12-1-2018

ANALYSIS OF THE CIS-REGULATORY ELEMENT LEXICON IN UPSTREAM GENE PROMOTERS OF ARABIDOPSIS THALIANA AND ORYZA SATIVA

Belan Khalil
Southern Illinois University Carbondale, belan.khalil@siu.edu

ANALYSIS OF THE CIS-REGULATORY ELEMENT LEXICON IN UPSTREAM GENE PROMOTERS OF *ARABIDOPSIS THALIANA* AND *ORYZA SATIVA*.

by

Belan M. Khalil

B.S., University of Salahaddin, 2002
M.S., University of Duhok, 2010

A Dissertation
Submitted in Partial Fulfillment of the Requirements for the
Doctor of Philosophy Degree

Department of Plant Biology
in the Graduate School
Southern Illinois University Carbondale
December, 2018

DISSERTATION APPROVAL

ANALYSIS OF THE CIS-REGULATORY ELEMENT LEXICON IN UPSTREAM GENE PROMOTERS OF *ARABIDOPSIS THALIANA* AND *ORYZA SATIVIA*.

by

Belan M. Khalil

A Dissertation Submitted in Partial

Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

in the field of Plant Biology

Approved by:

Matt Geisler, Chair

Andrew Wood

Aldwin Anterola

David Lightfoot

Ahmad Fakhoury

Graduate School
Southern Illinois University Carbondale
July 11, 2018

AN ABSTRACT OF THE DISSERTATION OF

BELAN M. KHALIL, for the Doctor of Philosophy degree in Plant Biology, presented July 11, 2018, at Southern Illinois University Carbondale.

TITLE: ANALYSIS OF THE CIS-REGULATORY ELEMENT LEXICON IN UPSTREAM GENE PROMOTERS OF ARABIDOPSIS THALIANA AND ORYZA SATIVIA.

MAJOR PROFESSOR: Dr. Matt Geisler

Gene expression in plants is partly regulated through interaction of trans-acting factors with the promoter regions of the gene. Trans-acting factor binding sites consist of short nucleotide sequences most often present in the upstream promoter region. These binding sites, the cis-regulatory elements (CREs), vary in structure, complexity, and function. In binding to trans-acting factors CREs connect genes to signaling and regulatory pathways that affect plant growth, development, and response to the environment. As words in a language, CREs and thus promoters can be analyzed by looking for spelling (patterns of nucleotides) associated with meaning (functions). Considering CREs as words in a language, this kind of analysis provides great opportunity for comprehensive understanding of promoter language. Identification and characterization of CREs is challenging either experimentally or bioinformatically, and has previously been accomplished by discovering degenerate words, with ambiguous nucleotides. This kind of result implicitly makes a hypothesis that binding of a specific trans-acting factor is somewhat promiscuous (or sloppy) and that all words represented by a degenerate pattern are equally good at binding. In this study, we unpack the "degeneracy hypothesis" through systematically considering each combination of letters independently for CRE function. Our results demonstrate that not all degenerate combinations of published CREs have the same effect on gene expression. A systematic search and comparison of all 65,536 possible 8 bp CRE words were searched in the 500 bp and 1000 bp upstream promoters of all genes in Arabidopsis thaliana and Oryza sativa, respectively. The function of each CRE was evaluated by statistically comparing the presence or absence of the element in the promoter with that genes response (induction or suppression) to stimuli in 1691 public availability transcriptomes of differential gene expression data. Arabidopsis, a model dicot plant had a much larger number of such data sets, than rice, however rice was chosen as a comparison as it had the largest number of datasets for a monocot, the most distantly related plant group with sufficient data available. A comprehensive list of 8 bp words associated with differential gene expression, linguistically known as lexicon, was retrieved for both species by establishing that the presence of a CRE significantly increased the likelihood for differential expression by at least one stimulus. The lexicons were composed of 641 and 856 CREs respectively in Arabidopsis and rice, and there were only 78 shared CREs between the two lexicons.

The CRE lexicon was then characterized for their strength and breadth of response, occurrence frequency, sequence complexity, and sequence conservation between two species. In Arabidopsis, evening element (EE) showed the strongest response to a cold stress transcriptome (p-value 10^{-99}). In rice, the element AAACCCTA showed strongest response to a tissue specific transcriptome (p-value 10^{-79}). The breadth of response varied between the two species due to number of transcriptomes used in the study. The element AAACCCTA and GCGGCGGA significantly correlated to 197 and 58 transcriptomes in both Arabidopsis and rice, respectively. On the other side of the breadth scale there were also many CREs with very restricted response. There were 291 and 258 CREs in Arabidopsis and rice, respectively, significantly correlated to a single stimulus. Occurrence frequency revealed that the most abundant CREs in Arabidopsis and rice genes were TATA box and TATA box like CREs. The structure of the CREs in the lexicon

was also varied. CREs were distributed on seven levels of complexity. Level one comprised CREs having 8 copies of the same nucleotide, level seven comprised CREs having two copies of the same nucleotide. In Arabidopsis, out of 641 CREs, 314 were of level 6 complexity, which means having 3 copies of the same nucleotide. In rice, the majority of the lexicon, 263 CREs were of level 5 complexity, which means having 4 copies of the same nucleotide.

Each CRE of the lexicon was correlated to at least one experimental condition in the differential gene expression data, but many were correlated to multiple and often related conditions such as drought, temperature and salinity. Therefore, each CRE was assigned a "meaning", i.e. the associated stimuli, thus providing a sort of CRE function dictionary in addition to the lexicon itself. Many CREs possessed different meanings (termed homographs in language), and in many cases the meanings of different CREs overlapped like language synonyms. Sharing meanings (synonyms) was often among CREs with strong sequence similarity (homonyms or homophones), however, not in all cases. Analyzed as a linguistic aspect, CRE homonymity and synonymity was applied to explore the hypothesis "all CRE synonyms are also homonyms and all CRE homonyms are also synonyms." To the end a single CRE was compared to all possible CREs with only one letter mismatch in their sequences are considered as homonyms. The CREs meaning was converted to a matrix of stimuli to generate clusters of synonyms that were analyzed for similarity of spelling (sequence). This analysis showed that not all homonyms are synonyms, however most synonyms are homonyms. Furthermore, despite a search of all one letter mismatches among homonyms, many of the functional homonyms shared smaller 4-5bp core sequence and only varied at the flanks. Synonyms being homonyms in the language of promoters raises a question, how did this evolve? Duplication of transcription factors in the genome generated transcription factor families where

each family member shares the same core domain, usually a DNA recognition site. We here propose that CREs also duplicate during gene duplication process building CRE families in parallel. Members of CRE families may show different connectivity and affinity to individual members of transcription factors in a transcription factor family. In environmental sensors and developmental decision panel, this association of two families of interaction factors is called dense overlapping region (or DOR) and is a highly overrepresented network topology in biological systems. This also explains the degeneracy of initially discovered CREs. The fact is only a portion of nucleotide combinations implied by a degenerate CRE is bioactive, it represents an overlap of different members of a CRE family which is part of the process of family expansion and diversification and done as compensatory mutations as the family of transcription factors expanded and diversified. We also extensively studied CREs involved abiotic stress and identifies shared elements among abiotic stresses as well as abiotic stress specific CREs. Furthermore, CREs follow a time sensitive response rule, which means some CREs participates in gene expression regulation only at certain period during the course of exposure to the abiotic stress.

OBJECTIVES OF THE STUDY

The main objectives of the current study can be summarized to:

1- Analysis and characterization and of the lexicons of CREs of both Arabidopsis and rice. This includes the comparison of the lexicon of the two species, determining frequency abundance, strength, response breadth and complexity of CREs in the lexicon. Furthermore, retrieving the well known CREs in plants like ABRE, DRE, and EE. In addition, discovering novel elements if present.

2- Comparing the strength and response breadth of CREs to test whether stronger CREs possess broader response.

3- Comparing CREs according to their sequence similarity (homonyms) and functional similarity (synonyms). Do CREs with similar sequences have similar expression pattern, or do CREs with similar expression pattern have similar sequences?

4- Determining if larger CREs and smaller CRE cores can be discovered using the 8 bp window of this study. Furthermore, identification of new cores and grouping CREs according to their cores.

DEDICATION

To my parents for being my first teacher and guidance

To my beloved wife for your support and encouragement
through all the years of studying

To my kids, Aryas and Aryan whom I saw hope
from your eyes

To my brothers and sister

ACKNOWLEDGEMENTS

I would like to thank my adviser Dr Geisler for his unlimited support and encouragement during the period of my studying

Thanks for the committee members Dr. Wood, Dr. Fakhoury, Dr. Lightfoot, and Dr. Anterola for their continual consultant and guidance

I would like to thank Dr Geisler-Lee for her advices and management roles in the lab

Many thanks to the chair of Plant Biology Department Dr. Sara Baer for her support during my financial crisis

Special thank to the government of Kurdistan region- Iraq for providing me such great opportunity to pursue my dreams

Thanks to Tony, Alicia, David and Bryan who were once Geisler and Geisler-Lee lab members

Thanks to all my friends and colleagues in Plant Biology Department

TABLE OF CONTENTS

CHAPTER 3 ANALYSIS OF PROMOTER LANGUAGE CHARACTERISTICS IN ARABIDOPSIS THALIANA AND ORYZA SATIVA

CHAPTER 4 ABIOTIC STRESS AND DEVELOPMENTAL CREs

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.0 Background

Plant gene expression is regulated in part by the interaction of trans-acting factors with the upstream promoter, adjacent to the gene being regulated. This promoter region contains many cis-regulatory elements (CREs) that are the sites of interaction with the trans-acting factors. CREs are different in length and in sequence, creating unique binding sites in a pattern unique to each promoter, and thus allowing different trans-acting factors to participate in the regulation of each unique gene. The promoter can be thought of as a set of expression instructions with connections to signaling and regulatory pathways. These instructions are like a language, with each CRE representing a word in that language. Different species may evolve different CREs just as they evolve different trans-acting factors. The closer the species are to each other, the more similar the CREs and their trans-acting factors will likely be. Thus, language of CREs and promoters might be thought of like the spoken languages of peoples; they diverge and evolve, but often share similarities. In this study, the lexicon of all CREs in Arabidopsis and rice is characterized for the 500 and 1000bp upstream promoter respectively of all genes.

1.1 Common known plant CREs

1.1.1 Abscisic acid responsive elements (ABREs)

ABREs are of the most well-known and studied CREs in plants. As its name suggests, ABRE is responsive to alterations in the concentration of abscisic acid (ABA), a plant hormone which plays a critical role in plant growth and development (Finkelstein 2013). ABRE was first found in the *Em* (early methionine) gene when promoter regions of many ABA-inducible genes were compared. ABRE contains an ACGT core, which is known to be the binding site of bZIP transcription factors (Izawa et al. 1993; Uno et al. 2000b; Foster et al. 1994). The structure and sequence of the element might vary from plant to plant. In Arabidopsis, the ABRE element possesses a core of seven letters ACGTG/TC (Yamaguchi-Shinozaki and Shinozaki 1994; Baker et al. 1994), while in rice the CGTACGTGTC is considered as an ABRE element (Tokunori et al. 1999) , whereas in maize the element is GACGTG (Kamp et al. 1997). In wheat and tobacco CACGTGGC and CCACGTGG which are slightly shifted to either side are ABRE elements, respectively (Guiltinan et al. 1990; Oeda et al. 1991). ABRE is widely involved in abiotic stress resistance and various plant growth and developmental aspects. It has been recorded that ABRE plays a critical role in ABA-dependent pathway during drought stress and osmotic stress (Fujita et al. 2005; Yamaguchi-Shinozaki and Shinozaki 2005). During abiotic stress the concentration of ABA increases triggering a series of transcriptional gene expression regulation starting with transcribing one of the many different Abscisic Acid Responsive Element Binding proteins or ABRE binding factors (AREB or ABF) (Fujita et al. 2005). These proteins, which are bZIP family transcription factors, bind to ABRE which eventually lead to transcribing stress tolerance proteins. In order for the ABRE/ABF to regulate gene expression, it needs an ABA-mediated signal, i.e. phosphorylation. In Arabidopsis, SnRK2 type protein, activated by AREB1, acts as a protein kinase that mediates gene expression and sequentially, regulates stomatal closure (Yoshida et al. 2002), which eventually leads to decreased transpiration, water loss, and increased drought tolerance of the plant.

1.1.2 Dehydration responsive elements (DREs)

Another critical and well characterized CRE in plants is DRE. This element was first characterized as a 9 bp (TACCGACAT) element in *RD29A* gene in Arabidopsis and is strongly related to dehydration, salt and cold stress (Yamaguchi-Shinozaki and Shinozaki 1994). The participation of DREs in abiotic stress is through both ABA-independent and ABA-dependent pathways (Dubouzet et al. 2003; Liu et al. 1998; Narusaka et al. 2003). Unlike ABRE, a single copy of DRE is sufficient to respond efficiently to stress stimuli in ABA-independent pathway (Yamaguchi-Shinozaki and Shinozaki 1994). There are other elements that contain A/GCCG core; however, they are not considered DREs. The C-repeat (CRT) and low-temperature responsive element (CTRL) are DRE-like elements which contain the DRE core (A/GCCG) and participate in gene expression regulation in cold responsive genes (Baker et al. 1994; Jiang et al. 1996). DRE represents the binding site for the ERF/AP2 type transcription factors. DRE binding proteins (DREBs) and C-repeat binding factors are among the well characterized ERF/AP2 type transcription factors that binds to DRE and play critical role in abiotic stress tolerance in plants (Stockinger et al. 1997; Liu et al. 1998). There are several groups and subgroups of DREBs/CBFs, which respond to various abiotic stress stimuli. DREB1s/CBFs group are divided further to six subgroups (Sakuma et al. 2002), among them DREB1A/CBF3, DREB1B/CBF1 and DREB1C/CBF2 which are involved in quick response to cold stress through upregulating numerous cold inducible genes that eventually encode cold tolerant proteins like late embryogenesis abundant (LEA) proteins as well as enzymes for sugar metabolism and fatty acid desaturation (Maruyama et al. 2009). Despite the critical roles of DREB1/CBFs in increasing cold stress tolerance, overexpression of such proteins have growth inhibitory effects in Arabidopsis (Liu et al. 1998) and rice (Ito et al. 2006).

DREB2s are another group of DREBs and divided to eight and five subgroups in Arabidopsis and rice, respectively. Functionally, this group of DREBs differs from the previous one regarding the type of abiotic stress they respond to. It has been reported that DREB2A and

DREB2B are drought, high salinity and heat responsive proteins (Liu et al. 1998; Nakashima et al. 2000), which indicates they are induced by dehydration in the ABA-independent pathway, while DREB2C, DREB2D and DREB2F are high salinity inducible, whereas DREB2E is ABA inducible (Sakuma et al. 2002). Unlike DREB1A, DREB2A overexpression displays no growth retardation and dwarfism in plants.

1.1.3 Evening element and evening like elements circadian element (EE)

As from its name, evening, this element is closely related to temporal gene expression in plants. Evening element is a plant specific element that was first identified by Harmer et.al. (Harmer et al. 2000a). Computationally, after surveying 450 circadian genes, they found a conserved motif of 9 bp (AAAATATCT) 46 times in 31, co-regulated, cyclic genes and expression of such genes reached their peak in the evening. To determine and confirm the importance of this element, a mutagenic analysis of *CCR2* (a circacadian gene) was performed and they demonstrated that circadian rhythmicity was sufficiently induced at 130- bp upstream of the transcriptional starting site which contained one copy of an evening element and another related 7 bp element. Furthermore, a partial mutation of the 7 bp element showed no rhythmicity alterations; however full sequence mutation caused notable rhythmicity reduction. This could suggest the importance of the whole sequence for proper circadian rhythmicity induction. Evening element is considered a binding site of MYB transcription factors like evening-expressing TIMING OF CAB EXPRESSION 1 (TOC) (Strayer et al. 2000), CIRCADIAN CLOCK ASSOCIATED 1 (CCA1) (Wang and Tobin 1998) and LATE ELONGATED HYPOCOTYL (LHY) (Schaffer et al. 1998). The aforementioned transcription factors comprise a circadian clock central loop in Arabidopsis (Jiao et al. 2007). The CCA1 and LHY suppress TOC1 through binding to EE in its promoter, while TOC1 acts as a positive regulator for further transcription of CCA1 and LHY. EE is also reported to participate in abiotic stress responses. A study by Mikkelsen and Thomashow (Mikkelsen and Thomashow 2009) revealed that cold induction of COL1 and COR27, two cold induced genes in Arabidopsis, requires EE and EEL elements coupled with ABRE like elements (ABREL). They observed the presence of a single EE element, two EEL (AATATCT) elements, and one CCA1-binding site (CBS) element, as well as six ABREL motifs in COL1 gene. Similar to COL1, COR27 promoter contained two EE element, two EEL element and three ABREL motifs. The presence of the EE and CBS in the promoter of either aforementioned genes indicated involvement of such genes in the circadian pathway. Furthermore, it reveals that cold responsive action of EE is not apart from its original circadian rhythmic function. The presence of EE and CBS was also confirmed in the promoter region of three DREB1/CBF genes (Dong et al. 2011), and since EE and CBS are binding sites of LHY and CCA1, respectively, they found that LHY and CCA1 act as positive regulators of DREB1/CBF, which is further confirmation of the crosstalk between circadian rhythm and cold stress. In addition, this harmony between circadian rhythm and cold stress may have protective effects against accumulation of excess DREB1s/CBFs which negatively affects the plant growth.

1.2 CRE structure and degeneracy

As CREs are part of the DNA sequence, they are composed of specific nucleotides (herein referred to as letters) and represent binding sites for certain trans-acting factors, for example transcription factors. There are different types of transcription factors comprising many transcription families. The members of a certain transcription factor family share characteristic features, among them a DNA binding domain. Despite sharing DNA binding domains, not all members of the same transcription family bind to the same CRE; however they have been shown to bind to CREs with similar core sequences. Thus, the members of a transcription factor family possibly have numerous specific binding sites. Unfortunately, to date, the exact sequences of many CREs are not fully identified with the possibility of other letters occupying the same position in a single CRE. This is known as element degeneracy. For instance, an element like A/GCCGACNT (Kyonoshin et al. 2004) is an 8 letter element and usually reported as a binding site for DREB1A; however, it is not a single element. The N letter indicates that of any pyrimidine or purine nucleotide can occupy this position. Therefore, the A/GCCGACNT is actually 8 elements which might represent binding sites for 8 transcription factors. This feature is more obvious in cis element databases like Jaspar (Sandelin et al. 2004) and PlantPan (Chow et al. 2016). In Jaspar database CREs discovered primarily from ChIP-seq experiments are displayed as a logo (sequence logo) having limited sequence length in x-axis and multiple letters of different sizes on y-axis. The bigger the size of a letter refers to the likelihood or the probability of existence of this nucleotide in this specific position is greater than the smaller letters (Figure 1.1). The bioinformatics methodology by which these CREs are discovered results in a position-weighted matrix; in which columns corresponds to positions of nucleotides in aligned binding sites and each row to a nucleotide (Figure 1.2). As shown in the figure, multiple letters are stacked in certain columns while others have single letters in a column. Therefore, this 9 letter degenerate element actually represents many different elements depending on the number of degenerate letters in each column. For this reason, at this moment it is doubtful that every single represented sequence is a functional CRE and can be considered false discoveries. Such false results are based on the hypothesis that all the possible numbers of CREs that originate from this logo have the exact same transcription factor binding to it with the same constant of association, and thus the function in the promoter. In another words the implicit hypothesis in each presented sequence logo is that all homonyms (different CREs coming from the same degenerate CRE) are also synonyms (binding to the same trans-acting factor with the same strength) at the same time. In this work, we test the hypothesis that all homonyms are also synonyms by first unpacking all degenerate CREs and examining all 1bp mismatches to a fixed word. Our results refute this hypothesis strongly. Not all homonyms from a degenerate or 1bp mismatches to a CRE are synonyms. However, we found that many synonyms are also homonyms. We also found that some synonyms had sequences different from each other.

1.3 Trans acting factors

1.3.1 Transcription factors

Transcription factors (TFs) are regulatory proteins, which are transcribed by genes to regulate genes expression of other or downstream genes. Transcription factors participate in gene expression regulation through binding to specific binding sites (CREs) of the promoter region of a gene. The binding typically occurs when the cell receives a signal either internally or from the environment. Then the TF is recruited through interactions with signal molecules or other proteins, possibly modified or re-localized to the nucleus to bind to specific CREs in order to increase or decrease the expression of a gene. TFs are usually unable to bind to a promoter by themselves. They often associate with one or more other proteins, and the whole process of transcriptional activation or suppression is carried out through interactions with a complex known as transcription initiation complex (Singh 1998) which includes RNA polymerase, the TATA box binding protein (TBP), and general non-specific TFs (such as TFIID) and other activators. Gene specific TFs that are part of differential gene regulation have a binding site that varies in both structure and length according to the type of the TF. There are TFs that require specific core sequences in the binding site to interact with, like the ACGT core for bZIP TF family (Uno et al. 2000b), while others require a binding site of certain length to interact functionally, like TACCGACAT for DREB1A TF (Liu et al. 1998). Understanding the nature of the TF binding sites is a critical part of the puzzle in understanding transcriptional gene regulation. In the post-sequencing and post-genomic era, TF binding sites recieved great attention from researchers which resulted in the identification and characterization of numerous binding sites.

1.3.2 microRNA (miRNA)

These are short sequences of non-translated RNAs and one of many small RNA molecules that are not protein-coding genes. miRNAs are often part of longer RNA molecules before they are processed and excised from loop regions through splicing to become independent molecules (Jones-Rhoades et al. 2006). There are up to 20 miRNA families conserved among Arabidopsis, rice and poplar. Plant miRNAs, unlike animal counterparts, are usually generated from areas in the DNA that are not related to protein coding genes. In plants, they regulate genes by binding directly to a messenger RNA, creating a double stranded RNA which is recognized and cut ("diced") by Dicer-like proteins. The mature miRNA is exported to cytoplasm by HASTY protein (an exportin protein) and then integrated to ARGONAUT (AGO) protein. Later the complex is guided by RNA-induced silencing complex (RISC) to their target sites and affect gene expression. miRNA could participate in gene expression regulation either through posttranscriptional pathway or through epigenetic pathway by causing epigenetic changes through DNA and histone methylation (Khraiwesh et al. 2012). miRNAs have been shown to have regulatory roles in plants and play critical roles in abiotic stress responses. miRNA targets in messenger RNA are often in the 3' UTR region or in introns of the coding region, and rarely found in the promoter. It has been reported that *miR168, miR171, and miR396* are drought, cold and high salinity responsive miRNA in Arabidopsis, respectively (Liu et al. 2008).

1.3.3 Methylation and chromatin remodeling

Methylation refers here to adding methyl groups in to either DNA sequences or specific amino acids of histone proteins at specific locations in the chromosome. In a DNA molecule, cytosine (C) is usually methylated and less frequently adenine (A), thus participating in gene expression regulation. This methylation usually occurs in the upstream promoter, in addition to the 5'UTR region and sometimes the entire gene and flanking regions. In plants, DNA methylation plays an important role in various aspects of plant development and abiotic stress responses. It has been reported that cold induction of *ZmMI1* gene in maize is significantly related to reduced DNA methylation (Steward et al. 2002). Also, histone modifications due to stresses can increase DNA methylation. Another type of modification, histone acetylation (adding an acetyl group to specific amino acids), has been reported in knockout mutants of *HDA6* of Arabidopsis and *HDA101* of maize as a result of histone methylation pattern and de-repression of silenced genes (Earley et al. 2006; Rossi et al. 2007).

