Institute.

StressLink
StressLink**[14]** is a database linking genome information to the primary literature on stress-related genes in Arabidopsis thaliana.

AtGenExpress
AtGenExpress **[15]** is a multinational effort designed to uncover the transcriptome of the multicellular model organism Arabidopsis thaliana **[15]**.

DATF
The Database of Arabidopsis Transcription Factors (DATF) **[16]** collects all Arabidopsis transcription factors and classifies them into 64 families. It uses not only locus (gene), but also gene model (transcript, protein) and the detail information is for each gene model not for locus. It adds multiple alignment of the DNA- binding domain of each family.

TAIR
The Arabidopsis Information Resource (TAIR) **[17]** maintains a database of genetic and molecular biology data for the model higher plant Arabidopsis thaliana. Data available from TAIR includes the complete genome sequence along with gene structure, gene product information, metabolism, gene expression, DNA and seed stocks, genome maps, genetic and physical markers, publications, and information about the Arabidopsis research community.

Construction of a Hidden Markov Model
Hidden Markov Model (HMM) is a probabilistic method, which is used for TFBS detection. The consensus (S) of length (L) was taken from the literature and the probabilistic score (P(S)) and log-odd score were calculated by using equation (1) and (2) (see supplementary material) respectively.

As plant sequences are rich in GC content, we gave higher weight to AT than GC in log-odd score (please see Figure 1 for an example).

STIF - TFBS search algorithm
The search program will accept nucleotide stretch that is upstream of an abiotic stress gene. A detailed flow-chart of the search algorithm is given in Figure 2. The TFBS search algorithm searches for cis-elements both in forward and reverse direction in the query sequence from the built models and the acquired hit gets a HMM score. The chromosomal position, UTR position, cis-element, orientation of the cis-element are also recorded and mentioned in the output.

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

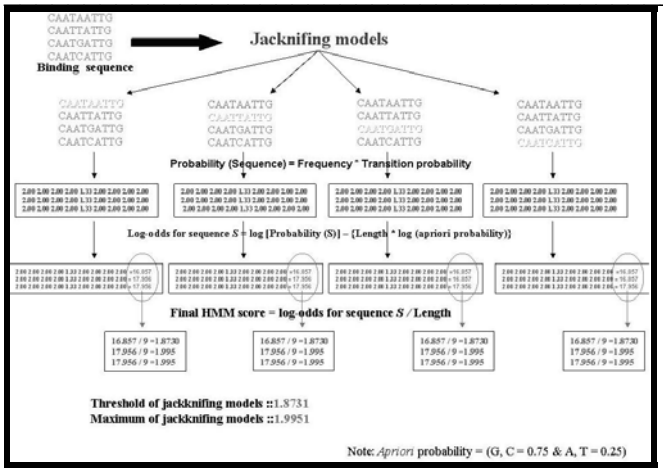

Figure 1: Construction of a Hidden Markov Model of transcription factor binding sites given the experimentally observed nucleotide patterns.