1.4 Methods of CREs discovery

1.4.1 Computational methods

1.4.1.1 Enrichment in co-regulated genes

The DNA microarray is a powerful tool and used to generate an enormous volume of whole genome-gene expression (transcriptomic) data. Though currently replaced by next generation RNA sequencing, it is still a powerful and versatile genomic tool to track and monitor gene expression throughout DNA sequence and provides high throughput and large scale genomic analyses (Trevino et al. 2007). Early application of DNA microarray technology was in studies related to human disease and tumor detection and soon found its way in plant science community (Wullschleger and Difazio 2003). One of the most important uses of microarray technology in plant science is detecting CREs in co-regulated genes. Microarrays simultaneously monitor expression of thousands of genes and provide a transcription profile for each gene individually, which determines the expression pattern of that gene in a time course, specific tissue and expermental condition (including environmental conditions). however, microarray alone is unable to provide information regarding regulatory elements; thus, in order to detect motifs and CREs, an additional tool is required; in this case, it would be algorithms (Moreau et al. 2002). As co-expressed genes are co-regulated at the transcription level, co-regulation occurs at the level of promoters and trans-acting factors. Co-regulated genes may therefore share regulatory elements in their promoter region. Bioinformatic analysis employs algorithms that can detect statistically overrepresented motifs and CREs in co-regulated genes. The most common types of algorithm used to detect CREs and motifs in co-regulated genes are Gibbs sampling and hidden Markov model.

1.4.1.2.1. Gibbs sampling

The methodology of Gibbs sampling is based on Markov chain Monte Carlo (MCMC) for optimization by sampling (Moreau et al. 2002). This model was first used by Lawrence *et al.* (Lawrence et al. 1993) to detect subtle local residue patterns in multiple sequences. A modified version of this model was used to detect motifs and CREs in co-regulated genes in Arabidopsis by Thijs et al. (Thijs et al. 2002). In this model, the algorithm is applied on input of aligned sequence (co-regulated genes) which compares them with background noise/model (original sequences of such genes). First, the algorithm detects the presence of certain words (CREs) in the background model with given scores, and then compares the presence of such sequences in co-regulated genes. Second, the algorithm loops over all sequences, estimating the position of the motif through computing probabilistic distribution, which leads to discovery of a single motif. To determine multiple motifs, the algorithm would be applied several times while masking the position of the pre-discovered motif to avoid redundancy. Gibbs sampling has also been used successfully in detecting CREs in Arabidopsis in different sets of microarrays. Geisler et al. (Geisler et al. 2006) further developed a universal algorithm based on Gibbs sampling to identify CREs in Arabidopsis through identification of CREs in genes and scoring the statistical correlation of such CREs with the expression files of genes in given microarrays. Fortunately, this methodology resulted in identification of several novel CREs with biological significance.

1.4.1.2.2 Hidden Markov model

In this model, two types of sequences are used, modules and background. The modules contain motifs in addition to background nucleotides, while background contains only background nucleotides. The HMM generates nucleotides or binding sites from the background model or motifs found by position weight matrix, with a value representing a probabilistic value regarding its position. This method is useful for analyzing one sequence at a time. Discovering motifs and CREs in multiple sequences requires coupling HMM with multiple alignment. Aligned sequences are searched for CREs and in case shared elements presented the region with such elements collapsed and displayed as collapsed lines (Zhou and Wong 2007)

1.4.1.2.3 Word enumeration

An enumerative word tool is applied to determine and detect the presence of specific words in certain regions of a gene or overall genome. In case of an eight-letter word (CRE) 4 different of nucleotides (A, C, G, and T), there will be 65,536 words to search for in DNA sequences. The word enumeration tool is used by Lichtenberg (Lichtenberg et al. 2009) to determine overrepresented words in the whole genome of Arabidopsis coupled with Gibbs sampling. This method applies a binomial model in which the probability of a word being present is independent of the position of the word in the sequence (Robin et al. 2005). A Markov chain model for maximum-order homogeneity is used to calculate the probability of the words, while the maximum likelihood method is used to determine probability transitions. Since the number of the words is extensively huge (65,536), a chi square is used to give each word a p-value in order to filter out words which are not enriched or less represented.

Despite detecting mere words and disregarding their biological significance, this methodology could still play an important role in understanding promoter language through comparing enriched words with motifs and CREs discovered by other methodologies. Word enumeration were also used in monocots like rice (Cserhati 2015), which can provide a baseline to compare words in dicots and monocots and understanding the physiological differences between plants in these groups.

1.4.2 Experimental methods

1.4.2.1 Promoter deletion assay

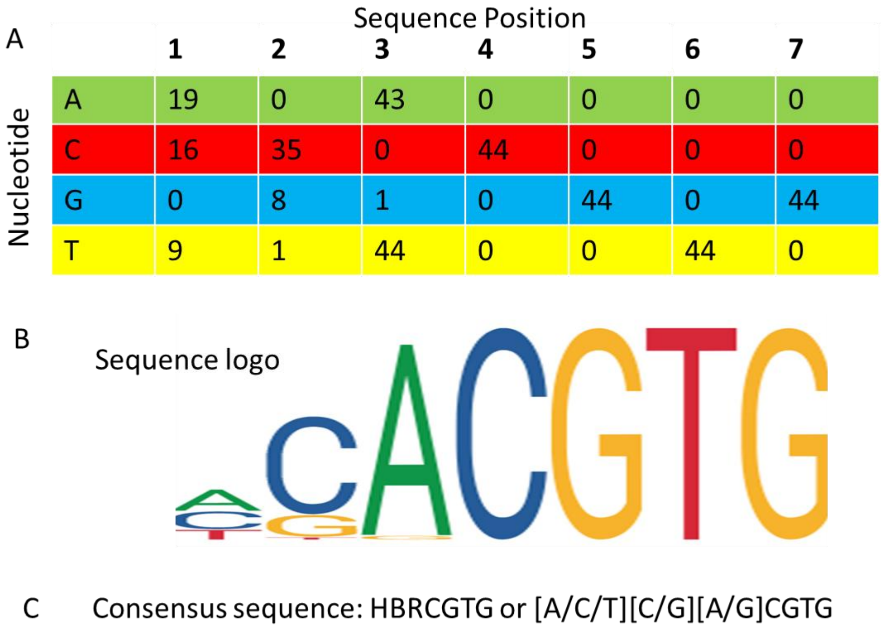
This is a molecular biology assay to detect CREs in promoter regions of genes. The technique is based on deleting ("bashing") the promoters in specific points using exonuclease enzymes, followed by cloning the remaining promoter region into plants to analyze the expression pattern of such promoter and determine how the deletion affected gene expression. In the pre-genomic era, this has been widely used to determine regulatory elements in plant and animal genes. The well known DRE element discussed previously was identified in *RD29A* through this technique. The promoter of *RD29A* was dissected (deleted) with exonuclease III in several places with specific time intervals to delete certain numbers of base pairs at a time depending on the enzyme. Then the promoter fragment is sequenced, followed by fusion with a reporter gene and cloning into the plant to monitor its expression pattern (Yamaguchi-Shinozaki and Shinozaki 1994). As seen from the methodology described above , this assay is very costly and time consuming especially in earlier times when sequencing was rather expensive compared to recent days. However, the advantage of this technique is the accuracy of the elements discovered, which provides powerful evidence for validation of predicted CREs.

1.4.2.2 Electrophoretic mobility shift assays (EMSA)

Known as gel shift or gel retardation assay, this technique used to detect protein-nucleic acid interactions, including both DNA and RNA. The principle of the methodology is based on the electrophoretic property of the DNA and protein -DNA, RNA complex. The DNA (or RNA) molecule has affinity toward positive charges, and migrate toward the positive electrode when a current is passed through a gel at different paces. When a protein binds to a DNA sequence, migration of the molecules will be slower than the control (protein and DNA separately. in other words, slower migration of molecules relative to control indicates protein-DNA interaction. Therefore, the mobility of the molecules which indicate molecule-molecule interaction (Garner and Revzin 1981). Eventually the complex is isolated from the gel and the DNA is sequenced to determine the specific sequences which bind to the protein. Accuracy is considered as an advantage for this technique; however, the methodology applies only to proteins acting as trans acting factors so, other regulatory trans acting factor that may bind to that specific DNA sequence will not be covered. In addition, in vitro studying of protein-DNA interaction is not necessarily a reflection of the in vivo interaction.

1.4.2.3 ChIP-Seq/ChIP-chip

These techniques measure and identify in vivo protein-DNA interaction through combining both chromatin immunoprecipitation and DNA microarray in a genome wide scale. The methodology is based originally on chromatin immunoprecipitation (ChIP) which is a powerful experimental assay to detect protein- DNA interaction in vivo. In ChIP-Chip technique, the immunoprecipitated DNA is combined with genomic DNA which serves as background or reference DNA. Later, the mixture is hybridized using microarray chips of the whole genome. The results retrieved will be an indication of DNA regions which are labeled or enriched with immunoprecipitation (Dey et al. 2012). The ability of probing vast number of genomic regions in one experiment is one of the most important advantages of this technology. Furthermore, it enables researchers to overcome the expense large scale qPCR assays byproviding platforms to study the localization of protein bindings and also parallel analysis of genes from different classes and families which makes statistical comparison easier. The ChIP-chip makes use of the DNA microarray to determine the DNA associated with the protein being tested, while ChIP-seq makes use of high throughput next generation sequencing.



A

Sequence Position

Nucleotide	1	2	3	4	5	6	7
A	19	0	43	0	0	0	0
C	16	35	0	44	0	0	0
G	0	8	1	0	44	0	44
T	9	1	44	0	0	44	0

C Consensus sequence: HBRCGTG or [A/C/T][C/G][A/G]CGTG

Figure 1.1 Matrix Model, Sequence Logo and Consensus Sequence of a CRE

A putative CRE of 7 letters from a ChIP-seq study (Choi et al. 2000) of ABF1 binding to Arabidopsis chromatin A: analysis as a position weight matrix model by position enrichment (Gao et al. 2017). Numbers represent the frequency of each nucleotide in positions from 1-7 of the CRE. The fractions are represented by height of letters in the sequence logo (B) and as degenerate nucleotides in a consensus sequence (C).

CHAPTER 2

CHARACTERIZATION OF THE LEXICON OF UPSTREAM PROXIMAL 8 bp CIS REGULATORY ELEMENTS IN ARABIDOPSIS AND RICE

2.0. Introduction

A cis regulatory element (CRE) is a short sequence (6-20 bp) mainly present in the promoter region of a gene. CREs are frequently found in the proximal and core promoter regions upstream of the transcription starting site (TSS). The TATA box is an example of a known CRE present in the core promoter region, which is considered a binding site of RNA polymerase II and participates in formation of the transcription initiation complex(TIC). CREs of biological interest, which comprise TF binding sites are mostly found in the 500 bp upstream region. in distal promoter region elements known as enhancers could be found and may have regulatory functions with the closest TSS. As there are thousands of genes in plants and animals and each gene contains at least one CRE, each organism may have a set of CREs, comprising a lexicon, analogous to words in a language. In our study, 1691 microarrays experiments (transcriptomes) were collected (Fitzek 2012) and Dr. Chai Ling Ho (unpublished) and combined with an algorithm script (Geisler et al. 2006) to perform high throughput CRE correlation analysis which resulted in building a CRE lexicon for both Arabidopsis and rice. The algorithm searched for shared CREs among corregulated genes, which finalized the set of CREs with certain scores. Then a Chi-square test (X^2) calculated whether CREs are induced, suppressed or neutral. The microarrays collected were of different categories in either species. In both Arabidopsis and rice abiotic stress comprised the majority of the transcriptomes that were significantly correlated to CREs in the lexicons. Tissue and pathogen categories were the second most abundant transcriptomes in both Arabidopsis and rice respectively. The powerful methodology used in the current study could be applied to other species to obtain further lexicons of other species, which could provide a strong baseline to understand the evolution of promoter language in plants.

2.1 Material and methods

2.1.1 Lexicon dataset

A sum of 1460 Arabidopsis and 231 rice transcriptomes were collected by Dr Fitzek (Fitzek, 2012) and Dr Chai Ling Ho (unpublished), respectively from GEO browser (http://www.ncbi.nlm.nih.gov/geo/) and Nottingham Arabidopsis Stock Centre's (NASC). A total of 65,536 CREs, which represents all possible combination of 8 bp word of 4 types of nucleotide, A,C,G and T were searched for in the corregulated genes in each transcriptome. Each gene had an M-value comprising its expression pattern, either induced or suppressed. An automated Chi2-test was used to score the strength of a CREs for either induction or suppression expression pattern. The strength of a CRE was determined by a p-value to specify the function of the CRE whether an inducer, suppressor, or neutral. A cut off (p-value > 10-^{08}) was used to eliminate sequences, like microsatellites, that might occur frequently using statistical enrichment tools. The p-value cut off was calculated by multiplying the number of transcriptome with 65,536. In Arabidopsis since 1460 transcriptomes were used the equation will be (1/(1460*65536)), in rice the equation will be (1/(231*65536)). Despite the fewer number of transcriptomes for rice, the cut off value was changed slightly. CREs passed the 10^{-8} P-value filter were searched for known elements like ABRE, DRE and EE in both species. The elements were searched for their expression pattern and functionality with regard to the already published counterparts. The Excel work sheet containing all the data set was used to identify the elements. Then lexicons of Arabidopsis and rice were filtered against each other to obtain shared elements between the two species. Excel graph charts were used to create the figures.

2.1.2 CRE overlap construction

In order to search for overlap between CREs in the Arabidopsis lexicon, the 641 CREs were subjected to multiple alignment using the online free version of Cluster X 2.1 (Thompson et al. 2003), a powerful tool for interpreting alignment data. The alignment parameters included gap opening (100), gap extension (6.66) and delay divergent sequence (30%). The default matrix, Gonnet series was used instead of PAM and BLOSUM series since it performed the alignment within a shorter period of time with similar results. There was no need to use rice data set for multiple alignment due the lack of rice database to compare and search for the overlapped CREs.

2.1.3 Functional annotation of CREs

Three known and two unknown elements retrieved from the Arabidopsis lexicon data set were run through Patmatch tool (Yan et al. 2005). The -500 bp upstream promoter regions in the TAIR10 version was used to search for the CREs in the Watson (given) strand. Patmatch parameter included maximum hit (75,000), mismatch (0), and minimum (1) and maximum (100) hit per sequence ,respectively. The genes retrieved through Patmatch were exported to David database (Dennis et al. 2003) an excellent functional annotation tool for clustering genes according to their functions. The list of genes containing the CRE were uploaded in the Enter Gene List section. The identifier selected was TAIR ID and list type identified as Gene List then the gene list were submitted. The results were retrieved from functional annotation clustering section which contains the number of clusters and their p-value and the number of genes in each cluster.

2.2 Results and discussion

CREs were discovered both genome wide, and systematically using all possible 8 bp nucleotide combinations (65,536 total). These were searched in 1,460 and 231different transcriptomes for Arabidopsis and rice, respectively collected from public databases and publications. Each potential CRE was scored if it was enriched or depleted in genes induced or suppressed. This was repeated for all 1460 and 231 transcriptomes. In Arabidopsis, 641 CREs were significantly correlated with 800 transcriptomes, which represent 0.97 % of the 65,536 total words, while 856 elements in rice were significantly correlated to 231 transcriptomes which represent 1.3% of total 65,536 words. The significance of selected CREs determined through a cutoff P-value of $>10^{-8}$ which corresponds to a Bonferroni correction for multiple hypothesis testing.

2.2.1 Retrieval of known elements

The availability of plant CRE databases cataloging discovered elements either experimentally of computationally enabled us to compare the lexicons built in the current study to those in the databases. In Arabidopsis, the most important and well characterized CREs like ABRE, DRE , and EE as well as ABRE-like (ABREL) and EE-like (EEL) elements were recovered with canonical and variant sequence spelling. The ABRE element, ACGTGTC, showed strong correlation to abiotic stress, ABA, and metabolic elicitor (KNO_3) as well as many other elicitors. The canonical ABRE is 7 bp long, but since the current study is based on 8 bp CREs, various versions of ABRE could exist. With the presence of four nucleotides (A, C, G, T), there is a possibility of recovering 8 versions of ABRE, but only 6 of them were present and GACGTGTC and TACGTGTC were excluded. The latter was a reverse complement of another CRE, that is why it has been removed, since reverse complements were disregarded due to similarity in expression pattern with the forward sequenced CREs. Collectively, the six versions of ABRE were significantly correlated to 492 transcriptomes which were mainly seed dormancy transcriptomes (63%) followed by abiotic stress (13%) (Figure 2.2A). Among the 492 transcriptomes, CACGTGTC was correlated to 173. Therefore for further analysis this version of ABRE is considered. ABREL elements, ACGTGGC, showed strong correlation to ABA, metabolic elicitors and considerably weaker correlation to abiotic stress elicitors. Four versions of ABREL was recovered in the current dataset. Collectively, the four versions of the element was correlated with 234 transcriptomes that were mainly seed dormancy (42%) followed by circadian transcriptomes, unlike ABRE (Figure 2.2B).

The 8 bp EE element was also recovered. The EE element had the highest and strongest correlation to abiotic stress (cold stress) transcriptomes with a p-value of 2.5^{-99}, which was the highest p-value in the whole dataset. Also, the EE element was correlated to 134 transcriptomes which were mainly of abiotic stress (51%) and circadian (31%) (Figure 2.2C). Similar to EE, EEL element showed strong correlation to cold stress with a p-value of 9.93^{-78}. EEL element were correlated to 99 transcriptomes which mainly were abiotic stress (58%) and circadian (32%) (Figure 2.2D). DRE element was shown in 2 versions, ACCGACA and TACCGACA since the canonical DRE is 9 bp element, and the one in the current dataset is 8bp. Together, they were correlated to 13 transcriptomes all of which were cold stress.

In rice, among the 28 CREs discovered, 18 were recovered by the current method, with correlation to the same published elicitors, except for a single CRE which responded to a different elicitor. Unrecovered CREs were due to either lack of significance (4 CREs) or infrequent presence in the genome which does not allow development a statistical model. ABRE, ABREL and EE, but not DRE were among the recovered elements. ABRE and ABREL, together, weakly correlated to only 5 transcriptomes, four of which were abiotic stress while the other was tissue transcriptome. EE was also weakly correlated to four transcriptomes of abiotic stress, developmental and ecotype comparison categories.

The huge difference in the strength and breadth of response between Arabidopsis and rice might refer to the nature of the two plants. Arabidopsis, is a dicot model plant with no economic significance, while rice is a monocot crop plant with important economic significance producing a high starch content product. Furthermore, the enormous amount of water required for rice irrigation may affect the significance of ABRE element and the moderate growing temperature may interpret the absence of DRE. In such crop plants, elements related to sugar loading and translocation protein accumulation could be highly significant. Unfortunately, our dataset did not include transcriptomes of such category.

2.2.2 CREs overlap and using an 8 bp word

In CREs discovered using ChIP-seq or other methods, the length of binding site varies in different trans-acting factors between 6 and 20 bp. For example, the well-known ABRE element is usually presented as 7 bp while EE element is 9 bp . TAAGAGCCGCC, which is an ethylene-responsive element found in tobacco class I chitinase gene, is 11 bp long (Shinshi et al. 1995). In our methodology an 8 bp (letter) word search pattern was used to systematically discover the lexicon CREs in both Arabidopsis and rice. However, the 8 bp word search in this study also discovered a sequence similar to ABRE and several shorter versions of EE that overlapped the entire known element. In other words using an 8 bp based word search can potentially find larger and smaller CREs. In case of shorter CREs there will be flanking sequences of either A, C, G, and T. However, not necessarily all the four nucleotides to be present as a flanking sequence. In other words, there may not be A-CRE or T-CRE present in our database due to 10-^{8} cutoff p-value. If not all 4 possible nucleotides are found in the flanks of shorter known CREs, this might indicate the importance of flanking sequence on CRE function. In case of CREs longer than 8 bp, a complete version of the real CREs could be dispersed between multiple overlapping 8 bp CREs. This results in detectable overlaps of systematically detected 8 bp CREs in our datasets. Looking systematically for every possibility of overlapin 641 and 856 CREs in Arabidopsis and rice respectively is very difficult; therefore, a few examples of known long elements were individually searched in each species to further explore this concept in addition to attempt to systematically search the entire CRE lexicon.

In Arabidopsis a multiple alignment for all 641 CREs was carried out using ClustalX2. As passing through alignment results, several overlapping regions were noticed which lead to construction of CREs longer than 8 bp. A 13 bp CRE, AATCCCGCCAAAA, was generated from the overlap of five of 8 bp CREs (Figure 2.3A). However, this hypothetical 13 bp element was not found in any Arabidopsis promoter using Patmatch tool in 500 bp gene upstream regions, indicating that such reconstructions of element could be misleading. Surprisingly, when the last A is deleted from the left side of the element making a 12 bp CRE, Patmatch found two genes that contain this element. As this type of deletion was continued and number of genes including the newly formed CREs were increasing until it reached to 27 genes for a 9 bp element, and the original 8 bp CRE was found to be present in 57 genes. Another overlapping hypothetical 12 bp CRE, TCACTCGAGTAA, was generated from overlap of four of 8 bp CREs (Figure 2.3B). None of the Arabidopsis genes were found to contain this sequence. However, when an A was deleted from the element, the 11bp element was found in a single gene. Further deletion resulted in further increase in number of genes. Nine and 29 genes were found to contain the 10 and 9 bp elements, respectively. The original CRE, TCACTCGA, were found in 333 genes.

The absence of 13 and 12 bp CREs might be due to the nature of the elements and the presence of an element this long might not frequently occur in promoter regions. This also points out that CREs significantly longer than 8 bp simply do not occur in enough genes to be able to make statistically relevant correlations to gene induction or suppression, which is a serious limitation of the method in finding larger elements. That is, using shorter element search windows like 8 bp (Grotewold and Springer 2009) is better at detecting CREs through high throughput tools and it potentially reduces false negative results (Lichtenberg et al. 2009).

2.2.3 Shared elements between Arabidopsis and rice

In our data set 641 and 856 CREs in Arabidopsis and rice were identified, respectively. We searched for shared elements between the species and found 78 shared CREs. This comprised 12.16% and 9.11% in the whole Arabidopsis and rice lexicon, respectively. Among the shared elements are versions of ABRE, ABREL, EVE, EVEL, AAACCCTA, and few TATA box elements. In general the shared elements are GC rich element which were closer to the structure of rice elements, unlike Arabidopsis which were AT rich.

In Arabidopsis, the 78 shared elements were present in 3361 transcriptomes and those elements where significantly related to mostly tissue transcriptomes (38%) followed by seed dormancy transcriptomes (16%) (Figure 2.4A). In rice, the shared elements were present in 632 transcriptomes only, which is almost one fifth of the Arabidopsis counterpart. Similar to Arabidopsis, the shared elements in the rice were mostly related to tissue transcriptomes (33%) followed by developmental transcriptomes (27%) (Figure 2.4B). Since there were no seed dormancy transcriptomes in rice dataset, seed dormancy elements were characterized. In

Arabidopsis, the highest p-value recorded by shared elements in rice was by EE (2.5^{-99}) followed by ABRE (2.03^{-77}), while in rice the elements AAACCCTA and AACCCTAG recorded highest p-values, 1.24^{-80}, and 6.68^{-80} respectively. As mentioned above, EE element showed highest p-value in Arabidopsis, while in rice the same element showed very low p-value (3.12^{-12}). Unlike EE element, the AAACCCTA CRE showed high p-value in either species. Further analysis of shared elements will be carried out in the next chapter.

2.2.4 Functional assignments of CREs

In earlier sections of this chapter, some known elements were retrieved in the current dataset. In this section the functionality of some known and unknown elements will be explained. Three of the well-known and characterized CREs (ABRE, EVE, and DRE) and two unknown elements were selected. Except for DRE, the categories of transcriptomes in which the CREs involved were greatly diverse. For this reason, the transcriptome category to which the CREs highly significantly related is selected. Therefore, Table 1 and 2 do not include all categories of the transcriptomes the CREs were significantly related to.

The current version of ABRE (CACGTGTC) is further analyzed and shown to be significantly correlated with 173 transcriptomes. The highest p-value recorded in metabolism transcriptome regarding 25mM of KNO3, while the lowest p-value recorded in seed dormancy transcriptome. As shown in Table 1, ABRE is highly correlated to metabolism microarrays, however, it is not induced in such transcriptomes. Unsurprisingly, the only induction occurred in hormone (ABA) and seed dormancy transcriptome, since the element responds to alterations in ABA concentration which plays a critical role in seed dormancy (Kucera et al. 2005).

Evening element (EE) is also further analyzed. As mentioned in a previous chapter, EE is identified as a key element in the circadian rhythm of plants and participates in cold stress tolerance, not apart from its original function. The highest p-value recorded for EE was in cold stress transcriptome (Table 2.1). The involvement of EE in oxidative stress with such high p-value 10^{-71}, is further supports our methodology in detecting and analyzing CREs, since circadian clock is acts as a regulator of reactive oxygen species (ROS) homeostasis and oxidative stress response (Lai et al. 2012). EE element was induced in all the five categories as shown in the Table. The third element is AAACCCTA, the unknown element found in the highest number of transcriptomes among all 641 CREs. It is obvious that this element is strongly related to tissue rather than any other aspects whether suppressed or induced. The second unknown element is GAAAAGTC. This element is strongly related to pathogen transcriptomes, however the p-values were not as high as in previous elements. Furthermore, the element showed induction in all the categories that were highly correlated to them, suggesting a positive regulatory role of this element. The last element in the current Table is DRE, the third well known CREs. The position of the element in the Table is due to the low number of transcriptomes it significantly related to them. DRE is well documented to be a cold responsive element (Yamaguchi-Shinozaki and Shinozaki 1994), and our study showed that it is only regulated by cold stress, and in all 5 cases in the table, DRE showed only inductive response.

In rice, the conditions were different. Since the well-known elements were related to very low number of transcriptomes, the elements with higher number of transcriptomes were selected (Table 2.2). The only CRE that is shared between Arabidopsis and rice in this section is AAACCCTA. The latter showed highest correlation with transcriptomes regarding rice seedling. Similar to the Arabidopsis version, this element is shown to be closely related to tissue transcriptomes; however, a pathogen transcriptome is also included in rice the version. Unlike the Arabidopsis version, this element showed neutrality toward transcriptome regarding developmental aspects, suggesting the tissue-restricted element hypothesis for this specific CRE. In general the rice elements selected in this section were mostly significantly related to tissue and developmental transcriptomes, except for CTCGCCGC and GAGGAGGA, which were more likely correlated to abiotic stress transcriptomes.

2.2.5 Functional clustering of CREs

The CREs analyzed for functional assignments in the last section were further investigated for functional clustering. Using David database (Dennis et al. 2003), the genes that genes containing those elements were annotated for their functions and clustered. ABRE element was found in 639 genes, and they were divided into 63 clusters. However, very few of them were significant cluster. The top cluster showed a score of 3.44, and the genes in this cluster were mostly chloroplast genes. There were 24 genes with highest p-value and were clustered in the chloroplast thylakoid membranes, followed by 32 genes clustered in the chloroplast stroma. The rest of the genes in this first cluster were distributed among transit peptides, chloroplast and plastid categories. This indicates that genes with ABRE elements are mainly involved in photosynthesis especially the light reaction part. In the second cluster, with enrichment score of 2.63, functional annotation showed enrichment of transcription factor activity and sequence specific DNA binding, revealing the main function of ABRE. The rest of the clusters were enriched for transcription factors, iron and sucrose metabolism, however, they lacked the significance power. EE element was found in 1,506 gens and functional annotation of the top cluster with an enriched score of 3.7 showed they are highly enriched for and highly significantly related to AP2/ERF DNA binding region and ethylene signaling pathway. As participates in cold induction response, EE might occur in promoter regions of cold induced genes like *DREBs/CBFs* (Dong et al. 2011) and *COL1* and *COR7* (Mikkelsen and Thomashow 2009). The unknown

AAACCCTA element was observed in 1,779 genes. Post clustering results showed that genes containing this element are highly significantly related to ribonucleoprotein and ribosome encoding genes with 95 and 82 genes in each, respectively. They were both included in the first cluster with enriched score of 19.77. This might be interpreted as involvement of this element in translational processes rather than transcription. The GAAAAGTC element was found in 496 genes. The annotation and functional clustering showed that the top cluster with an enriched score of 2.94 were enriched for leucine-rich repeat protein coding genes and the genes under this category were mostly involved in coding disease resistant genes like TIR-NBS-LRR -class family of genes. Collectively, with the results in Table 1, we suggest that this element is a pathogen resistant element. DRE was found in 153 genes. The first cluster of 1.9 enrichment score showed stress response including water deprivation, cold and osmotic stress abscisic acid response genes. That explains the restrictive response breadth of this element. The rice element were not subjected to further analysis due to lack of rice functional annotation tools.

2.2.6 Relationship of CREs to genes

The abundance of genes with predicted CREs was scored for all CREs that strongly significantly correlated (*P*-value<10^{-8}) to gene expression in a microarray in both Arabidopsis (641 CREs), as well as rice (856 CREs). In Arabidopsis the most abundant CREs observed in the induced genes were, in sequence, similar to that of TATA-box elements. This is unsurprising since 29 % of Arabidopsis genes showed to contain TATA elements and usually spin around -32 bp upstream to transcription start site.(Molina and Grotewold 2005). The presence of the TATA-box is essential to increase transcription level in certain promoters. It has been reported that presence of three copies of the TATA-box in promoter regions significantly increases transcriptional level of β-Phaseolin promoter, although with differences in transcriptional activity depending on nature of TATA-box and spacing between three copies (Grace et al. 2004). The presence of TC motifs were not observed in Arabidopsis 500 upstream region in top 20 abundant genes containing CREs for both induced and suppressed categories. However, it was reported that 18% of Arabidopsis genes contain TC motifs and its conserved in rice (Bernard et al. 2010). Additionally, the structure of these most abundant CREs were very simple in which, except for the ACTTTTTT element, were composed of only two nucleotides, TC, TA, CA, or GA. The presence of TATA-box elements in the most abundant category could be considered as an evidence of the validity of our methodology. Simplicity of CREs was observed in the most abundant rice CREs as well. However, TATA-box elements were not as frequent as in Arabidopsis. Instead GC rich and CT elements were noticed frequently. Except for ATATATAA word, rice and Arabidopsis shared no other elements in this top 20 most, although they shared other CREs but in less abundant genes abundant category. The TC element was present frequently in the top 20 abundant genes containing CREs (Figure 2.4 c). However, TC elements recorded in current data differs from those reported (Bernard et al. 2010). A trinucleotide TC element of 2 T and one C, TTCTTC, CTTCTT, TCTTCT, were documented (Bernard et al. 2010).

Beside the most abundant gene containing CREs, there were also least abundant genes containing CREs. There were 34 CREs observed in only one gene for induction in Arabidopsis, while 44 CREs were not observed in any genes at all. In rice, the least number of genes where CREs were present was 11 genes, and there was no instances of CREs being absent in any genes. Similar CREs pattern and structure were observed in suppressed genes in both Arabidopsis and rice. Opposite to abundant CREs in suppressed gene, in Arabidopsis 29 CREs were observed in only one gene for suppression and 44 other CREs were not shown in any suppressed genes. In rice, the least number of genes for CREs to be observed for suppression was 3, and no instances of CREs being absent in genes were recorded.

TABLE 2.1 FUNCTIONAL ASSIGNMENT OF CREs IN ARABIDOPSIS.

CRE	No. of MA	Category	*p*-value	Enrichment
CACGTGTC	173	metabolism	76.692	biregulation
		metabolism	57.724	suppression
		hormone	55.574	induction
		metabolism	51.385	suppression
		seed dormancy	48.199	induction
AAATATCT	134	cold	98.602	Induction
		oxidative	71.794	Induction
		circadian	65.626	Induction
		cold	59.91	Induction
		drought	51.278	Induction
AAACCCTA	197	root tip	64.453	induction
		root hair	64.453	suppression
		tricellular pollen	56.094	suppression
		uninucleate microspore	56.094	induction
		stem	54.783	induction
GAAAAGTC	60	pathogen	24.074	Induction
		ozone	20.053	induction
		pathogen	19.841	induction
		hormone	19.531	induction
		circadian	19.295	induction
ACCGACAT	9	cold	22.329	Induction
		cold	20.934	Induction
		cold	18.081	Induction
		cold	17.942	Induction
		cold	14.173	Induction

The -log10 of p-values were calculated. The type of function of the CREs in each category is illustrated as either biregulated (both up and down regulated), induced, suppressed or neutral.

MA= Microarray

TABLE 2.2 FUNCTIONAL ASSIGNMENT OF CRES IN RICE.

CRE	No. of MA	Category	*p*-value	Enrichment
AAACCCTA	53	seedling	79.906	suppression
		developmental	43.391	neutral
		pathogen	41.486	suppression
		root	40.935	induction
		panicle	40.086	suppression
AGCTAGCT	45	root	72.583	suppression
		callus	59.391	biregulation
		shoot	46.123	suppression
		developmental	42.006	induction
		embryo	37.974	biregulation
GCGGCGGA	58	developmental	45.322	neutral
		root	43.289	induction
		embryo	40.714	neutral
		shoot	39.679	induction
		developmental	38.563	neutral
CTCGCCGC	48	heavy metal	34.034	neutral
		heavy metal	32.406	neutral
		mutant	32.062	neutral
		shoot	31.787	induction
		salt	31.049	neutral
GAGGAGGA	44	developmental	48.422	neutral
		mutant	46.081	neutral
		salt	45.275	induction
		salt	41.462	neutral
		heavy metal	41.211	neutral

The -log10 of p-values were calculated. The type of function of the CREs in each category is illustrated as either biregulated (both up and down regulated), induced, suppressed or neutral.

MA= Microarrays

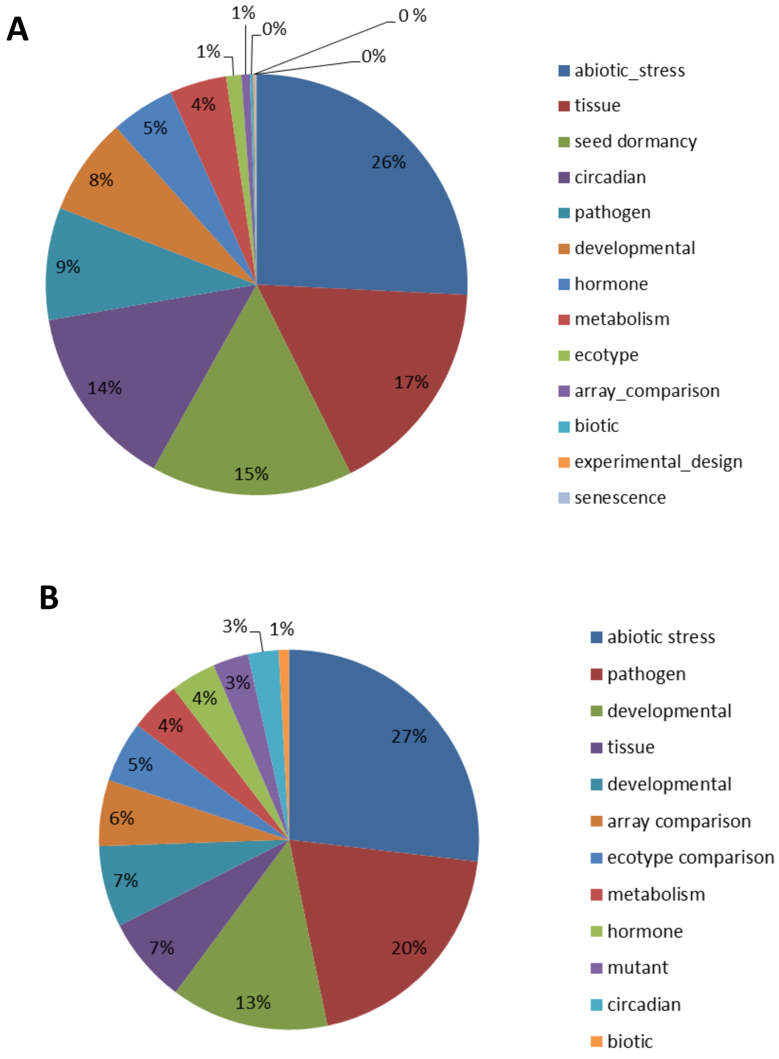


Figure 2.1 Transcriptomic experiment designs used to identify CREs in Arabidopsis and rice. Experiments were placed into design categories in (A) Arabidopsis and (B) rice. Most experiments in both species were abiotic stress microarrays. Other common experiments were exposure to pathogens, development and tissue comparisons, and application of hormones. A total of 804 microarray based transcriptomic experiments for Arabidopsis and a further 231 for rice were collected and normalized for comparison in subsequent systematic CRE discovery and analysis.

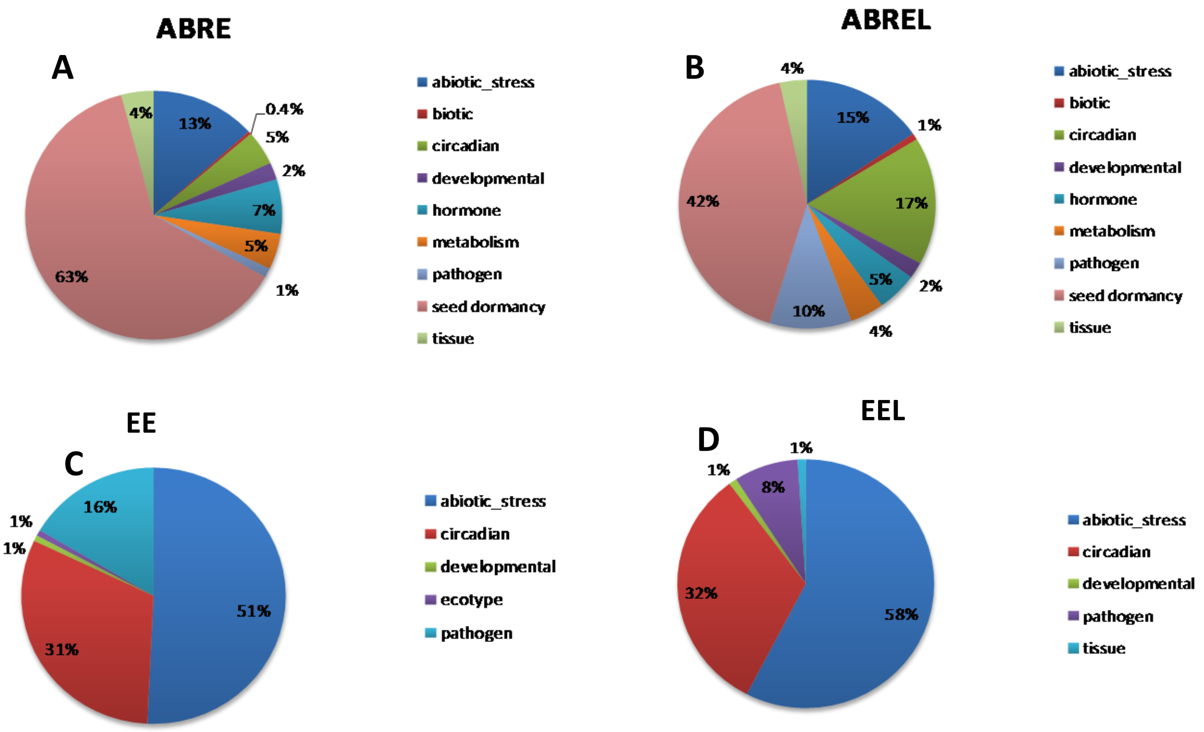
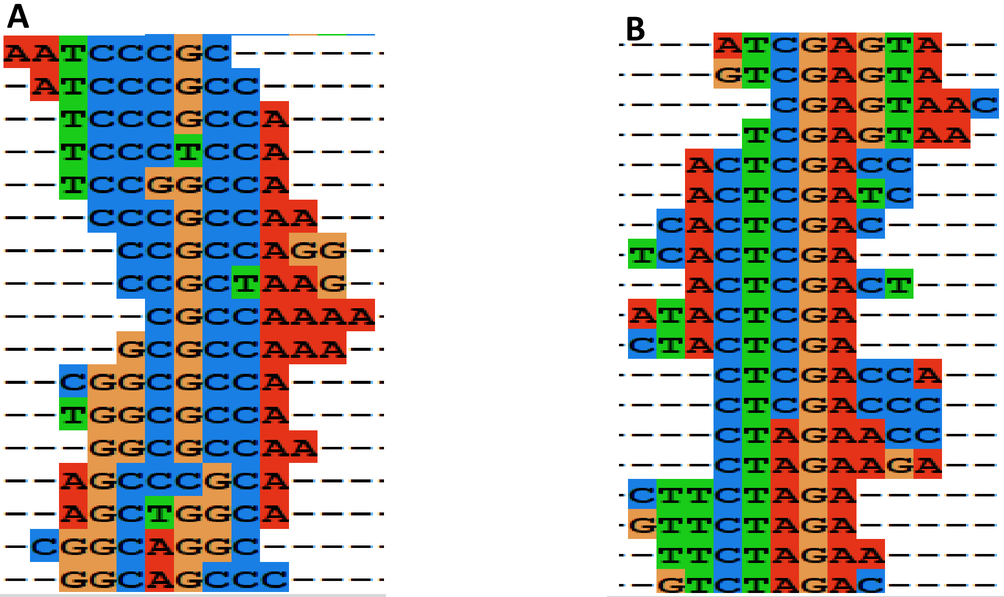


Figure 2.2 Transcriptome categories of commonly known CREs in Arabidopsis. The ratio of transcriptome categories significantly correlated with (A) Abscisic acid responsive element (ABRE) , (B) Abscisic acid responsive like element (ABREL), (C) Evening element (EE) and (D) Evening like element (EEL) elements in Arabidopsis. The L letter in ABREL and EEL refers to like, since they differ from the original ABRE and EE respectively.



CRE	Base pair	No. of genes
AATCCCGCCAAAA	13	0
AATCCCGCCAAA	12	2
AATCCCGCCAA	11	4
AATCCCGCCA	10	11
AATCCCGCC	9	27
AATCCCGC	8	57

CRE	Base pair	No. of genes
TCACTCGAGTAA	12	0
TCACTCGAGTA	11	1
TCACTCGAGT	10	9
TCACTCGAG	9	29
TCACTCGA	8	333

Figure 2.3 CRE overlap in Arabidopsis. A multiple alignment of 641 CREs in Arabidopsis using Clustal X2 alignment tool. (A and B) two examples of CRE overlap which resulted in construction of one 13 bp, AATCCCGCCAAAA and one 12 bp TCACTCGAGTAA CRE. The rest of the CREs represent shorter version of the 13 and 12 bp after one nucleotide deletion at time to test the frequency of each CER in -500bp upstream region in Arabidopsis genes. number of genes obtained through Patmatch tool. The original 13 and 12 bp were not present in the whole Arabidopsis genome. Deletion of nucleotides at a time resulted in increasing the number of gene that contain the post-deletion CREs.

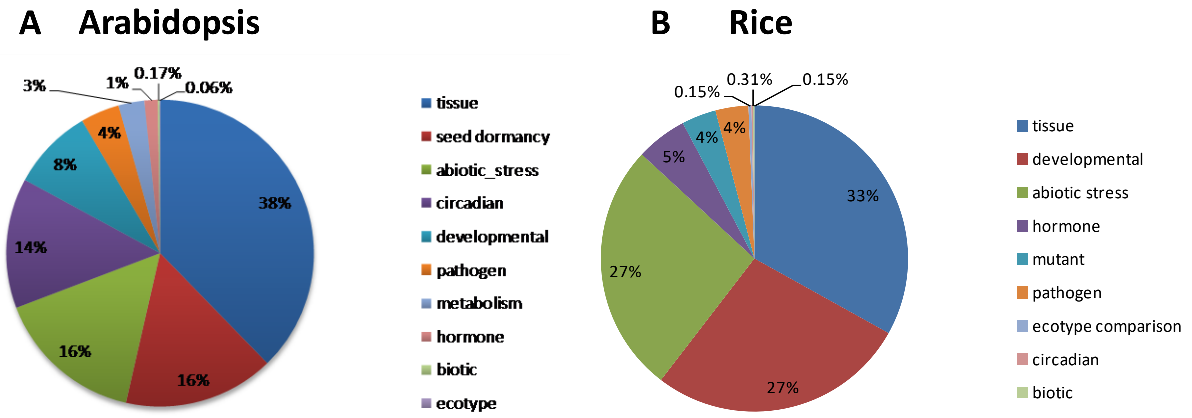


Figure 2.4 transcriptome categories of shared element. Arabidopsis and rice shared 78 CREs. The number and frequency of transcriptomes significantly correlated to CREs is varied. A) Categories of 3361 transcriptomes significantly correlated to shared elements in Arabidopsis. B) Categories of 632 transcriptomes significantly correlated to shared elements in rice.