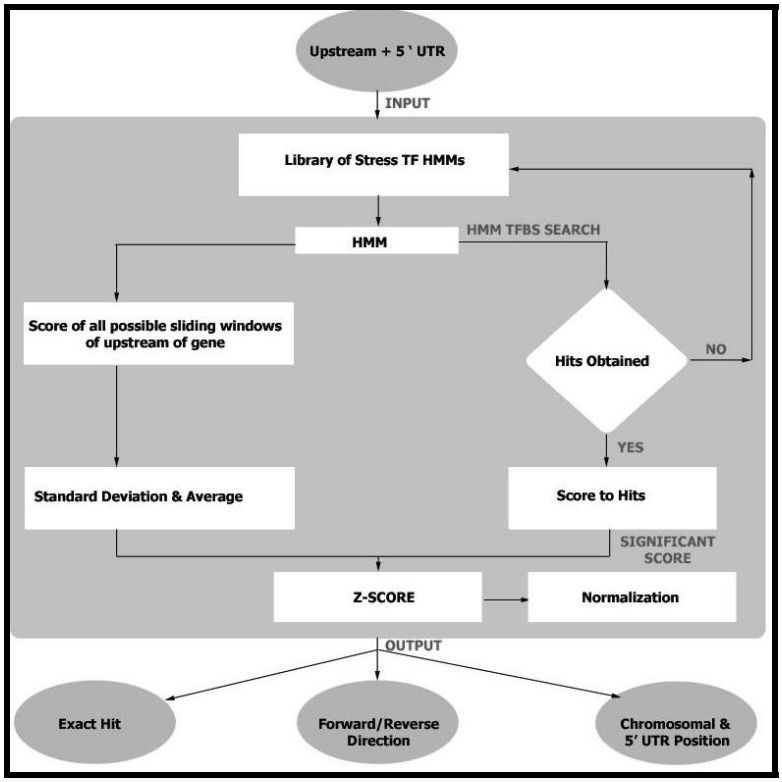

Figure 2: Flow chart diagram of STIF search algorithm.

Validation

Validation was performed using a leave-one-out approach (Jackknifing method). The threshold and maximum scores of each HMM model was further decided by this statistical test of jackknifing (Figure 1). The parameters (coverage, sensitivity and specificity) chosen for the statistical tests were calculated using the equations (3), (4) and (5) given under supplementary material.

Implementation

We have used Perl based programs exclusively developed for STIF algorithm to perform HMM related computation, searching, calculation of statistics and input - output parsing. Scripts for parsing, searching, statistics and other calculations like Z-Score and normalization are coded for the STIF algorithm in Perl as in equation (6) and (7) respectively (under supplementary material). The source code is available from the corresponding author upon request.

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

Discussion:

A computational method, STIF, has been developed to search for potential transcription factor binding sites of stress-specific transcription factors, starting from Hidden Markov Models of nucleotide binding site patterns of cis-elements that are well-known to respond during stress situations in plants. The 19 models of cis-elements, based on abiotic stress transcription factor families, were built as Hidden Markov Models and were validated using Jackknifing method. We had applied our HMM-based search algorithm, STIF, to search 100 base pairs upstream of the gene with its 5′UTR. We identified 60 abiotic stress genes from well-known microarray databases based on the high stress-induced expression profiles. These genes were known to be upregulated during stress and their validated TFBS information is also clearly available. To evaluate the method further, we also searched against 1000 base pairs with its 5′UTR.

In our validation set, at a Z-score of 2.0 when searched 100 base pairs with 5′UTR, the sensitivity of the method is very high, since we identify 18 out of 20 hits (95% coverage) with only two false negatives Table 1 (see supplementary material). We therefore propose that a Z-score of 2.0 or more could be effective in not missing out the associated TFBS when searched for 100 base pairs with 5′UTR. In several instances, more than one transcription factor has been recorded for a stress gene of interest (for instance, COR15a has both DREB_AP2_EREBP and G_ABRE_bZIP (Figure 3a and Table 1 in supplementary material). The 60 stress genes that we have considered for validation are known to be upregulated during different types of stress - such as cold, dehydration, salinity etc. It is possible that, during a particular type of stress, any one of these transcription factors would selectively respond by binding upstream of the gene of interest.

We also noticed that there are very few 'validated' TFBS which are mapped 100 base pairs upstream of stress genes. Therefore, we extended this validation to searches 1000 base pairs upstream of the gene and likewise a Z-score threshold of 1.5 is appropriate for 1000 base pairs with 5′UTR (Figure 3b and Table 2 under supplementary material). 90% sensitivity is achieved in STIF, where 71 out of 78 hits could be correctly identified with Z-scores above the threshold. As with most other algorithms, we notice that the method is not highly specific and can generate false positives. The specificities for searches in the validation set, by searching 100 base pairs and 1000 base pairs, is 57 and 18.6 (for Z-score threshold o1.5) and 54 and 20.4 (for Z-score threshold of 2.0), respectively. The difficulty in obtaining high specificities has been due to simple and short nucleotide patterns that describe some of the transcription factors like bHLH. Such TFs, would respond frequently and that too with very good match with HMM and are reflected as high scores. We have proposed an alternate normalized score for these frequently responding TFs. STIF employs Hidden Markov Models of binding site information of well-known plant transcription factors in stress. Microarray results of key stress upregulated genes in plants have shown that a large number of these genes are upregulated in response to a variety of genes generating redundancy in the dataset of stress upregulated plant genes. Further, the experimentally 'validated' results also indicate that more than one transcription factor can turn ON the stress genes in our dataset. The scoring schemes and thresholds established should be useful for dealing with redundancy and occurrence of multiple true positives.