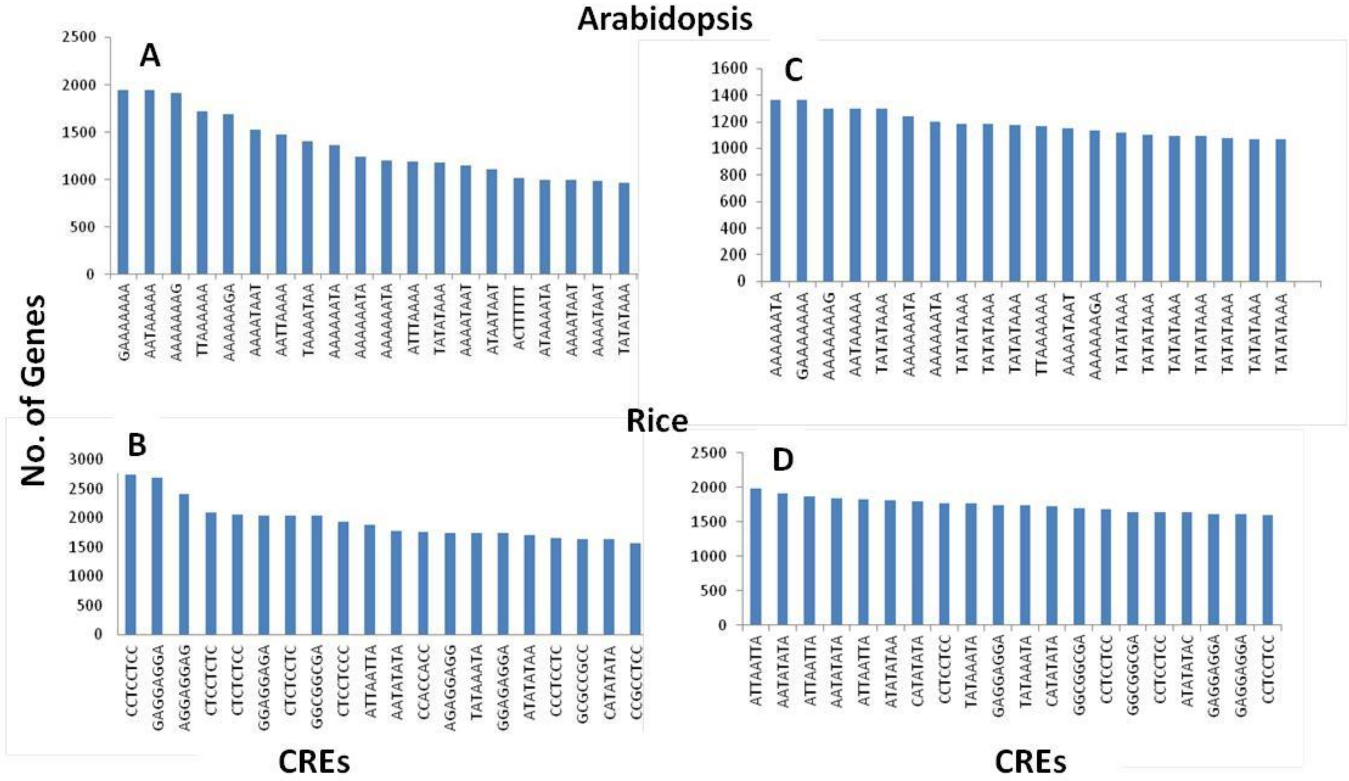


Figure 2.5 Most abundant predicted CREs in Arabidopsis and rice promoters. CREs were scored by exact match in forward strand of the 500bp upstream (Arabidopsis) or 1000bp upstream (rice) regions of protein coding genes. A and C) CREs in induced and suppressed genes respectively in Arabidopsis. B and D) CREs in induced and suppressed genes in rice.

CHAPTER 3

ANALYSIS OF PROMOTER LANGUAGE CHARACTERISTICS IN ARABIDOPSIS THALIANA AND ORYZA SATIVA

3.0 Introduction

Plant gene promoters function to control gene expression by interacting with trans-acting factors that can include diverse mechanisms such as transcription factors, DNA methylases, miRNAs and others. Each trans-acting factor is associated with one or more cis-regulatory elements (CREs) in the promoter and can act to suppress or enhance gene expression. CREs and promoters can be treated as a language consisting of 4 letters (A, C, G, T), and all possible 8bp words were analyzed for function by associating each word with a pattern of increased or decreased gene expression in gene containing them in different tissues and developmental stages or experimental treatments. Decryption of the promoter language was done by enumerating all 8bp words and looking for patterns linked to gene regulation. This gave a *CRE lexicon*, a list of all functional words (641 in Arabidopsis, and 856 in rice) that represents about 1% of all possible words, and separated them from nonsense words (~64,000 remaining 8bp words). In addition to filtering for these functional words, a *CRE dictionary* was constructed by mapping each word to a series of functions. In the previous chapter the lexicon was analyzed for language patterns (patterns in the spelling and occurrence of words). Here the dictionary is now analyzed for meaning, looking a strength and breadth of response and how these tie to complexity. The concept of CRE synonyms and homonyms is also introduced. A CRE synonym shares the same correlated elicitors; for example two CREs that are both correlated to gene expression in cold stress induction. A CRE homonym or homophone is one that has similar sequence and may be a member of a degenerate sequence discovered in bioinformatics analysis of co-expression or

ChIP-seq data. A CRE can have single or multiple meanings, in that they can correlate with differential gene expression of a single elicitor, or multiple different elicitors. Signaling in plants is non-linear, and involves branched and cross-talking pathways. Some promoters act as integrators of different signals, and pass on that integrated signal to the next promoter via their transcription factor. The CRE that has multiple meanings likely indicate where it is located in the branching and cross talking signaling pathways and that it is bound to by an integrating trans-acting factor. As we begin to explore promoters as a language, rice and Arabidopsis can be thought of as distantly related languages, like Spanish and Italian both belong to the same language family, and some of the words are similar enough to be understood in both languages.

Arabidopsis is the non-economic dicot plant which is used to understand basic developmental and growth aspects of plants within a brief period of time, and thus enjoys a large number of publically available experiments for analysis (Meinke et al. 1998). On the other hand, rice although an economically important plant, is less well studied by transcriptomics, in part due to the increased difficulty in growing and conducting experiments. However, rice was selected for this study as the most intensively studied plant among monocots, which might vary from dicots as it is evolutionarily more distant and allows us to explore conserved promoter features (Izawa and Shimamoto 1996). The higher number of pathogen experiments of rice is a reflection of the economic importance of studying disease resistance in rice. Potential bias in experiments has been reduced to an extent by matching comparable studies in the two organisms. Another feature easily observed in the publicly available experiment categories in both species is the high ratio of abiotic stress experiments (compared to animal studies which are more focused on development and biotic diseases). Since plants are sessile and unable to mover, the various environmental conditions may adversely affect its growth and development, and with increased threats of climate change studying adverse effects of abiotic stress extensively is a priority among plant scientists.

3.1 Material and methods

3.1.1 CREs strength and response breadth

The CRE lexicon built for both Arabidopsis and rice were sorted according to their p-value and breadth response to identify list of strongest and broadest response CREs in both species. For each category top 20 CREs were selected for further analysis.

3.1.2 CRE complexity

CRE lexicons were dissected into 8 independent characters in an Excel spreadsheet. The function command =countif(A:H, nucleotide) was used to determine the number of A, C, G and T in each CRE. Later, a combination of two functional commands =if ((max (I:L,number)),1,0)) to determine the complexity level of a CRE. In latter function, the number indicates the frequency of each nucleotide in a given CRE. For instance if the max number of certain nucleotide was 8, the complexity level will be one, which means that CRE consists of single type of nucleotide. Furthermore, if the max number was 2, this means there are two copies of each type of nucleotide, therefore, complexity level will be seven.

3.1.3 Correlation between strength and breadth

Highest p-value (strongest response) copy of each CREs was considered and compared with its response breadth. A linear regression graph was created with a linear equation for lexicons of both Arabidopsis and rice.

3.1.4 Homonym and synonym analysis

Ten CREs were selected to analyze and study the homonym concept in Arabidopsis. The criteria of the CREs has to have accepted response breadth to allow expression pattern comparison with their homonyms. The Patmatch tool was used to retrieve all possible homonyms for a given CRE. The -500 bp upstream promoter regions in the TAIR10 version was used to search for the homonyms in the Watson (given) strand. Patmatch parameter included maximum hit (75,000), mismatch (1), and minimum (1) and maximum (100). Due to limited number of transcriptomes in rice homonym analysis for rice was not performed. Synonymic analysis was conducted through determining the expression pattern (in this case transcriptome categories) for all 641 CREs in Arabidopsis lexicon. Two CREs sharing more than 50% transcriptome categories considered as synonyms.

3.2 Results and discussion

In this project transcriptomes of both Arabidopsis and rice subjected to hormone and abiotic stress experiments and/or tissue dissections were collected to carry out a systemic search for CREs that are significantly correlated to a treatment, stress, or specific tissue. A sum of 1,460 and 231 experiments were collected for both Arabidopsis and rice respectively. These experiments were placed into different categories as illustrated in Figure 1a,b, which included both stress and developmental pathways. The number of Arabidopsis experiments exceeds the rice counterpart by seven fold, as this model organism is simply more often studied. However 10 Arabidopsis experiments were evaluated to be an approximate match to rice experiments in terms of tissues examined and/or treatments applied. In this chapter, the relationship between the CRE and the experimental condition were connected by identifying CREs that were enriched genes differentially expressed in that condition. This produced a CRE lexicon (list of all bio-active CREs), as well as a CRE dictionary, in which each CRE were associated with their elicitor(s) and thus their potential meaning/function. All CREs are clustered for similar elicitors/function, which we called synonyms, and for similar spelling (homonyms) which thus creates a CRE thesaurus.

3.2.1 Characterization of the CRE dictionary by breadth of response.

We searched for CREs with broadest response in terms of the number of different experiments the CRE was associated with, and with most abundance among genes in their 500bp upstream promoter sequence, and finally CREs with the most statistically significant (strongest) correlation between occurrence and differential gene expression with experiments. The breadth of response was calculated as the presence of a CRE in a set of genes that correlated to strong significant pattern of differential expression (p-value$< 10^{-8}$) in largest number of experiments. For example, in *Arabidopsis thaliana* the presence of the element AAACCCTA in gene 500bp upstream sequences was observed to be correlated to differential expression 197 different experiment experiments while ACGTGGCA was similarly observed in 113 experiments (Figure 3.1A). As seen in Figure 2, the elements that correlated to multiple experiments are of a complex sequence pattern (as defined in Chapter 2). Some of the elements are highly variable and contain all the four kinds of nucleotides (A, T, C, G). However others are less variable and restricted to three or two kinds of nucleotides. No element that correlated with differential expression was observed to have only one kind of nucleotide. Another observation of the elements shown here is that they all contain at least one adenine or thymidine (A/T) in their structure but no other nucleotide (C or G) was universally present. CRE that resemble known, experimentally validated CREs like CACGTGTC, which is similar to the known abscisic acid responsive element (ABRE), and AAATATCT, which similar to the evening element (EE) were found to be correlated to gene expression in a large number of experiments, beyond what these elements have been previously reported to be responsive to. It was not surprising for these elements to be in broadest response category of our data set, since they are well known to be involved in multiple stress and developmental processes. This analysis recaptures all known environmental associations (e.g. ABRE with ABA, EE with circadian rhythm), but also shows many novel, previously unknown associations.

In rice (*Oryza sativa*), the breadth of the response was not as numerically high as in Arabidopsis due to lesser number of experiments available for study. GCGGCGGA was observed to be significantly correlated to differential gene expression in 58 experiments and CGCCTCCG in 46 experiments (Figure 3.1B). Rice elements observed in this broad response category are GC rich elements with simpler (less complex as defined in Chapter 2) patterns of sequence compared to broad responding Arabidopsis elements. No ABRE or EE –like elements were observed among the 20 most broadly responding elements. The rice ABRE and EE like elements were found, but they were much more narrowly responsive than their Arabidopsis counterparts. Rice ABRE and EE-like elements were found only in 2 and 4 experiments. GC-box (GGCGG) and GCC core containing elements were observed frequently in this set of broadly responsive elements. In the top 20 broadest response category one element, AAACCCTA was found in both rice (53 experiments) and Arabidopsis (197 experiments). Overall, only 79 elements were shared between rice and Arabidopsis, representing about 10% of the total. Thus finding 1 out of 20 elements shared across these angiosperm families gives the evolutionary distance between the promoter languages across angiosperms, as rice (monocot) and Arabidopsis (dicot) represent an early branch in the angiosperm tree of life.

The expression patterns of newly discovered CREs, AAACCCTA and ACGTGGCA in Arabidopsis and GCGGCGGA and CGCCTCCG in rice, were further explored. In Arabidopsis the element AAACCCTA is correlated with differential expression in most studies of tissue comparisons (68 transcriptomes), which might indicate that this CREs is a tissue specific element (Figure 3.2A). On the other hand, ACGTGGCA hit mostly seed dormancy experiments followed by circadian experiments (Figure 3.2B). In rice elements, GCGGCGGA and CGCCTCCG were both correlated with differential expression in mostly experiment studies of abiotic stress comparisons. Unlike Arabidopsis, tissue experiments were not frequently hit by the rice elements mentioned above, however, following abiotic stress, developmental experiments were second mostly hit experiments by GCGGCGGA and CGCCTCCG (Figure 3.2C and D). This grouping of experiments by category and correlation to shared CREs might thus show underlying regulatory mechanisms they have in common, and those processes which are more separated.

Only a few CREs exhibited a large breadth of response in both rice and Arabidopsis, the far greater proportion of detected CREs was restricted to differential expression in only one or a few thematically related experiments. In Arabidopsis from 641 CREs detected overall, 291 responded to a single experiment. Of these single responders, 92 of the experiments were in perturbation of metabolism while another 89 experiments were of abiotic stress (Figure 3.3A). In rice, a similar proportion (278 out of 856) of CREs, responded to one experiment. Of these responders, 117 were correlated to tissue experiments followed by 85 CREs hitting abiotic stress experiments (Figure 3.3B). There were no metabolism experiments in rice microarrays, for this reason no CREs correlated with such experiments.

3.2.2 Characterization of the CRE dictionary by CREs strength of response

Strongest response category of CREs was evaluated by calculating the highest significant correlation of the presence of a CRE to significant induction or suppression of the gene by an experimental condition. This is also considered as an important category to be analyzed in our data set, as they might represent the most obvious and confident regulatory connections discovered by this method. In Arabidopsis a CRE with the 8bp pattern AAATATCT that overlapped the 9 bp EE (Mikkelsen and Thomashow 2009) element had the highest significant response after 12 hour exposure of cold in shoot tissue (Table 3.1). This element, although previously characterized, was not used in any way to train the discovery method, thus was a true rediscovery rather than being an expected artifact. This 8 bp version of EE was also correlated significantly to other experimental conditions like oxidative, salt, drought and wound stress, as well as circadian clock. Despite the fluctuation in strength, EE had higher significant correlation to abiotic stress and circadian experiments. The second element having highest significance relationship to experiments was another CRE with a different overlap to the 9 bp EE element AAAATATC. This might indicate that our use of all combinations of 8bp words could potentially discover elements that were larger or smaller than 8bp. This 8 bp EE element look alike (AAAATATC) had a similar pattern of responses to experiments except for the one tissue experiment hit by this element. In addition, this element recorded its highest significant correlation to the same experiment hit by EE. This might indicate that even shifting one letter from original CRE its response could still be high and possibly with similar experimental condition. Another element (CACGTGTC) that matched the characterized 7 bp ABRE (Hattori et al. 2002) element (ACGTGTC) was also shown to be among CREs with the strongest response, which is yet another rediscovery. The ABRE-like element had highest significant correlation to metabolism experiment when seeds imbibed for 48 hours in 25 mM KNO_3. This is a coincidence since the role of ABA in seed development and maturation is well-documented (Rock and Quatrano 1995; Frey et al. 2004; Raz et al. 2001; Karssen et al. 1983). In addition, the ACACGTGT element similar to ABRE and ABRE-like elements was highly correlated to experiment the same as ABRE-like element. The element that had the broadest response, AAACCCTA, was also present in the strongest response category with EE and ABRE but has not been previously described and is presented here as a novel, broadly and strongly responsive CRE detected by this method. This element showed highest correlation to experiment of root tissue when the root reached its full diameter and 15 mm away from root tip. Furthermore, derivatives of this element like AAAACCCT and AACCCTAA were also among top 20 strongest CREs, and involved in tissue experiments in pollen and root tissues respectively. As seen from the Table, the structure of the CREs in the strongest responsive category is of complex pattern having at least three types of nucleotides in each CRE. In rice, no EE element or its derivative or ABRE element were present. However the novel AAACCCTA and one of its 1-bp mismatch derivatives were shown as the highest significant CREs (Table 3.2). Either element showed strongest response in tissue experiments in rice seedlings. As observed in the Table 3.2, CREs structures are simpler than those observed in Arabidopsis with multiple CREs having only two types of nucleotides. In Arabidopsis and rice, root tissue comprised the majority of tissues where the CREs showed their strongest response.

3.2.3 Characterization of the CRE dictionary by CRE complexity

Another aspect that may affect CREs properties and expression pattern is element complexity. Many well characterized element show high sequence complexity, with a non-repeating pattern of all 4 DNA nucleotides. In this study CREs were categorized into seven levels of complexity. Level 1 comprises the simplest CREs having 8 replicates of the same type of nucleotide going up to level 7 comprising the most complicated CREs having 2 replicates of the same type of nucleotide, regardless of the position of these replicates. In Arabidopsis no CRE was observed under level 1 category. The least complex CREs were of level 2 category with 6 CREs out of 641 (Figure 3.4A). Conversely, 25 CREs were among the most complicated elements, i.e. level 2 categories. In between, the rest 630 CREs were distributed on the rest of the complexity levels, with the majority of CREs, 316 elements, under level 6 category having 3 replicated of the same type of nucleotide (Figure 3.4A). As mentioned earlier CREs expression pattern could be complexity-dependent, the 25 most complicated elements were investigated for their functions. Collectively, the 25 CREs hit 423 experiments and the majority of the experiments related to seed dormancy followed by abiotic stress (Figure 3.5A).

Unlike Arabidopsis, CRE complexity in rice was quite different. CREs structure in rice tends to be simpler than their Arabidopsis counterparts. There was only one element in level 1 category, a full C element (Figure 3. 4B). Furthermore, the level 2 category has 4 fold CREs compared to Arabidopsis, 26 CREs. In the far side of complexity, out of 856 CREs 22 CREs were observed among most complicated CREs, level 7. In addition, level 5 comprised the majority of rice CREs, 261 elements, and compared to Arabidopsis the number of CREs in level 4 is dramatically greater. From our observations, we can conclude that rice CREs detected by this method are simpler than those of Arabidopsis.

As Arabidopsis and rice, share 78 CREs complexity levels of the shared elements were investigated (Figure 3.4C). It was observed that most of the shared elements were from level 5 category (33 CREs) followed by level 6 (30 CREs). Surprisingly, this number dropped to only one CRE (CTGATCGA) in level 7 category, which reveals that the most complex CREs might be species specific. CREs complexity revealed that maximum complexity of CREs is tissue specific phenomenon. Arabidopsis and rice possessed 25 and 22 highly complex CREs, respectively, and except for CTGATCGA; they shared no other CRE in level 7 category. Another notable aspect regarding level 7 complex CREs searching for known common cores in like ACGT. This common core is the common binding site for bZIP family transcription factors.

In Arabidopsis among 25, most complicated CREs 7 had ACGT cores. As the aforementioned 25 CREs collectively hit 434 experiments, 369 of them were hit by ACGT-containing CREs (Figure 3.6A), while the rest 18 CREs hit 65 experiments. For this reason, no other cores were investigated since ACGT-CREs hit most of the experiments. In contrast, out of 22 most complicated CREs in rice, only 3 of them had ACGT cores. Collectively these three ACGT containing CREs hit 29 experiments explaining the huge difference between Arabidopsis and rice CREs. Furthermore, since ACGT containing CREs hit very few of the predicted CREs from this method, the presence of other common cores was also investigated. A multiple alignment of the 22 CREs resulted in revealing other common cores like GATC and CTAG (Figure 3.6B). The latter, represented by 4 CREs, hit 128 experiments, while GATC-containing CREs, represented by 6 CREs, hit only 88 experiments.

3.2.4 Response breadth versus Response strength

The relation between response breadth and strongest response of CREs was investigated. In Arabidopsis as the majority of CREs hit single experiment a big mass of blue squares (CREs) accumulated in the left corner of the graph (Figure 3.7A). Moreover, CREs having strict response breadth also have low p-values. As shown in the figures, as CREs responses become broader p-value increases as well. However, there were few exceptions like EVE element (blue arrow) which had highest p-value among the 641 CREs but not the broadest response as AAACCCTA (black arrow). Similar pattern was observed in rice (Figure 3.7B) with more CREs accumulating around the regression line. However, the number of outliers observed in rice were not comparable with that of Arabidopsis.

3.2.5 Core sequences in CREs

A core sequence refers to a sequence that is shared by a significant number of CREs. For the 4 letters of a CRE (A, T, C, and G) there are 256 possible ways to make a 4 letter core in 8 letter CREs. So, in all the 65,536 CREs (words) there are 256 core sequences. In Arabidopsis, 253 core sequences were observed and cores like GGTA, TGCT, and TTAC were missing in the entire 641 CREs. Despite excluding only 3 cores from the 256 core sequences, 93 of the observed cores showed enrichment (Figure 7a), while the rest were depleted. The latter two terms refer to the ratio between observed and expected value. Enrichment means that observation values were greater than expected values, while depletion means observed values were less than expected values. The highly enriched core sequences in Arabidopsis CREs are listed in Table 3.3. As shown in the Table, CGCG, despite not being the most observed core, is the most enriched core sequence. Furthermore, ACGT a b-ZIP transcription factor family binding site (Izawa et al. 1993; Menkens et al. 1995; Uno et al. 2000a), came as the second most enriched core sequence, which is not a surprise. Additionally, a version of DRE (RCCG) core sequence, a binding site for transcription factors with ERF/AP2 binding domain (Sakuma et al. 2002; Liu et al. 1998; Yamaguchi-Shinozaki and Shinozaki 1993), ACCG was also reported to be highly enriched (Table 3.3).

In rice, out of 256 core sequences 17 cores were completely missing in the entire 856 CREs. In the rest, 103 showed enrichment , while the rest were depleted. Among core sequences that were not present is AACA, which is considered as a binding site for MYB proteins and plays an important role in glutelin storage in rice (Suzuki et al. 1998), and endosperm-specific gene expression (Wu et al. 2000). The simplicity of rice CREs were more illustrated by being CCCC as the most observed and enriched core sequence. In addition, GCs were more frequently observed than A and /or AT. In contrast to Arabidopsis, no b-ZIP and ERF/AP2 binding sites (ACGT and ACCG) were present among the most enriched cores; however, both cores were present in rice CREs. Although, ACGT was slightly enriched and observed in 25 cases, ACCG, was depleted and observed in 15 cases only. Furthermore, there were only 3 cores (CGCG, GCCC, CCCA) shared by Arabidopsis and rice among the most enriched core sequences as indicated by asterisk. However, the two species shared 45 enriched sequences, which means 48.3% and 43.6% of total enriched cores in Arabidopsis and rice respectively, with variations in enrichment values. In our study and as mentioned earlier, 641 and 856 CREs were discovered to be highly significantly correlated to differential gene expression in Arabidopsis and rice respectively, and the two species shared 78 CREs (Figure 7B). This means 12.3% and 9.2% of total active CREs in both Arabidopsis and rice respectively. This observation may indicate that the two species share cores frequently rather than 8 word CREs. In other words, Arabidopsis and rice tend to share a part of CREs and thus, accordingly, the two species might share families of CREs disregarding the rest of the CREs (the remaining 4 letters).

As mentioned before, not all core sequences were enriched. There were 160 and 136 cores depleted in Arabidopsis and rice, respectively. The most depleted core sequences are listed in Table 2. As shown in the Table there was no shared cores between two species. Furthermore, the nature of the core in this category was different from most enriched cores. The cores in this case were T rich and there were at least one T in each core in either species. This could be explained, as Ts may be a factor for reducing enrichment and depletion. However, no studies or other observations were recorded earlier to cement our discovery.

3.2.6 Synonyms, homonyms, and homophones in promoter language

In any language in the world, words are made to refer to specific items. A word could be a noun, and derivatives of the same word may become a verb, adjective and adverb among the others. A single word may have multiple meanings. For example, date which refers to a fruit usually grows in warm areas, and at the same time refers to a romantic meeting and a specific day. Furthermore, rose, type, net are also words that are spelled and pronounced the same way; however, having different meanings. In linguistics, these words are called "homonyms." In addition, there are words that are pronounced the same but spelled differently. For example, ate and eight, rain and reign, course and coarse are words with different meanings and spellings; however, same pronunciation. These are known as "homophones." Furthermore, there are words pronounced differently depending on the phrase. For instance, minute could mean a small or tiny thing as well as a 60-second of time. The meaning changes radically when its pronunciation changed. These are "homographs." On the other hand, words with different spellings and pronunciation but with the same or similar meanings, like rich and wealthy, are called "synonyms."

Considering promoter as a language and CREs as words, these terms could be applied to understand the evolution of CREs and how mismatches of even a single letter may change the expression pattern of CREs. In this case, homonyms may refer to CREs that are similar in spelling but with different meanings, while synonyms will be CREs having different letters but similar meanings. Since CREs cannot be pronounced like ordinary words, applying the homograph concept may differ from its original linguistic definition. Despite the different spelling of CREs, they might sound the same to a specific transcription factor. For example, differently spelt CREs may respond to the same abiotic stress after specific time of exposure and tissue.

To achieve this goal, 10 CREs with their one-letter mismatch counterpart were chosen to follow their functional expression pattern. Several strategies were followed regarding CREs selection. CREs with broad response patterns like AAACCCTA, known elements, and cores like EVE or ABRE elements or ACGT and DRE cores were selected. However, elements with very narrow response patterns were disregarded. One letter mismatch of the aforementioned CREs were retrieved through Patmatch tool (Yan et al. 2005) as available on www.Arabidopsis.org. Since there are four types of nucleotides, a one-letter mismatch of an 8 letter CRE resulted in 24 derivatives; however, not all of them are present due to two reasons. First, they did not pass a 10^{-10} raw *P*-value filter from the chi-squared test (which gives 0.01 Bonferroni corrected *P*-value after correcting for multiple hypotheses). Second, the derivative might be reverse complements of other active CREs in the dataset. Therefore, the derivatives of the selected CREs in Figure 8 are the ones that passed the 10^{-8} *p*-value filter.

The 8-letter ABRE element (CACGTGTC) responded to 14 different experiments, six of which were abiotic stress related experiments (Figure 3.9). From 24 ABRE one-letter mismatch derivatives, four CREs present among the 641 active CREs in Arabidopsis. When the first C letter substituted with A (AACGTGTC), the expression pattern narrowed to 5 experiments excluding all the abiotic stress experiments except for osmotic stress. When the C letter was restored and the second T replaced by G, the new element restored all the experiments, except for dark stress and biotic, hit by ABRE. Few experiments were missed when the third C letter was replaced by A (CACGTGTA). However, when the second G was replaced by T, the expression pattern was restricted to 2 experiments only. An overall observation of ABRE and its derivatives is that despite the occurrence of mismatches, the derived elements maintained their involvement in developmental and biological processes (especially tissue experiment); however, when the last G letter was replaced with T, the new element (CACGTTTA) disappeared in all developmental and biological experiments. This could be a good example of the effect of replacing one nucleotide with another. It is obvious how the C and G letter (**C**ACGT**G**TC) is critical to maintain the cis element functional (available) for different transcriptional factors to bind. More specifically, despite the mismatches occurred, the ABRE derivative element conserved the core ACGT, which is essential for sequence for ABRE functionality. As one letter mismatch of ABRE resulted in derivatives some of which with considerably high breadth response, except for one, in EE (AAATATCT), which responded to 134 transcriptomes, one letter mismatch restricted the expression pattern of the three derivatives to 5, 4, and 2 transcriptomes. When the C was substituted with A, the expression pattern restricted to two tissue transcriptomes only. This indicates the importance of cytosine in this exact position. Furthermore, this incident explains an evolutionary aspect that substitution of C with A, decreases the affinity of transcription factors toward the new element, and the CRE may lose its main function as well. Another important feature regarding homonyms in our data set was the observation that there were no derivative CREs after one letter mismatch. The element CTTATCCA was significantly correlated to 37 transcriptomes of 9 different categories, but when the 24 derivatives were obtained by Patmatch tool, none of those derivatives were significantly correlated to transcriptomes. In other words, CTTATCCA possesses no homonyms. As letter mismatch may alter and diverge the expression pattern of CREs, functional similarity (similar expression pattern) could be observed among CREs with similar or rather different sequences. Such CREs can be defined as synonyms. In Arabidopsis, due to the availability of huge number of transcriptomes, studying synonyms could be easier and more doable, unlike rice which has comparably fewer transcriptomes to derive such information. To reveal synonyms in the Arabidopsis lexicon, 20 CREs with considerable response breadth were used to track down similarity and mismatch in expression patterns. Unlike the homonym concept for detecting synonyms, the mismatch in expression is considered regardless of the number of mismatches in CREs structure. For this reason, a cutoff of >50% similarity in expression pattern was made as a standard to identify synonyms. In other words, if a CRE was correlated to 14 expression pattern categories, its synonyms must be correlated to at least 8 of them. The number of synonyms were greatly diverse among the 20 selected CREs. AAGGCCCA and AACACGTG possessed the highest number of synonyms, 30 for each, while GAAAAGTC and AAATATCT (EE) possessed least number of synonyms, 6 and 5 respectively (figure 3.10). In addition to the small quantity, the quality of the GAAAAGTC synonyms was poor as well. Its closest synonyms shown mismatches in 5 expression pattern. In contrast, AATGGGCC, AAGGCCCA, AAGCCCAA, AACACGTG possessed closest synonyms. Their top five closest synonyms showed zero mismatches in one of the synonyms and one mismatch in the rest of the four synonyms. Not all of their synonyms showed sequence similarity. Two of closest synonyms of AAGGCCCA, were rather different in structurally, GCCCAATA, AATGGGCT, with one expression pattern mismatch.

Similarity in expression patterns among CREs indicates the presence of CRE families similar to transcription factor families. TFs contain DNA binding domain, which mediates TF-DNA interaction, in other words, TF-CRE interaction. Evolutionary mutagenesis could result in the formation of a new DNA binding domain through duplication (Chen and Rajewsky 2007). Duplication of TFs is, often, followed by mutation in DNA binding domains resulting in formation of a new domain. The newly formed DNA binding domain may bind to new CRE as well as its original binding site; however sometimes newly formed binding domains are unable to restore its original binding site. Simultaneously, genes may undergo duplication and have different fates. The newly formed gene through duplication may acquire a novel function and the other copy may still restore its original (ancestral) function. This is known as neofunctionalization. although the main function may be partitioned between the two duplicated genes. This is called subfunctionalization. Another alternative fate would be restoring the original function by both copies, which results in expression pattern robustness of the two genes (Gu et al. 2003; Force et al. 1999). Since CREs are located in the promoter region of the genes, gene duplication also means CRE duplication, and the duplicated CREs could have similar fates as duplicated genes. Thus, the likelihood of TF-CRE overlap could be strongly enriched, which means CREs may group in families similar to TFs.

TABLE 3.1 CRES WITH HIGH SIGNIFICANT CORRELATION TO EXPERIMENTS IN ARABIDOPSIS.

CREs	**p-value**	**Exp. No.**	**Category**	**Condition**	**Tissue**
AAATATCT	98.602	ex682	abiotic_stress	Cold stress-12 hr	shoot
AAAATATC	77.003	ex682	abiotic_stress	Cold stress-12 hr	shoot
CACGTGTC	76.692	ex1423	metabolism	25 mM KNO3 48 hr	seeds
AAGGCCCA	66.142	ex100	tissue	root tip full diameter	root
AAACCCTA	64.453	ex100	tissue	root tip full diameter	root
GCCCATTA	64.407	ex1210	tissue	mature pollen	pollen
AATGGGCC	64.110	ex1209	tissue	tricellular pollen	pollen
AGGCCCAT	61.801	ex100	tissue	root tip full diameter	root
ACACGTGT	55.836	ex1423	metabolism	25 mM KNO3 48 hr	seed
ATTGGGCC	53.547	ex100	tissue	root tip full diameter	root
AAAGCCCA	51.910	ex1210	tissue	mature pollen	pollen
AGGCCCAA	51.854	ex1207	tissue	mature pollen	pollen
GCCCAATA	50.065	ex100	tissue	root tip full diameter	root
ACGTGTCA	49.131	ex1423	metabolism	25 mM KNO3 48 hr	seed
AAGCCCAA	47.698	ex100	tissue	root tip full diameter	root
AACCCTAA	46.855	ex100	tissue	root tip full diameter	root
AAAGTCAA	43.999	ex695	metabolism	10 µM CHX	whole plant
TAGGCCCA	43.838	ex100	tissue	root tip full diameter	root
AAAACCCT	43.672	ex1206	tissue	tricellular pollen	pollen
AAACGCGT	43.004	ex695	metabolism	10 µM CHX	whole plant

The p-values were converted from (-) to (+) through taking -log of base 10

TABLE 3.2: CRES WITH HIGH SIGNIFICANT CORRELATION TO EXPERIMENTS IN RICE.

CREs	**p-value**	**Exp. no.**	**Category**	**condition**	**Tissue**
AAACCCTA	79.906	181	tissue		seedling
AACCCTAG	79.175	181	tissue		seedling
AGCTAGCT	72.583	47	tissue		root
GCTAGCTA	66.080	47	tissue		root
ACCCTAGC	54.364	181	tissue		seedling
CCTCCTCC	52.928	15	developmental	root vs. wax layer	root
CTAGCTAG	51.924	47	tissue		root
GAGGAGGA	48.422	15	developmental	root vs. wax layer	root
AAAACCCT	46.327	181	tissue		seedling
GCGGCGGA	45.322	30	developmental	11-20 DAP	seed
AGGAGGAG	44.182	25	abiotic stress	salt	root
GCGCCGCC	41.642	181	tissue		seedling
CGGCGGAG	41.255	47	tissue		root
GGCGGCGA	39.732	47	tissue		root
CCTCCGCC	38.528	47	tissue		root
CCGCCGCG	38.209	4	abiotic stress	Heavy metal-arsenate	root
AAGCTAGC	37.779	226	tissue	2,4-D, light	callus
CCGCCTCC	37.061	47	tissue		root
CCCGCCGC	34.385	47	tissue		root
CTCGCCGC	34.034	4	abiotic stress	Heavy metal -arsenate	root

The p-values were converted from (-) to (+) through taking -log of base 10

TABLE 3.3: MOST ENRICHED CORE SEQUENCES IN ARABIDOPSIS AND RICE.

Arabidopsis				Rice			
Core sequence	**Observed**	**Expected**	**Enrichment**	**Core sequence**	**Observed**	**Expected**	**Enrichment**
CGCG *	52	12.05	4.236	CCCC	108	13.37	8.074
ACGT	53	12.5	4.236	CCGC	112	16.60	6.746
CGTG	53	12.5	4.236	CGCC	97	16.60	5.842
GCCC *	50	12.5	4.236	CCCA	94	16.70	5.626
ATAT	46	12.05	4.236	CCTC	80	16.60	4.818
GGCC	44	12.5	4.236	CGCG	76	16.09	4.722
AAAA	34	10.01	4.236	CTCC	75	16.60	4.517
CCCA *	42	12.5	4.236	CGGC	72	16.60	4.337
CACG	39	12.5	4.236	GCGG	64	16.60	3.855
GTCA	37	12.5	4.236	GCCC	63	16.70	3.771
ACCG	36	12.5	4.236	CCAC	60	16.60	3.614
AACC	35	12.5	4.236	GGCG	59	16.60	3.553

* The asterisk indicates shared core sequences in Arabidopsis and rice.

TABLE 3.4: MOST DEPLETED CORE SEQUENCES IN ARABIDOPSIS AND RICE.

Arabidopsis				Rice			
Core sequence	**Observed**	**Expected**	**Enrichment**	**Core sequence**	**Observed**	**Expected**	**Enrichment**
TTCA	2	12.5	-6.254	CTGT	1	16.705	-16.705
TTGC	2	12.5	-6.254	CTTG	1	16.705	-16.705
CTCT	1	12.05	-12.05	GACT	1	16.705	-16.705
GTAG	1	12.43	-12.43	GAGT	1	16.705	-16.705
TAGT	1	12.43	-12.43	GTGC	1	16.705	-16.705
TTGT	1	12.43	-12.43	TCTA	1	16.705	-16.705
CTAT	1	12.5	-12.5	TGAG	1	16.705	-16.705
TAAC	1	12.5	-12.5	TGCC	1	16.705	-16.705
TACG	1	12.5	-12.5	TGTA	1	16.705	-16.705
TATG	1	12.5	-12.5	TTCG	1	16.705	-16.705
TGAA	1	12.5	-12.5	TTGA	1	16.705	-16.705
TTCC	1	12.5	-12.5	TTTC	1	16.705	-16.705

Note: No shared core sequences were found in this category in Arabidopsis and Rice.

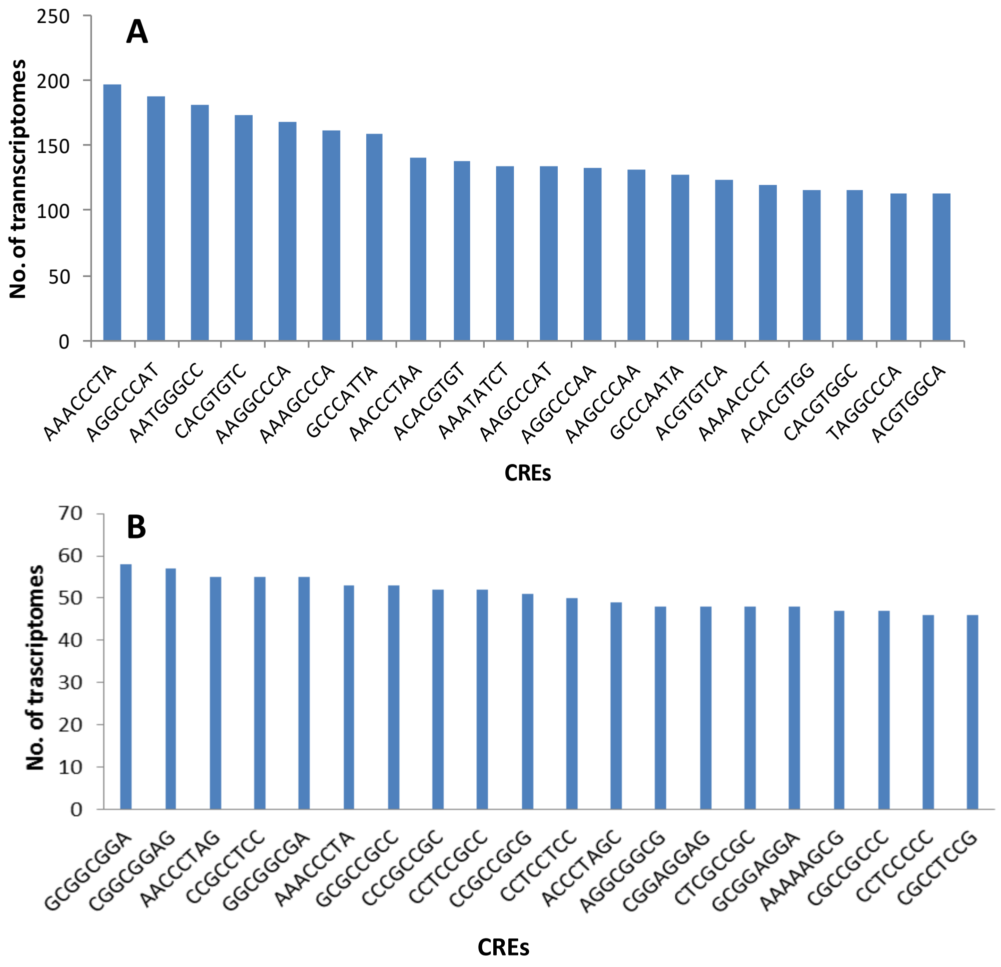


Figure 3.1 CREs response breadths in Arabidopsis and rice. A) Top 20 CREs with broadest response to experimental conditions in Arabidopsis. The element AAACCCTA and ACGTGGCA were significantly correlated to 197 and 113 microarrays respectively. B) Top 20 CREs with broadest response in rice. The elements GCGGCGGA and CGCCTCCG were significantly correlated to 57 and 46 microarrays respectively.

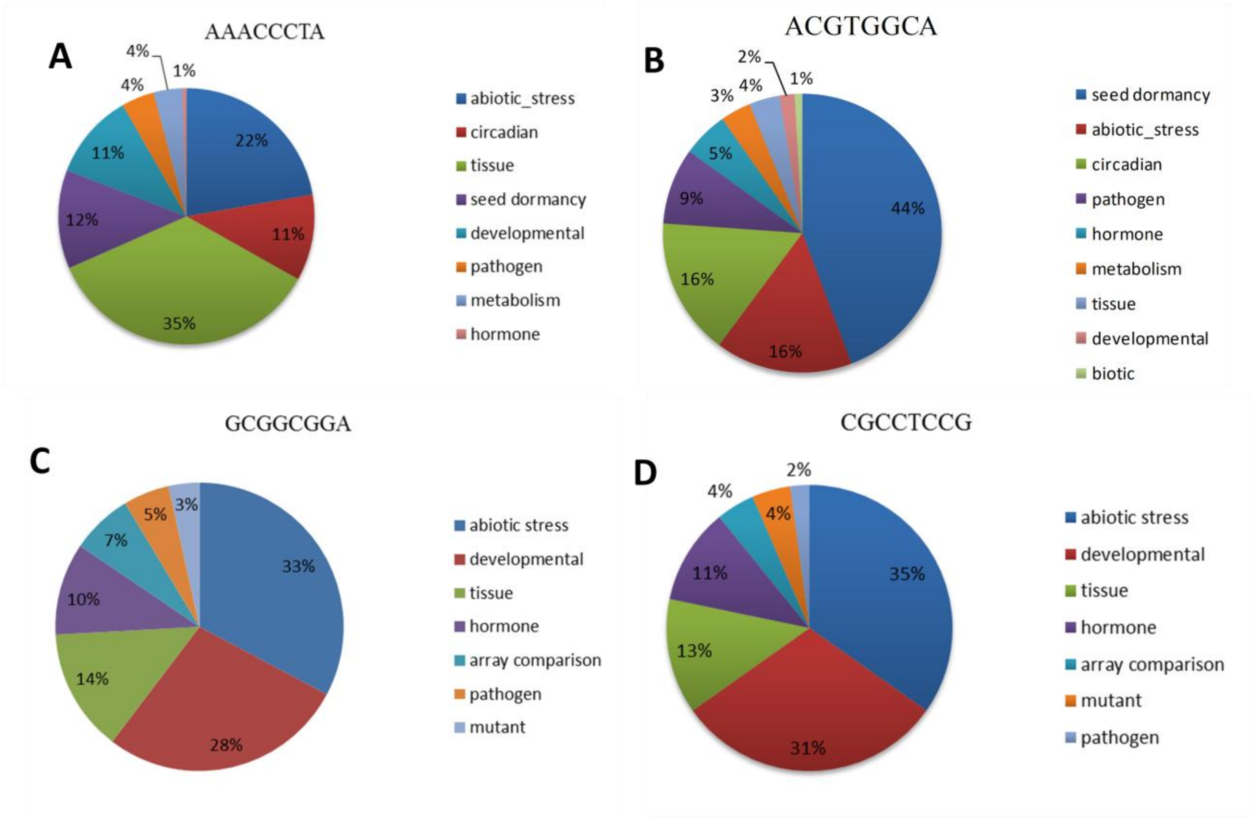


Figure 3.2 Response patterns of selected elements by experimental design. A) More than one third of the microarrays correlated to the element AAACCCTA were of tissue comparison category. B) The element AAACGCGT is most frequently correlated to developmental stages, but not tissue comparison microarrays. (e and f) Response patterns of GCGGCGGA, CGCCTCCG. In both cases, the abiotic stress represented highest percentage of microarrays.

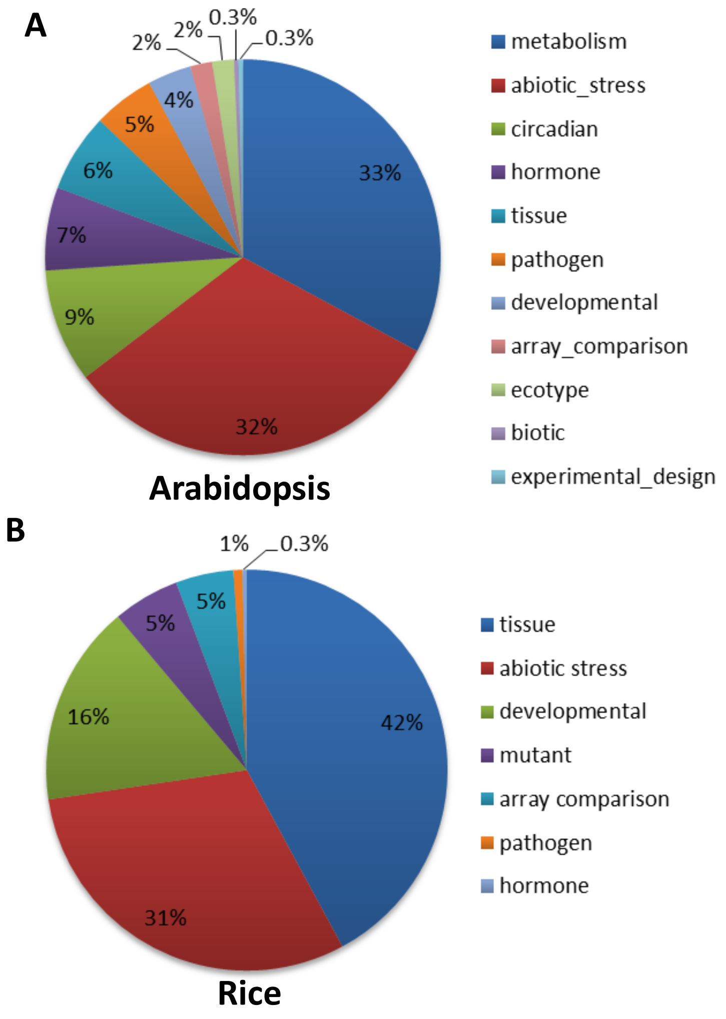


Figure 3.3 Experimental designs correlated with single-response CREs in Arabidopsis and rice. (A) Response pattern of 291 CREs that responded to only a single experimental design in Arabidopsis. The majority of this class of CREs responded to metabolism and abiotic stress conditions. (B) Response pattern of single-microarray CREs in rice. Note the increase in tissue comparison experiments with single tissue responsive CREs in rice (42%) when compared to Arabidopsis (6%).