We have built each HMM model and provided a stringent threshold for the scoring schemes. STIF algorithm, along with its database, is highly specific for plant stress cis-elements. However, this can be easily applied and extended to general systems after updating the HMM library and carefully standardizing the scoring scheme thresholds. The availability of such sensitive and specialized search algorithms can be very useful for addressing particular biological problems.

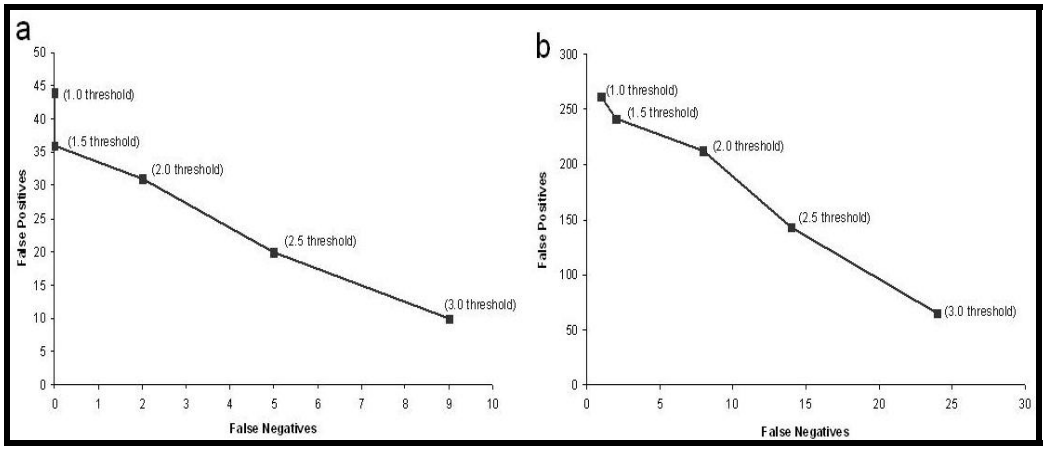

Figure 3: (a) The validation set of 11 stress responsive genes when searched for 100 base pairs with its 5′UTR with 11 stress responsive genes. The total number of false positives obtained during the search is compared against the total number of false negatives for various Z-score thresholds applied for the statistical tests. (b) Same as Figure 3a but for a validation set of 29 stress genes where search for TFBS was performed 1000 base pairs with its 5′UTR.

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

Conclusion:
Computational Transcription Factor Binding Site (TFBS) prediction is a mature domain in the field of Bioinformatics. Various algorithms, stand-alone software and web servers are available for the effective prediction of transcription start from sequence information using knowledge based and motif based methods **[5, 18]**. A wide array of TFBS prediction programs are available based on different biological contexts. For example a novel method for prokaryotic promoter prediction based on DNA stability that utilises structural properties of DNA is developed and analysed across different organisms **[19]**, time-delay neural network based method (NNPP), is available specifically for the analysis of *Drosophila melanogaster* promoter regions **[20]**. An HMM based method based on markov chain optimization is available for the identification of hepatocyte nuclear factor 4-alpha in human genome **[21]**. Due to availability of such generic as well as specific TFBS prediction algorithms specific algorithms for prediction of transcription factor binding sites, users are recommended to use multiple programs to obtain a consensus result. STIF algorithm explained in this manuscript which uses HMM models of known Abiotic stress factors will be useful for further analysis and understanding of stress gene regulation in the plant model system *Arabidopsis thaliana.* Since no bioinformatics tool provides a complete solution for the transcription factor identification problem, it is always better to analyse the promoter regions with more than one algorithm or program that based on the biological context.