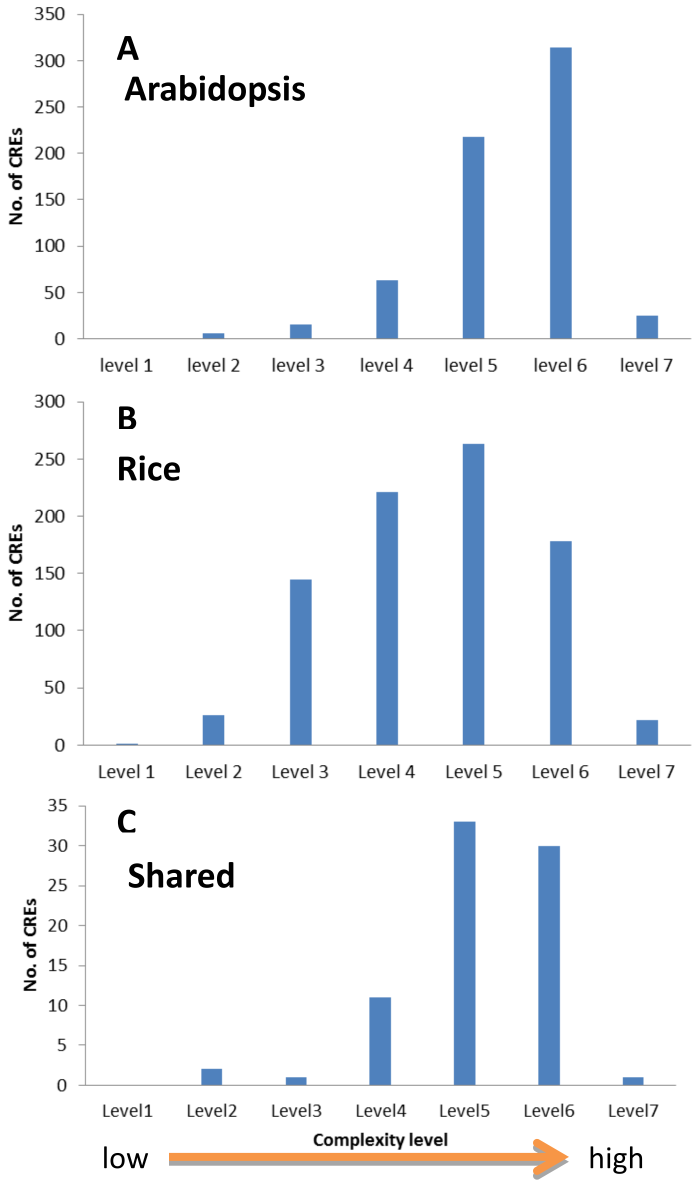


Figure 3.4 Complexity levels of CREs. Complexity of CREs was divided into seven levels (see methods). Level one representing least complex CREs (single nucleotide, eg. AAAAAAAA), level seven comprises most complex (CREs with two of each nucleotide, eg. CATGATCG). (A) No level 1 CREs was observed in Arabidopsis, and the majority of the xxx observed CREs were of level 6. In rice one CREs had level 1 complexity, and the majority of CREs were of level 5 (B). 80 shared CREs occurring in both rice and Arabidopsis were from complexity levels 2-7, with a modal complexity of 5.

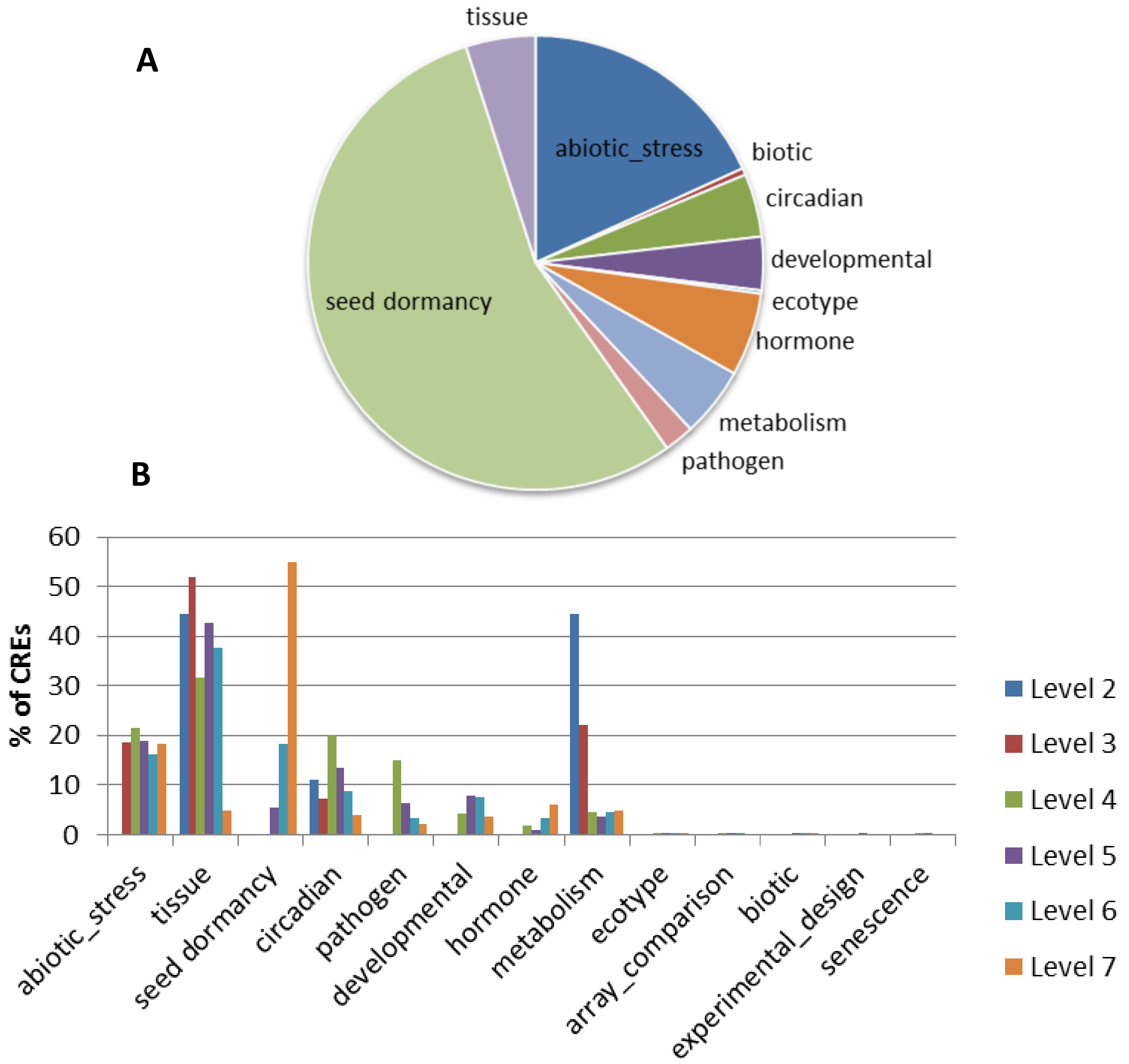


Figure 3.5 Experimental designs correlated with CREs of different complexity level. (A) level 7 complex CREs in Arabidopsis, seed dormancy and abiotic stress comprised the majority of correlations to high complexity CREs. (B) Correlation of different complexity CREs to experiments by experimental design. Experiments on metabolic pathways correlated with low complexity CREs, while pathogen and circadian rhythm correlated with medium complexity CREs.

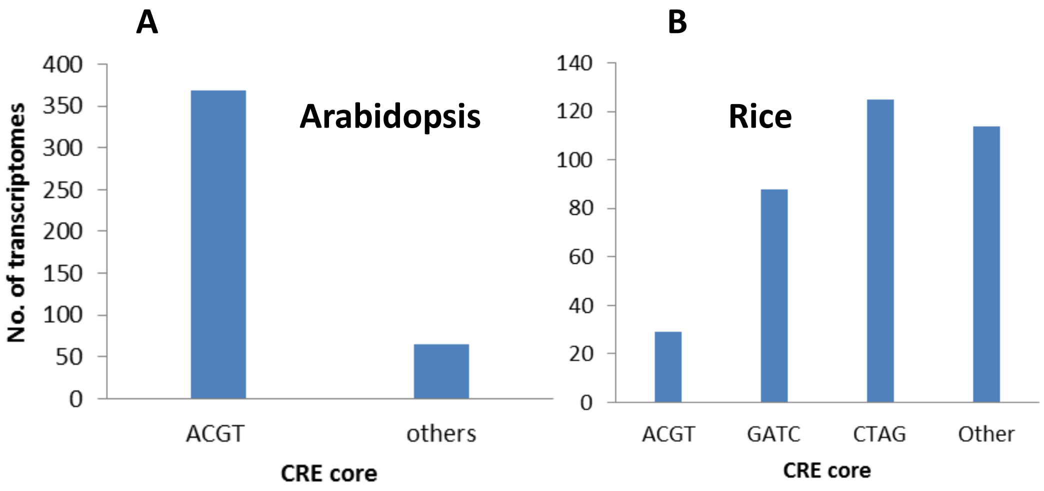


Figure 3.6 Common cores found in high complexity CREs. ACGT was the most common CRE core (4 letter common feature) among the most complex (level 7 complexity) CREs. Prevalence is shown by the number of experiments they in Arabidopsis. In rice (right panel) 3 common cores including ACGT were found in high complexity CREs in rice, however GATC and CTAG cores were more prevalent than in Arabidopsis.

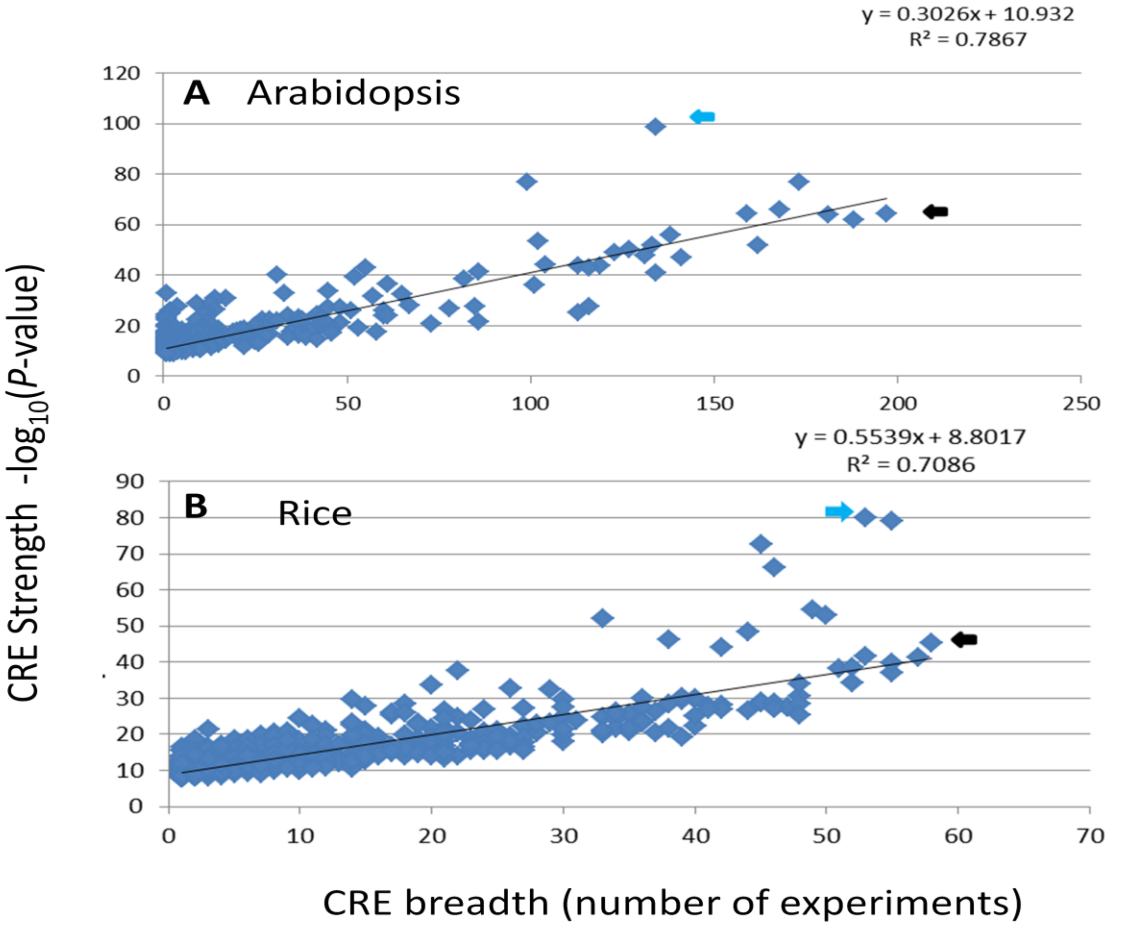


Figure 3.7 Correlation of *p*-value and response breadth. (A) In general increased p-value accompanied with increases response breadth in Arabidopsis. However, EVE element (blue arrow) with highest p-value has narrower response than AAACCCTA (black arrow). (B) Similar pattern observed in rice, however, AAACCCTA with highest p-value has narrower response than GCGGCGGA.

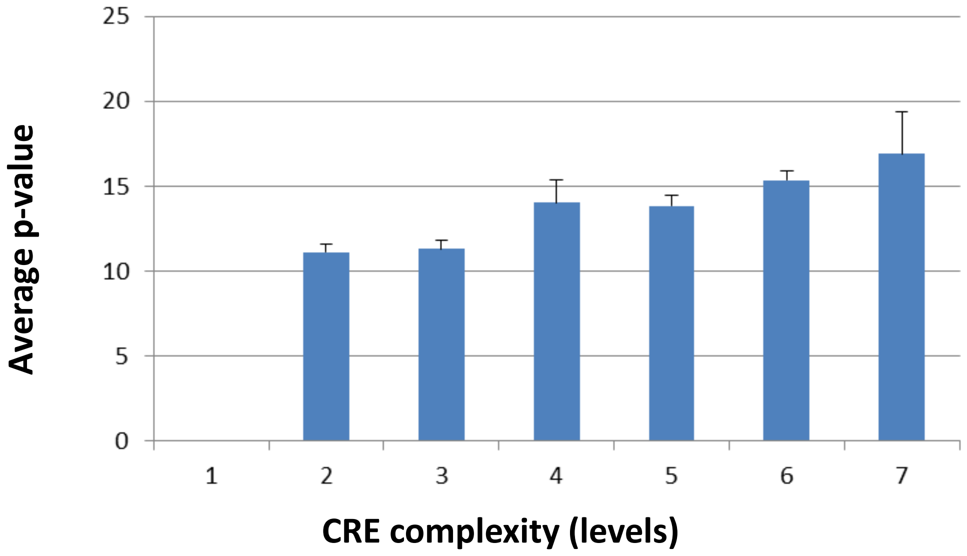
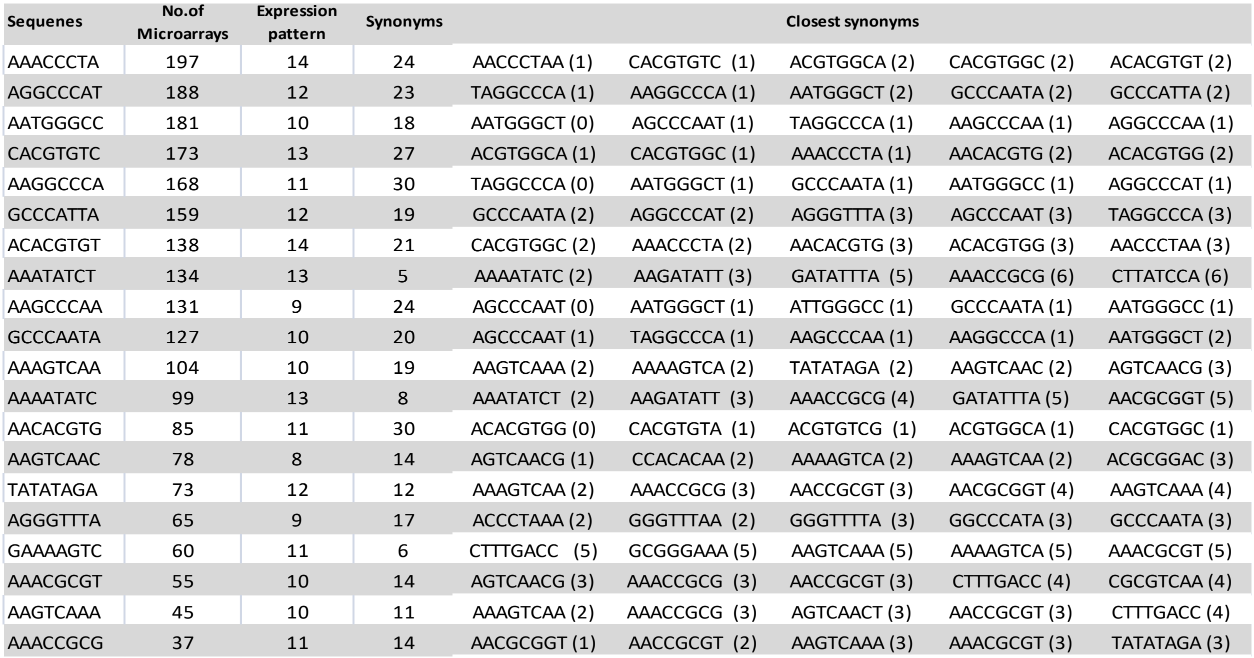


Figure 3.8 CREs correlation strength (*P*-value) versus complexity levels. The average of highest scoring *P*-values (Y-axis shows CRE Strength as -log10(P-value)) of CREs from a hypergeometric test correlating the presence of the element to differential expression by experiment was tested for CREs in each complexity level. This P-value was considered to show the strength of CREs at gene regulation (higher likelihood of gene regulation). The strength is increasing while CREs structure becomes more complex.

sequence	mismatch	frequency	complexity level	cold	drought	osmotic	salt	heat	oxidative	wound	CO_2	dark	ozone	light	uv-light	flood	biotic	circadian	developmental	ecotype	hormone	metabolism	pathogen	seed dormancy	senescence	tissue
CACGTGTC	0	173	L-6	🟩	🟩	🟩	🟩	🟩				🟩					🟩	🟩	🟩		🟩	🟩	🟩	🟩		🟩
CACGTGGC	1	116	L-6	🟩	🟩	🟩	🟩	🟩										🟩	🟩		🟩	🟩	🟩	🟩		🟩
CACGTGTA	1	61	L-7	🟩		🟩	🟩											🟩	🟩		🟩	🟩		🟩		🟩
AACGTGTC	1	39	L-7			🟩															🟩	🟩		🟩		🟩
CACGTTTC	1	2	L-6			🟩																	🟩			
ACACGTGT	0	138	L-7	🟩	🟩	🟩	🟩	🟩			🟩				🟩			🟩	🟩		🟩	🟩	🟩	🟩		🟩
ACACGTGG	1	116	L-6	🟩	🟩	🟩	🟩	🟩			🟩				🟩		🟩	🟩	🟩		🟩	🟩	🟩	🟩		🟩
ACACGTAT	1	15	L-6	🟩		🟩	🟩														🟩	🟩		🟩		🟩
AAACGTGT	1	13	L-6	🟩	🟩		🟩						🟩					🟩			🟩		🟩			
ACACGTGA	1	6	L-6			🟩													🟩		🟩	🟩	🟩			
ACACGCGT	1	3	L-6	🟩																	🟩	🟩				
ACACGTGC	1	1	L-6																			🟩				
AACCGCGT	0	82	L-6	🟩	🟩				🟩	🟩			🟩		🟩	🟩		🟩	🟩		🟩	🟩	🟩			🟩
AAACGCGT	1	55	L-6	🟩					🟩	🟩			🟩			🟩		🟩	🟩		🟩	🟩	🟩			🟩
AAGCGCGT	1	9	L-6	🟩						🟩						🟩		🟩	🟩			🟩				
AAATATCT	0	134	L-5	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩			🟩	🟩			🟩	🟩	🟩			🟩			
AAATATAT	1	2	L-4																							🟩
AAATATCA	1	5	L-4	🟩	🟩		🟩		🟩																	
AAATATCC	1	4	L-5	🟩	🟩					🟩																
AAACCCTA	0	197	L-5	🟩	🟩	🟩	🟩	🟩			🟩	🟩	🟩					🟩	🟩		🟩	🟩	🟩	🟩		🟩
AAACCCTT	1	10	L-6																🟩							🟩
AAGCCCTA	1	4	L-6																							🟩
AAGCCCAA	0	131	L-5	🟩						🟩	🟩	🟩						🟩	🟩			🟩		🟩		🟩
AAGCCCAT	1	134	L-6	🟩			🟩			🟩	🟩	🟩						🟩	🟩			🟩	🟩	🟩		🟩
AGGCCCAA	1	133	L-6	🟩			🟩			🟩	🟩	🟩						🟩	🟩			🟩	🟩	🟩		🟩
AAGCCCTA	1	4	L-6																							🟩
GAAAAGTC	0	60	L-5	🟩			🟩		🟩		🟩		🟩		🟩			🟩	🟩		🟩	🟩	🟩		🟩	🟩
AAAAAGTC	1	7	L-4																				🟩			
AAAGTCAA	0	104	L-4	🟩	🟩		🟩						🟩		🟩			🟩	🟩		🟩	🟩	🟩		🟩	🟩
CAAGTCAA	1	3	L-5																			🟩	🟩			
GGGCCTTA	0	48	L-6								🟩	🟩							🟩					🟩		🟩
GGGCTTTA	1	43	L-6								🟩	🟩										🟩		🟩		🟩
GGGCCTAA	1	38	L-6	🟩						🟩		🟩						🟩								🟩
CTTATCCA	0	37	L-6	🟩		🟩	🟩		🟩	🟩	🟩							🟩				🟩	🟩			

Figure 3.9 homonymic analysis of CREs in Arabidopsis. Ten CRE and their homonyms with their expression pattern in Arabidopsis. The original CREs are in bold text and homonyms are in

non-bold texts. Mismatch degree is indicated by 1 and the 0 refers to original CREs. Frequency is the response breadth (number of experiments hit by the CREs). L illustrates complexity levels and the numbers refer to the certain levels of complexity. The green color indicates the experiment category hit by the CREs, the pale yellow color indicates experiment categories missed by the CREs.