Acknowlegement:
The project is funded by a grant from the Department of Biotechnology, India. Susan Mary Varghese acknowledges the CSIR for the Senior Research Fellowship during the course of this research work. We thank UAS and NCBS (TIFR) for infrastructural support.

References:

[01] The Arabidopsis genome initiative, *Nature,* 408: 796 (2000) [PMID: 11130711]
[02] E. Lander *et al.*, *Nature,* 409: 860 (2001) [PMID: 11237011]
[03] N. O. Steffens *et al.*, *Nucleic Acids Res.,* 1: D368 (2004) [PMID: 14681436]
[04] K. Singh *et al.*, *Curr Opin Plant Biol.,* 5: 430 (2002) [PMID: 12183182]
[05] M. L. Bulyk *et al.*, *Genome Biol.,* 5: 201 (2003) [PMID: 14709165]
[06] E. Wingender *et al.*, *Nucleic Acids Res*., 29: 316 (2001) [PMID: 10592259]
[07] H. Zhang *et al.*, *Nucleic Acids Res*., 30: e121 (2002) [PMID: 12409480]
[08] A. Sandelin *et al.*, *Nucleic Acids Res.,* 1: D91 (2004) [PMID: 14681366]
[09] D. E. Schones *et al.*, *Bioinformatics,* 21: 307 (2005) [PMID: 15319260]
[10] R. Pudimat *et al.*, *Bioinformatics,* 21: 3082 (2005) [PMID: 15905283]
[11] W. Zhang *et al.*, *Bioinformatics,* 21: 3074 (2005)[PMID: 15890746]
[12] T. Sakurai *et al.*, *Nucleic Acids Res.,* 1: D368 (2005) [PMID: 14681436]
[13] D. K. Button *et al.*, *Nucleic Acids Res.,* 1: D712 (2006) [PMID: 16381965]
[14] G. J. Warren, *Curr Biol*., 8: R514 (1998) [PMID: 9705923]
[15] M. Schmid *et al.*, *Nat Genetics.,* 37: 501 (2005) [PMID: 15806101]
[16] A. Guo *et al.*, *Bioinformatics,* 21: 2568 (2005) [PMID: 15731212]
[17] S. Y. Rhee *et al.*, *Nucleic Acids Res*., 31: 224 (2003) [PMID: 12519987]
[18] M. Tompa *et al.*, *Nat. Biotechnol.,* 23: 137 (2005) [PMID: 15637633]
[19] A. Kanhere and M. Bhansal, *BMC Bioinformatics,* 6: 1 (2005) [PMID: 15631638]
[20] M. G. Reese, *Comput Chem.,* 26: 51 (2001) [PMID: 11765852]
[21] K. Ellrott *et al.*, *Bioinformatics,* 18: S100 (2002) [PMID: 12385991]

Edited by P. Kangueane
Citation: Ambika *et al.*, Bioinformation 2(10): 431-437 (2008)

Supplementary material

$$P(S) = F * T \quad \rightarrow \quad \textbf{(1)}$$

where P(S) – Probability of consensus
F – frequency (i.e. No: of particular nucleotide/ Total no in column)
T – transition probability

Log odd-score for consensus

$$(S) = \log P(S) - L\,(AT) \log 0.375 + L(GC) \log 0.125 \quad \rightarrow \quad \textbf{(2)}$$

$$\text{Coverage} = \frac{TP}{\text{Total Number of hits}} \rightarrow \quad (3)$$

TP = Hits acquired which is equal to experimental validation + greater than threshold value of the dataset.
Total number of hits = Total number of hits acquired which is equal to experimental validation.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \rightarrow \quad (4)$$

TP = True Positive
FN = False Negative (total hits occurring below threshold value)

$$\text{Specificity} = \frac{TP}{TP + FP} \rightarrow \quad (5)$$

TP = True Positive
FP = False Positive (total hits occurring above threshold value)

$$\text{Z-Score} = \frac{\text{Score - Mean}}{\text{Standard Deviation}} \rightarrow \quad (6)$$

Z – Z-score
score – HMM score of the acquired hit
mean – average of all possible sliding windows of upstream of stress gene
std deviation – Standard Deviation of all possible sliding windows of upstream of stress gene.