Sequenes	No.of Microarrays	Expression pattern	Synonyms	Closest synonyms				
AAACCCTA	197	14	24	AACCCTAA (1)	CACGTGTC (1)	ACGTGGCA (2)	CACGTGGC (2)	ACACGTGT (2)
AGGCCCAT	188	12	23	TAGGCCCA (1)	AAGGCCCA (1)	AATGGGCT (2)	GCCCAATA (2)	GCCCATTA (2)
AATGGGCC	181	10	18	AATGGGCT (0)	AGCCCAAT (1)	TAGGCCCA (1)	AAGCCCAA (1)	AGGCCCAA (1)
CACGTGTC	173	13	27	ACGTGGCA (1)	CACGTGGC (1)	AAACCCTA (1)	AACACGTG (2)	ACACGTGG (2)
AAGGCCCA	168	11	30	TAGGCCCA (0)	AATGGGCT (1)	GCCCAATA (1)	AATGGGCC (1)	AGGCCCAT (1)
GCCCATTA	159	12	19	GCCCAATA (2)	AGGCCCAT (2)	AGGGTTTA (3)	AGCCCAAT (3)	TAGGCCCA (3)
ACACGTGT	138	14	21	CACGTGGC (2)	AAACCCTA (2)	AACACGTG (3)	ACACGTGG (3)	AACCCTAA (3)
AAATATCT	134	13	5	AAAATATC (2)	AAGATATT (3)	GATATTTA (5)	AAACCGCG (6)	CTTATCCA (6)
AAGCCCAA	131	9	24	AGCCCAAT (0)	AATGGGCT (1)	ATTGGGCC (1)	GCCCAATA (1)	AATGGGCC (1)
GCCCAATA	127	10	20	AGCCCAAT (1)	TAGGCCCA (1)	AAGCCCAA (1)	AAGGCCCA (1)	AATGGGCT (2)
AAAGTCAA	104	10	19	AAGTCAAA (2)	AAAAGTCA (2)	TATATAGA (2)	AAGTCAAC (2)	AGTCAACG (3)
AAAATATC	99	13	8	AAATATCT (2)	AAGATATT (3)	AAACCGCG (4)	GATATTTA (5)	AACGCGGT (5)
AACACGTG	85	11	30	ACACGTGG (0)	CACGTGTA (1)	ACGTGTCG (1)	ACGTGGCA (1)	CACGTGGC (1)
AAGTCAAC	78	8	14	AGTCAACG (1)	CCACACAA (2)	AAAAGTCA (2)	AAAGTCAA (2)	ACGCGGAC (3)
TATATAGA	73	12	12	AAAGTCAA (2)	AAACCGCG (3)	AACCGCGT (3)	AACGCGGT (4)	AAGTCAAA (4)
AGGGTTTA	65	9	17	ACCCTAAA (2)	GGGTTTAA (2)	GGGTTTTA (3)	GGCCCATA (3)	GCCCAATA (3)
GAAAAGTC	60	11	6	CTTTGACC (5)	GCGGGAAA (5)	AAGTCAAA (5)	AAAAGTCA (5)	AAACGCGT (5)
AAACGCGT	55	10	14	AGTCAACG (3)	AAACCGCG (3)	AACCGCGT (3)	CTTTGACC (4)	CGCGTCAA (4)
AAGTCAAA	45	10	11	AAAGTCAA (2)	AAACCGCG (3)	AGTCAACT (3)	AACCGCGT (3)	CTTTGACC (4)
AAACCGCG	37	11	14	AACGCGGT (1)	AACCGCGT (2)	AAGTCAAA (3)	AAACGCGT (3)	TATATAGA (3)

Figure 3.10 Synonymic analysis of CREs in Arabidopsis. A list of 20 CREs in Arabidopsis and their synonyms. Response breadth is illustrated by number of experiments. Expression pattern indicates the categories of experiments hit by CREs. The number of synonyms indicates the number of CREs which show similarity in expression pattern to the main CREs by more than 50%. The closest synonyms refer to top five synonyms related to the main CREs and the number between the brackets indicates the number of mismatches in expression pattern between main CREs and their synonyms.

CHAPTER 4

ABIOTIC STRESS, TISSUE, AND DEVELOPMENTAL CRES

4.0 Introduction

Environmental conditions play a great role in plant growth and development. Since plants are sessile organisms, resistance to adverse environmental conditions (abiotic stress) could be a challenge. Response to such conditions may involve hormonal and non-hormonal signal transduction (Yoshida et al. 2014), which can lead led to massive adjustment of transcriptional gene expression (J. et al. 2017) and protein synthesis (Nouri et al. 2015). Furthermore, with having their own response pathways, the various abiotic stresses may present interactional gene regulation and sharing CREs (Yamaguchi-Shinozaki and Shinozaki 2005). Exposure of plants to such abiotic stress may vary during their life cycle or throughout the year. However, certain abiotic stresses may occur repeatedly throughout a plant's life cycle and may have virtual and destructive effects on crop yields in case of crops (Mittler 2006). Cold stress, in particular, is the most common abiotic stress (Yadav 2010) which may appear repeatedly, and due to climate change and increased temperature of the globe, drought stress becomes a critical abiotic stress which could limit plant's yield.

The identification of molecular bases of abiotic stress resistance of plants were main goals of plant scientists. Extensive research studies resulted in the discovery of CREs that play critical roles in gene expression regulation during abiotic stress. ABRE (Hattori et al. 2002) and DRE (Yamaguchi-Shinozaki and Shinozaki 1994) are two well-known CREs identified experimentally and *in silico*. ABRE (ACGTGTC) is a major abiotic stress CRE in ABA-dependent gene expression pathway. Due to increased concentration of ABA during drought and high salinity, the AREB/ABF (ABRE binding protein) (Choi et al. 2000; Uno et al. 2000a), interacts with ABREs and trigger ABA-dependent pathway. In Arabidopsis two ABRE element is necessary for full activation and response, while in rice a coupling element is required (Hobo et al. 1999). DRE (RACCGACAT), on the other hand is another critical CRE that plays an important role in abiotic stress through responding to cold, drought and ABA. However, it may still be induced even with the absence of ABA. This makes this element a crosstalk between ABA-dependent and ABA-independent pathway. Crosstalk is not an abiotic stress specific feature, but it may occur between abiotic stress and other biological and developmental processes. The well-known Evening Element (AAAATATC) (Harmer et al. 2000b), involved in regulation of gene expression in circadian genes, were also reported to be induced by cold stress through coupling with ABRE like element (Mikkelsen and Thomashow 2009). Abiotic stress may also be critical in developmental processes during a plant's life such as flowering. Exposure to cold stress (low temperature), a phenomenon called vernalization, is necessary for repression of FLC which eventually leads to flowering (Amasino 2004). On the other hand, gene expression may occur regardless of the environmental condition, but requires appropriate time, the right place and appropriate abundance. This could be essential in producing transgenic plants with increased expression of such genes or CREs. The two CREs GSE1 and GSE2 were reported to be positively regulate gene expression in all green tissues in rice (Ye et al. 2012).

Here in this chapter, we illustrate the combination and crosstalk between CREs involved in the abiotic stress and various biological processes. In addition, we focus on the CREs that are specific for the above categories, and the complexity levels of CREs that may determine CREs specificity to the above mentioned categories.

4.1 Materials and methods

4.1.1 Microarray experimental data sources

In abiotic stress transcriptomes, the AtGenExpress project data were considered (Joachim et al. 2007), due to uniform conditions of the experiments, which means there were other non-AtGenExpress abiotic stress transcriptomes. However, excluded due to different experimental conditions and methodology. In AtGenExpress project, Arabidopsis seeds were sterilized then nine seeds were placed on floats closed by vented lid , which later was transferred into transparent growth boxes containing MS media supplemented with B5 vitamins, 0.5% sucrose, 0.5% agar and pH 5.7. Plants were grown at 24ºC and 16 h light intensity of 150 µmol.photons m^{-2} sec^{-1} and 8 h dark regime. Eighteen day old seedlings were used to initiate stress treatment. For each stress treatment, 0 h , 0.5 h, 1 h, 3 h, 6 h, 12 h and 24 h time points were considered, except for UV-B light stress, in which stress application lasted for only 15 min. Later, root and shoot tissues were harvested for RNA extraction.

4.1.2 Transcriptome categorization

Depending on the type of the experiments conducted, current transcriptomes were categorized into abiotic stress, tissue, seed dormancy, circadian and developmental transcriptomes. For abiotic stress plants subjected and treated with certain abiotic stresses like cold, drought, osmotic, salt, wound, oxidative light, CO_2 and dark stress (Joachim et al. 2007). Tissue transcriptomes include comparison of gene expression in any tissue or organ. Developmental transcriptomes include tissue and organ comparison in various life cycles. Studying differential gene expression in tissues and organs with respect to time categorized as circadian transcriptomes. Seed dormancy category represent imbibing seeds for different time course and varied environmental and illumination conditions (Fitzek 2012).

4.1.3 Generation of heat maps for correlation by shared CREs

Identification of shared elements among different abiotic stresses and abiotic stress and developmental related categories were indentified using Microsoft Excel program. A VLOOKUP function (=vlookup(Ax, list range, column index number, False (exact match)) was used to determine shared CREs among all the categories. Such CREs are considered as multi-stress in case of abiotic stress, and multi-category CREs in case of other categories. Simultaneously, CREs not shared among any of the above mentioned categories are uni-stress CREs in case of abiotic stress, and category specific CREs in case of other categories. For abiotic stress sub-categories, shared and non shared elements were displayed using heat map (conditional formatting) tool in Excel.

4.1.4 Time series analysis

As mentioned earlier, abiotic stresses were conducted in different time courses. Therefore, the data contained CREs correlated to induction or suppression of different time series with varied strength (p-value). CREs vs time series data were clustered using a free version of K-mean clustering with centered correlation parameters in Cluster 3.0 (de Hoon et al. 2004). The output was various clusters depending on the number of CREs in each stress.

4.1.5 CRE complexity analysis

CRE complexity was measured the same as mentioned in Chapter 3.

4.2 Results and discussion

4.2.1 Abiotic Stress cross talk determined by related CREs

In this study, abiotic stress microarrays corresponding to 10 different categories of environmental variables (including temperature, osmotic, oxidative, and wounding) were studied for their interactions among CREs. Of the 1460 microarrays surveyed 803 (55%) showed strong significant correlation (p-value $<10^{-8}$) for differential regulation to at least one of the 65,536 8-word CREs tested. Of these 803, a subset of 200 microarrays were selected in which an abiotic stress experiment was performed, and these were further divided into categories of abiotic stress. The 10 most common abiotic stress categories were selected to further investigate CREs involved in such stresses and to explore the interaction among the stresses through shared CREs.

Each CRE was scored for strong significant correlation with differential regulation by each of 10 categories of stress on at least 1 microarray experiment of that condition using a chi-squared test. The number of CREs per each individual abiotic stress category is illustrated in Figure 4.1. Most CREs were correlated to regulation by multiple stresses, indicating that these points within the network may represent cross talk or integrating nodes. Stress categories were linked by the number of CREs that showed regulation by both stresses (Figure 4.1). Some stress comparisons, such as wounding, cold and drought showed high numbers of shared CREs. Due to the great number of the stresses, drawing a Venn diagram to represent all the interactions was not suitable. As shown in Figure 4.1, salt and osmotic stress shared the highest number of CREs with and enrichment value of 7.78, followed by dark stress and CO2 stress with an enrichment value of 7.85. . The CO_2 and dark categories are referred to plants grown under different levels of CO_2 and 96 hour dark period under 22ºC and 4ºC. In contrast, the least abiotic stress to share CREs was light stress in which plants kept in dark for 16 hours and 8 hours of light intensity at either 400 or 100 μmol m-1 s-1. Light level stress shared three CREs with cold and drought stress (AAATATCT, CCTTATCC, and ACCTTATC) and one multi-stress CRE (AAATATCT) that correlated to all stresses except for dark stress. This might indicate responding to light stress is governed by specific CREs apart from other abiotic stresses and the pathways do not cross-talk very much. The multi-stress responsive element identified in this study that light stress shared with other abiotic stresses (except dark) has significant sequence overlap with an experimentally characterized known element. The 8-letter word discovered here (AAATATCT), is a 1 bp shorter version of original reported evening element AAAATATCT responsible for circadian rhythm (Mikkelsen and Thomashow 2009). Shared CREs between cold and wound stress was also highly enriched (4.12). Since cold and wound stress include membrane damage and electrolyte leakage, the response and repair machinery might be similar as well. This means similar gene expression regulation for cold and wound stress. Furthermore, abiotic stresses that might crosstalk in response pathway showed a high enriched value for shared CREs. For example, cold stress and drought stress had a high enrichment value of 4.53 for shared CREs. In contrast, abiotic stress that have different response pathways like drought stress and dark stress, had very low enriched value (1.83). There were only two incidents of zero sharing of CREs between stresses; these were between dark and oxidative stress, and dark and light stress.

4.2.2 Temporal response of abiotic stress CREs in Time Series Experiments

The AtGenExpress data included in the transcriptomic data obtained for this study had valuable time series experiments that could help us understand the way CREs can respond to stimuli (stress) depending on the duration of stress. Here we analyzed a series data for certain abiotic stress experiments in Arabidopsis. Time series data helps place the order of events into early, middle and late phases of a signal transduction chain to the downstream proteins that help the plant survive the stress. The hypothesis is that earlier events in a specific stress stimulus response will be more specific to that stress, and that cross talk and integration steps in multiple stress response pathways come at middle and later time points. Thus, CRE that responds to multiple stresses would do so at later time points than a CRE that responds specifically to that stimulated stress, as measured by the group response of the cohort of genes containing these

CREs. Unfortunately, time series data is difficult and expensive to generate, and as of yet, such datasets only exist for Arabidopsis, and no such data has been generated for rice.

4.2.2.1 Cold stress temporal response pattern.

In cold stress, there were 91 CREs that correlated to differential regulation in at least 1 time point in a 24 hour cold stress measured at 0 (time control), 0.5, 1, 3, 6, 12 and 24 hours after shifting plants to 4ºC. These were compared to plants in identical conditions but held at room temperature. Separate transcriptomes were generated for shoot and root tissues. In this entire series, 84 CREs correlated with differential expression in at least one time point in the shoot transcriptomic data and 23 in roots. There were an overlap of 16 CREs between shoot and root. The behavior of these CREs were quite different in shoot and root tissue, indicating some tissue specificity of the response pathway, or the promoter context. Due to the low number of responsive CREs in root tissues, only CREs in shoot tissue were considered for clustering by type. The 84 CREs were clustered using K-mean clustering (Clustal 3.0) software, and sorted according to their p-values. This resulted in 16 different clusters of CREs which were separated into three graphs (Figure 4.2A,B,C and D) by correlation pattern with time. Thirteen CREs were found to be correlated with differential regulation of expressed genes in the first 30 minutes of exposure to cold stress However one of them (ATTATATA) was specific to this time point, while the other 12 CREs were also correlated with differential regulation at other time points. Similar number of CREs was responded to cold stress after one and three hours of exposure (14 CREs). Furthermore, there were only three CREs expressed (turned on) only after one hour exposure, followed by expression of four specific CREs of 3-hour time point. Further exposure to cold stress accompanied by increasing numbers of expressed CREs. In the 6-hour time point, the number of CREs responded to cold stress increased to 22, then doubled in 12 hour exposure

42 CREs. Despite the larger number of induced CREs in 6-hr time point, only 2 time point specific CREs was observed. Surprisingly, time point specific CREs increased to 10 after 12 hr of cold exposure. Induced number of CREs increased further at 24-hour time point. It was observed that 58 CREs induced after 24 hour of exposure to cold stress with 31 time point specific time point CREs.

Sharing of CREs among time points was different than expected based on crosstalk sharing, and was highly time point independent. There were only four CREs (AAAGTCAA, AAGTCAAC, ACCTTATC, CCACACAA) shared by all time points, indicating those CREs were switched on during the stress without interruption. On the other hand, there were three CREs (AAACCCTA, AACCCTAG, ACGTGGCA) shared by all time points except for one. In addition, the shared CREs had the same regulatory effect (either induction or suppression of genes) in all time points they correlated to. In other words, if an element induced genes in one time point, it remained correlated to gene induction (and not suppression) in other time points if it correlated at all.

4.2.2.2 Drought stress temporal response pattern

Time series response pattern was illustrated for drought stress in the same way as cold stress. In drought stress, 42 CREs were correlated to differential regulation in at least 1 time point in a 24 hour drought stress measured at 0, 0.25, 0.5, 1, 3, 6, 12 and 24 hours. Among 42 CREs, 33 were correlated with differential expression in shoot transcriptomic data and 9 in roots. For the same reason as in cold stress, root CREs were not clusters, while clustering shoot CREs resulted in 8 clusters as shown in Figure 4.3A. Unlike cold stress, the distribution of CREs in time points were sparse and scattered and there was one more time point in drought stress, 0.25 hour. In 0.25-hour time point, 11 CREs were differentially expressed, with only one CRE specific to this time points. the involvement of such number of CREs in the first 15 min of exposure to drought stress may be to the immediate protective action the plant must follow to reduce water loss through closing stomata. The number of CREs increased to 17 in both 0.5 and 1 hour with 2 and 1 specific time point respectively. CREs numbers declined gradually to 16, 14, 13, and 4 in 3, 6, 12 and 24 hours time points respectively. The number of time specific CREs remained as low as before, however, the last time point showed no specific CREs. This was the main, but not the only one, difference between drought and cold stress. As mentioned earlier, in cold stress there were 31 time point specific CREs after 24 hour of exposure. Furthermore, the very low number of CREs involved in 24 hour time point was another distinguishing feature between the two stresses.

Sharing elements were also present in drought stress. Out of 33 CREs, only 2 of them (ACCTTATC and CCACACAA) were shared by all time points. Sharing elements between time points was varied and time point specific. Time points 0.5 and 1 hour shared the 5 CREs, the number of sharing elements between time points. Another feature of sharing elements were switching on and off CREs up to half way of time points. The element AACCCTAA was switched on in the first three time points then off in the rest, while, AGTGGTCC switched on only in the last four time points.

Drought and cold stress share 27 CREs as shown in Chapter 3 (Fig 4.1.). However, only 25 CREs were shared in shoots, and the time pattern of expression of some of these shared CREs was compared. First, among the 4 CREs involved in 24 hour time point (Figure 4.4), only ACCTTATC and CCACACAA have similar time pattern in drought and cold stress. In both cases these two CREs induced genes at all time points. Other CREs had differing effects by time, for example AACCACAC correlated to genes responded to drought stress at 0.5, 3, and 24-hour time points in drought stress, while in cold stress this element acts only at the 1-hour time point. There are also some elements that only act in one stress during the 24-hour time series. This indicated that these elements may act at later time points, or responded to other experimental parameters (ie. soil vs. gel-based experiments). Extreme shifting in time points in both stresses was also observed. TATATAGA was expressed in a single, but different, time point in either stresses. In drought stress, it was expressed very early, after 0.5 hour of exposure to drought, while, in cold stress it expressed after 24 hour of cold exposure. Another example, AAGTCAAA was expressed only after 3 hour of exposure to drought stress, while it required 24 hour of exposing to cold to switch on. This shifting in time may indicate something about how and where the drought and cold pathways intersect. For example if cold temperature results in drying of tissues like 3 hrs of drought, but only after 24hrs exposure, or that the cold stress signaling pathway intersects directly with the drought stress signaling pathway, but with a 21 hour delay. In general the highest proportion of shared elements mostly acted at early time points in drought stress, while in cold stress the same elements shifted to the late time points.

4.2.3 Sequence complexity of CREs involved in cold stress response.

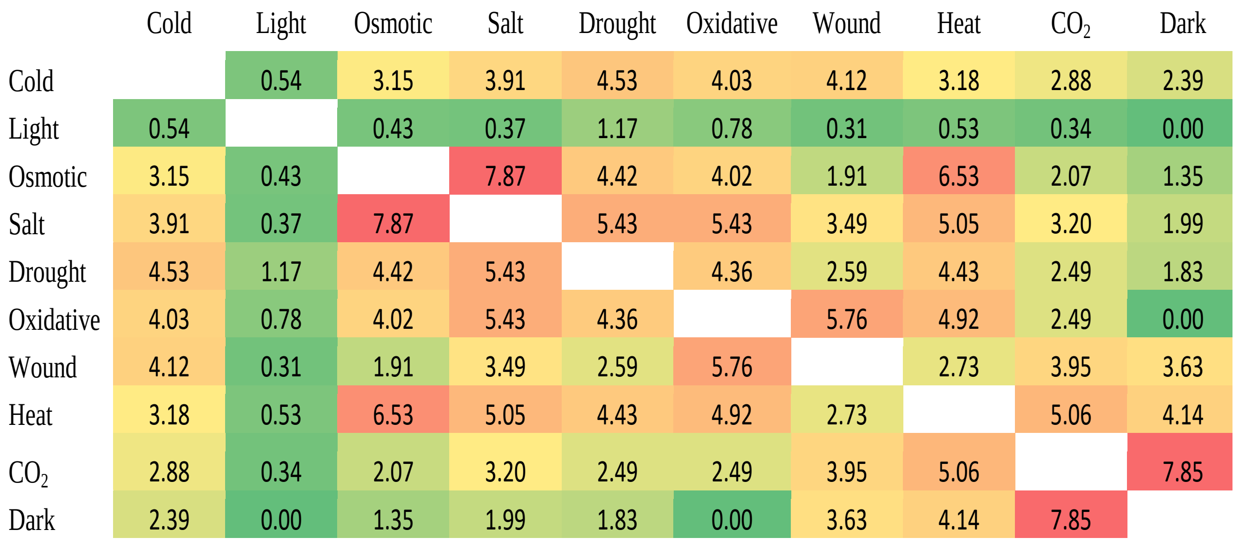
The complexity of Cold CREs were investigated (Figure 4.5A). As mentioned in the previous chapter, complexity refers to the presence of the number of copies of the same nucleotide in a single CRE. As the selected CREs (significantly related to transcriptomes) are subsets of larger sets of mathematically calculated words (65536 words), their complexity pattern may be compared to those of 65K words. As shown in Figure 4.5 B, the number of CREs increased in each level up to level 6, which means a single CRE have 3 copies of the same nucleotide, then it sharply falls down in level 7 (this is mathematically correct). Despite the random selection of CREs depending on significant relationship to transcriptomes, this complexity pattern was reflected and applied to the whole active CREs and cold stress CREs (Figure 4.5C), with the absence of CREs of level 1 and 1 and 2, respectively. However, the percentage of the CREs in the three cases could be slightly different. As demonstrated in Figure 4.3b the distribution of CREs on complexity levels were similarly represented and balanced especially between 65K and active CREs, however there were exceptions. In level 5, cold CREs were underrepresented (28.5%compared to 65K and active CREs 34 %). In contrast, the situation was reversed in further two levels with overrepresentation of cold stress CREs rather than 65K and active CREs. This was clearly observed in level 7 as 6 cold stress CREs were of this level, which means 25% of active CREs of level 7 were involved in cold stress.

The cold stress CREs complexity was further investigated regarding their specificity to cold stress and tissue. The number of CREs involved in a single stress (uni-stress responsive) and multiple stress (multi-stress responsive) varied depending on the abiotic stress (Figure 4.6A). As shown, multi-stress responsive CREs in each abiotic stress significantly outnumbered the uni-stress responsive CREs, except for light stress. The increased number of multi-stress responsive CREs in each abiotic stress category indicates tight correlation of these stresses and the multi-functionality of transcription factors controlling the aforementioned stress. In another word, a single transcription factor may participate in controlling gene expression in a variety of stresses. On the other hand, the presence of uni-stress responsive CREs reveals the fact of presence of abiotic specific transcription factors as well. The complexity of either uni-stress or multi-stress responsive was also investigated. In uni-stress responsive CREs, except for cold, the level 1, 2 and 3 (regarding the fact that level 1 was already absent) were completely depleted, and level 5 and 6 were more dominant and continuously presented throughout the abiotic stresses. In contrast, level 4 and 7 were less dominant their presence was distributed throughout the abiotic stresses. Level 4 CREs were present in five abiotic stresses, while level 7 existed only in three of them, and the presence of the two levels was not mutual. Except for cold stress, the presence of level 4 and 7 were alternative, which means presence of CREs of one of the levels accompanied by the absence of the other one. Moreover, the co-presence of level 5 and 6 were numerically interchangeable. Except for oxidative, wound and carbon dioxide stress, the level 6 CREs were always greater than level 5. These demonstrations reveal that uni-stress responsive CREs are of medium complexity category since level 7 was not frequently present. Similar results were observed in multi-stress responsive CREs with few differences. Level 6 CREs were always dominant level 5 except for light stress (Figure 4.6B). Unlike uni-stress responsive, Level 5 and 7 CREs were more frequently present in abiotic stresses and they were observed in 8 and 6 abiotic stresses respectively. In addition, unlike uni-stress responsive CREs, the presence of level 7 CREs were always accompanied by the presence of level 5 CREs (Figure 4.6C). The presence of more complex CREs in this section could be due to the necessity of the abiotic stresses to such CREs to in order to attract and recruit multifunctional transcriptional factors.