Normalization score

$$\text{The normalization formula is} = \frac{\dfrac{\text{Top 1st rank of z-score of binding site for that TFBS and that stress gene}}{\text{Total No: of binding sites for that TFBS}}}{\dfrac{\text{Total no: of binding sites for all TFBS library and stress gene}}{\text{Total no: of binding for all TFBS library and of all stress genes}}} \rightarrow \quad (7)$$

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

Tables

S. No	Family name	Sub-family	Stress signal	Reference (Stress signal)	Name of the Cis-element	Cis-element	Reference (Cis-element)
1	ABI3/ VP1		ABA	Plant J. 2000; 24(1):57-66	distB ABRE	GCCACTTGTC	Plant J. 2000; 24(1):57-66
2	AP2/ EREBP	EREBP-ERF	Cold, Drought	The Plant Cell, 1998; 10:1391–1406.	GCC-box	GCCGCC	The Plant Cell, 1998; 10: 1391–1406.
		DREB	Cold, Drought	Proc. Natl. Acad. Sci., 1997; 94:1035-1040	CRT/DRE	(A/G)CCGAC	Proc. Natl. Acad. Sci., 1997, 94:1035-1040
3	ARF		Auxin	PNAS, 1999; 96(10): 5844-9	AuxREs	TGTCTC	PNAS, 1999; 96(10): 5844-9
4	BHLH/ myc		NACL, ABA, Drought	The Plant Cell, 2003; 15: 63–78	N box	CACG(G/A)C	The Plant Cell, 2003; 15: 63–78
					G box	CACGTG	The Plant Cell, 2003; 15: 1749–1770
5	bZIP		ABA, Drought	Current Opinion in Plant Biology 2000; 3:217–223	G box1	CCACGTGG	The Plant Cell, 1992; 4: 1309-1319
					G box2	TGACG(T/C)	The Plant Cell, 1992; 4: 1309-1319
					G/ABRE	(C/T)ACGTGGC	Journal Of Biological Chemistry, 2000; 275(3): 1723–1730
					C/ABRE	CGCGTG	Journal Of Biological Chemistry, 2000; 275(3): 1723–1730
6	HB		ABA, Drought	*Plant Molecular Biology,* 1998; **37:** 377–384.		CAATNATTG	Nat. Struct Biol, 1999; 6:464-470
7	HSF		Drought, Cold, Heavy-metal stress and oxidative stress	Plant Physiol. 1998; 117: 1135–1141	HSE	TTCNNGAA GAANNTTC	Nat. Struct Biol, 1999; 6:464-470
8	MYB		Dehydration, Wounding	The Plant Cell, 1993; 5:1529-1539		(T/C)AAC(G/T) G	*Genes & Dev.* 1990; 4: 2235-2241
						CC(T/A)ACC	Genetics, 1998; 149: 479–490.
						TAACTG	Plant Journal,1996; 10(6): 1145-1148
						CC(TA)AACC	Genetics, 1998; 149: 479–490.
						(C/T)AACN(A/G)	The Plant Journal, 2003; 33: 259–270
9	NAC		Drought, high salinity and ABA	The Plant Cell, 2004; 16: 2481–2498.		CATGTG	Plant Mol Biol. 2002; 50(2):237-48.
10	WRKY		Biotic stress (pathogen attack) Abiotic Stress (wind, rain, hail)	*Plant Physiology*, 2002, 129: 661–677	W box	(T)TGAC(C/T)	***Plant Molecular Biology* 51: 21–37, 2003.**

Table 1: Abiotic stress responsive transcription factor families.