4.2.4 CREs involved in Developmental processes

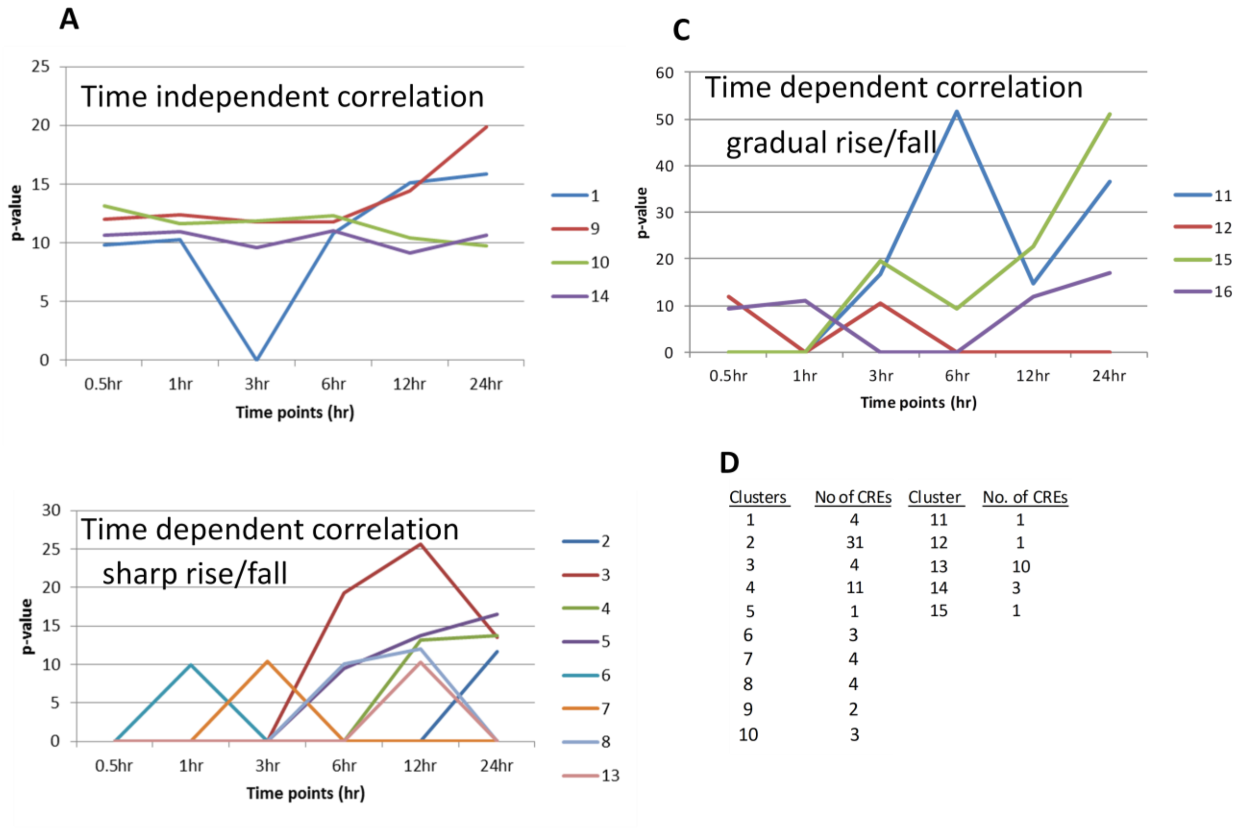
The life of a plant begins with single dormant seed in which after a suitable environmental conditions breaks the dormancy and starts its journey. As a multicellular organism, plants are composed of different tissues and organs, vegetative or reproductive, internal or external aerial or underground. The development and differentiation of such tissues required spatiotemporal gene expression and reprogramming influenced by environmental conditions, which includes induction and suppression of many genes and involvement of various transcriptional factors and regulatory proteins. From dormancy to maturation, various genes may continue participating in development process up to maturity, while other genes are specialized to certain period and developmental stages then destined to certain tissues. Therefore, signaling of genes between plant compartments is common. In this study, the involvement of CREs in whole plant development was investigated. Figure 4.8 illustrates the number of CREs in each category and the crosstalk of CREs between them. Seed dormancy and developmental categories had the highest enrichment value of 4.54, followed by circadian and developmental categories. Despite the high number of CREs in tissue and abiotic stress transcriptomes, they showed the least enriched value for shared CREs (1.03).

The number of CREs according to the transcriptomes were greatly different. As shown in Figure 4.9A, there were 13 tissue specific transcriptomes containing 131 experiments and 228 CREs were significantly correlated to the experiments. Since some CREs responded to the same experiment the frequency of experiments containing all the CREs were 2796. The same observation recorded for other categories (Figure 4.9A). Regarding the number of CREs and experiments, showed most specificity, as with 200 abiotic experiments and 291 abiotic stress CREs, repetition (frequency) of CREs in the experiments were 1317. Furthermore, abiotic stress showed highest number of category specific CREs and, while seed dormancy CREs showed least specificity with only 8 seed dormant CREs as mentioned above (Figure 4.9B).



	Cold	Light	Osmotic	Salt	Drought	Oxidative	Wound	Heat	CO_2	Dark
Cold		0.54	3.15	3.91	4.53	4.03	4.12	3.18	2.88	2.39
Light	0.54		0.43	0.37	1.17	0.78	0.31	0.53	0.34	0.00
Osmotic	3.15	0.43		7.87	4.42	4.02	1.91	6.53	2.07	1.35
Salt	3.91	0.37	7.87		5.43	5.43	3.49	5.05	3.20	1.99
Drought	4.53	1.17	4.42	5.43		4.36	2.59	4.43	2.49	1.83
Oxidative	4.03	0.78	4.02	5.43	4.36		5.76	4.92	2.49	0.00
Wound	4.12	0.31	1.91	3.49	2.59	5.76		2.73	3.95	3.63
Heat	3.18	0.53	6.53	5.05	4.43	4.92	2.73		5.06	4.14
CO_2	2.88	0.34	2.07	3.20	2.49	2.49	3.95	5.06		7.85
Dark	2.39	0.00	1.35	1.99	1.83	0.00	3.63	4.14	7.85	

Figure 4.1 Heatmap for correlation of abitoic stress by number of shared CREs. cis-regulatory elements that correlated to differential expression in at least 1 experiment involving 10 categories of stress are indicated as numbers in the matrix diagonal (blue squares) if they correlated to only 1 stress. The number of CREs correlated to two stresses are placed in the red to green squared at the intersection of the two stresses. Green indicates larger numbers, and red indicates fewer numbers of shared CREs between 2 stresses. Cold is a 24hr treatment of 4C, light is 16 hour dark and 8 hours of 400 and 100 µmol m-1 s-1 light intensity. Osmotic, salt, drought and oxidative are 24hr treatment of 300 mM mannitol, 150 mM NaCl, drought, and oxidative stress respectively. Heat stress indicates plants treated with light period for 3 hour then harvested after different recovery periods. Wound stress indicates injuring plant tissues and harvesting after 24 hrs. CO_2 is treating plants with elevated, low and ambient levels of CO2 per different light intensities. Dark refers to 96 hr dark condition at 22 and 4ºC



Clusters	No of CREs	Cluster	No. of CREs
1	4	11	1
2	31	12	1
3	4	13	10
4	11	14	3
5	1	15	1
6	3		
7	4		
8	4		
9	2		
10	3		

Figure 4.2. Time series analysis of CREs in cold stress. Temporal response of CREs significantly correlated to cold induced genes in Arabidopsis shoot. Eighty four CREs were involved in cold stress response. Colored lines represent different CRE clusters. According to their strength and temporal response, CREs were clustered. One -16 represents the number of clusters. A) CREs that are constitutively responded to cold stress. B) CREs with fluctuating response to cold stress. C) CREs showing no or slight response in early exposure to cold stress. D) Number of CREs in each cluster

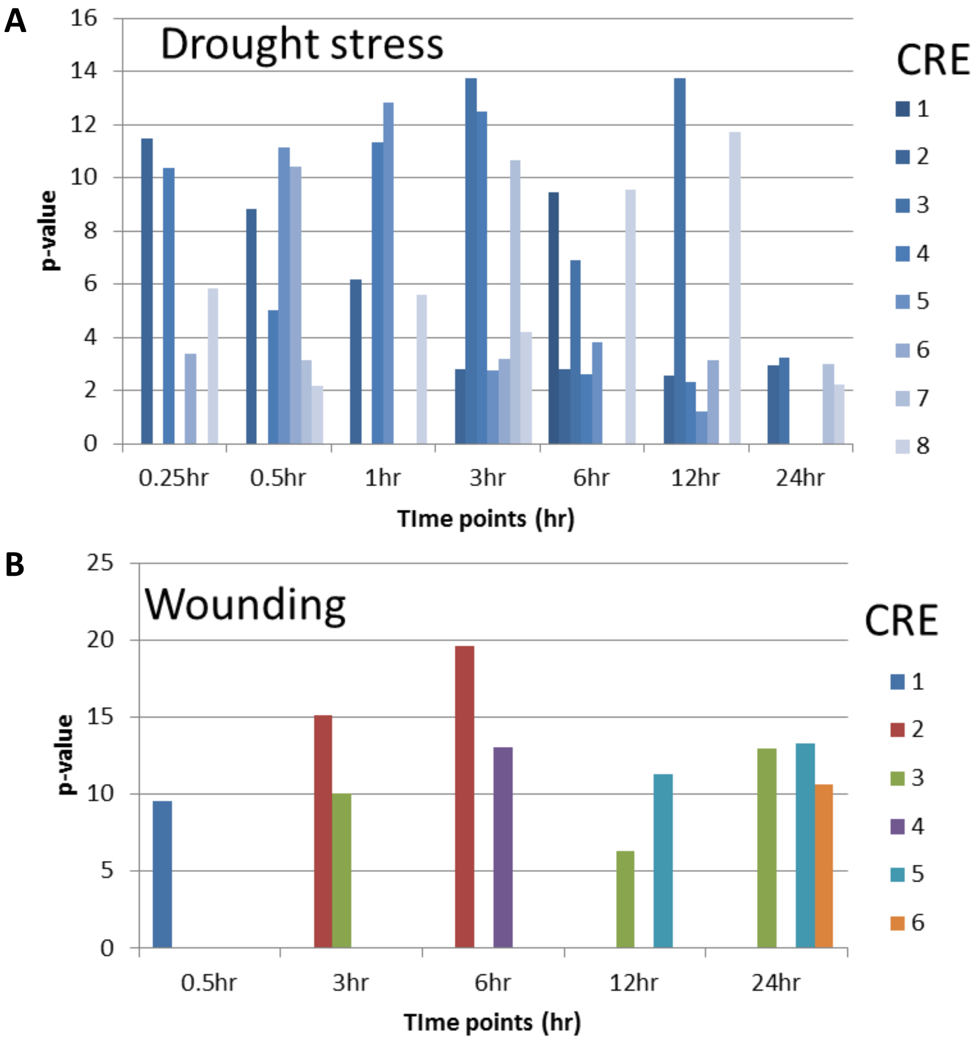
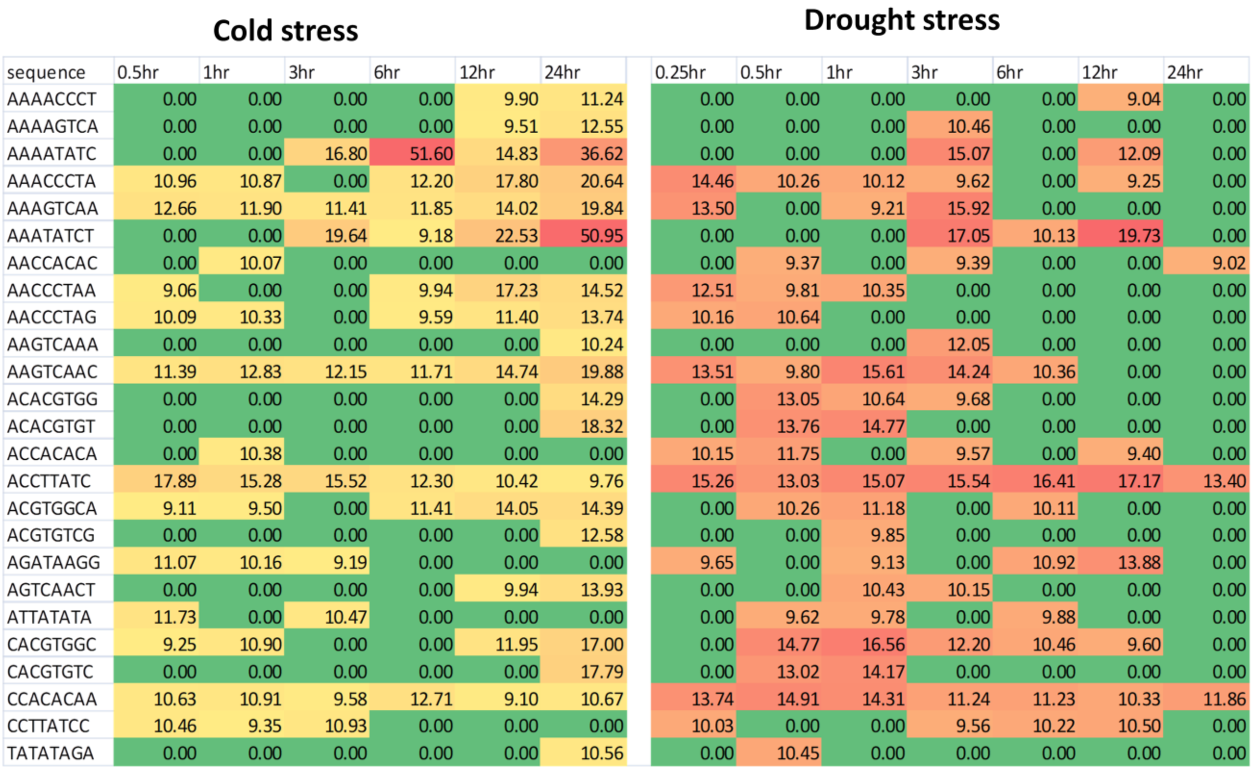
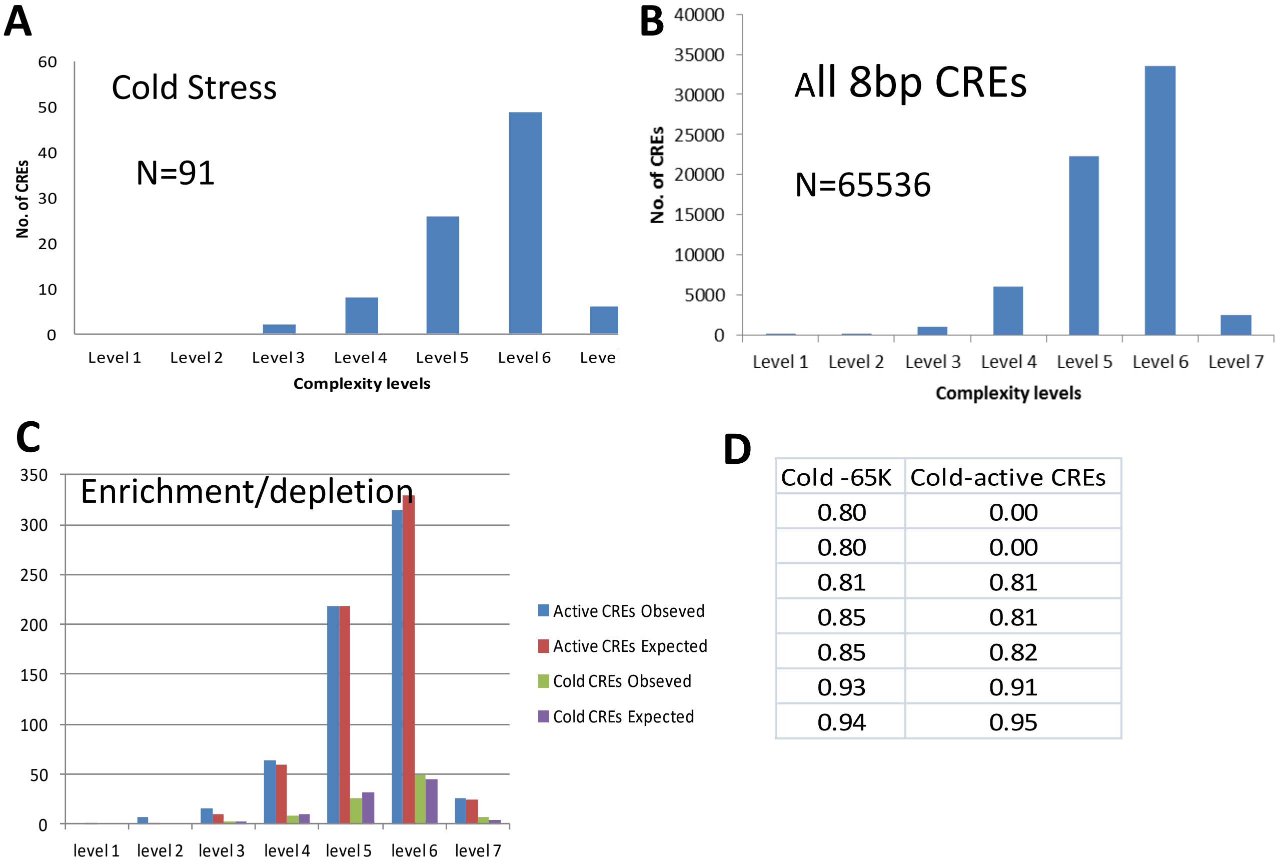


Figure 4.3. Time series analysis of cis-regulatory elements in that respond to drought stress and wounding. A) Correlation of the presence of 33 CREs with differential gene expression at different time points clustered into 8 clusters in plants grown in agar in a drought stress experiment where water was withheld for 24 hours. B) Similar analysis of 6 CRE clusters correlated with differential gene expression in a mechanical wounding experiment. Note that some CREs only correlate with differential expression at specific time points, some at early time points (e.g. Cluster1 and 2 in wounding), others at middle and later time points (e.g. Cluster 3 and cluster 7 in drought stress) See methods for experimental details.



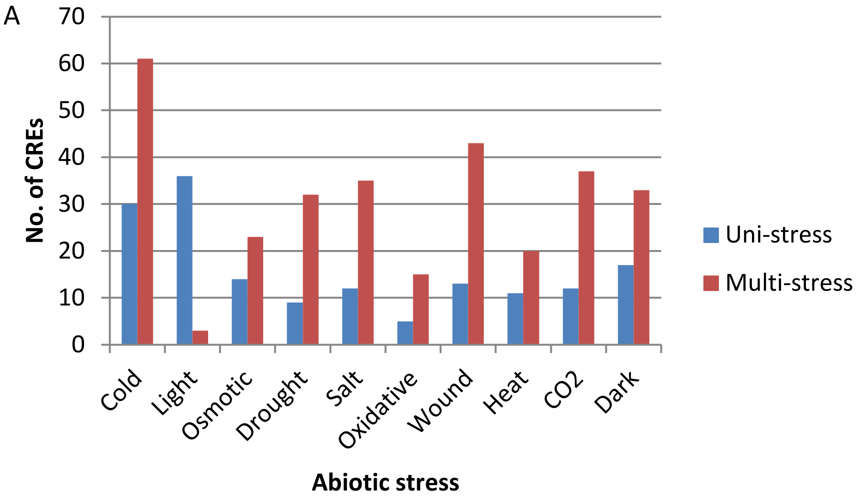
	Cold stress						Drought stress						
sequence	0.5hr	1hr	3hr	6hr	12hr	24hr	0.25hr	0.5hr	1hr	3hr	6hr	12hr	24hr
AAAACCCT	0.00	0.00	0.00	0.00	9.90	11.24	0.00	0.00	0.00	0.00	0.00	9.04	0.00
AAAAGTCA	0.00	0.00	0.00	0.00	9.51	12.55	0.00	0.00	0.00	10.46	0.00	0.00	0.00
AAAATATC	0.00	0.00	16.80	51.60	14.83	36.62	0.00	0.00	0.00	15.07	0.00	12.09	0.00
AAACCCTA	10.96	10.87	0.00	12.20	17.80	20.64	14.46	10.26	10.12	9.62	0.00	9.25	0.00
AAAGTCAA	12.66	11.90	11.41	11.85	14.02	19.84	13.50	0.00	9.21	15.92	0.00	0.00	0.00
AAATATCT	0.00	0.00	19.64	9.18	22.53	50.95	0.00	0.00	0.00	17.05	10.13	19.73	0.00
AACCACAC	0.00	10.07	0.00	0.00	0.00	0.00	0.00	9.37	0.00	9.39	0.00	0.00	9.02
AACCCTAA	9.06	0.00	0.00	9.94	17.23	14.52	12.51	9.81	10.35	0.00	0.00	0.00	0.00
AACCCTAG	10.09	10.33	0.00	9.59	11.40	13.74	10.16	10.64	0.00	0.00	0.00	0.00	0.00
AAGTCAAA	0.00	0.00	0.00	0.00	0.00	10.24	0.00	0.00	0.00	12.05	0.00	0.00	0.00
AAGTCAAC	11.39	12.83	12.15	11.71	14.74	19.88	13.51	9.80	15.61	14.24	10.36	0.00	0.00
ACACGTGG	0.00	0.00	0.00	0.00	0.00	14.29	0.00	13.05	10.64	9.68	0.00	0.00	0.00
ACACGTGT	0.00	0.00	0.00	0.00	0.00	18.32	0.00	13.76	14.77	0.00	0.00	0.00	0.00
ACCACACA	0.00	10.38	0.00	0.00	0.00	0.00	10.15	11.75	0.00	9.57	0.00	9.40	0.00
ACCTTATC	17.89	15.28	15.52	12.30	10.42	9.76	15.26	13.03	15.07	15.54	16.41	17.17	13.40
ACGTGGCA	9.11	9.50	0.00	11.41	14.05	14.39	0.00	10.26	11.18	0.00	10.11	0.00	0.00
ACGTGTCG	0.00	0.00	0.00	0.00	0.00	12.58	0.00	0.00	9.85	0.00	0.00	0.00	0.00
AGATAAGG	11.07	10.16	9.19	0.00	0.00	0.00	9.65	0.00	9.13	0.00	10.92	13.88	0.00
AGTCAACT	0.00	0.00	0.00	0.00	9.94	13.93	0.00	0.00	10.43	10.15	0.00	0.00	0.00
ATTATATA	11.73	0.00	10.47	0.00	0.00	0.00	0.00	9.62	9.78	0.00	9.88	0.00	0.00
CACGTGGC	9.25	10.90	0.00	0.00	11.95	17.00	0.00	14.77	16.56	12.20	10.46	9.60	0.00
CACGTGTC	0.00	0.00	0.00	0.00	0.00	17.79	0.00	13.02	14.17	0.00	0.00	0.00	0.00
CCACACAA	10.63	10.91	9.58	12.71	9.10	10.67	13.74	14.91	14.31	11.24	11.23	10.33	11.86
CCTTATCC	10.46	9.35	10.93	0.00	0.00	0.00	10.03	0.00	0.00	9.56	10.22	10.50	0.00
TATATAGA	0.00	0.00	0.00	0.00	0.00	10.56	0.00	10.45	0.00	0.00	0.00	0.00	0.00

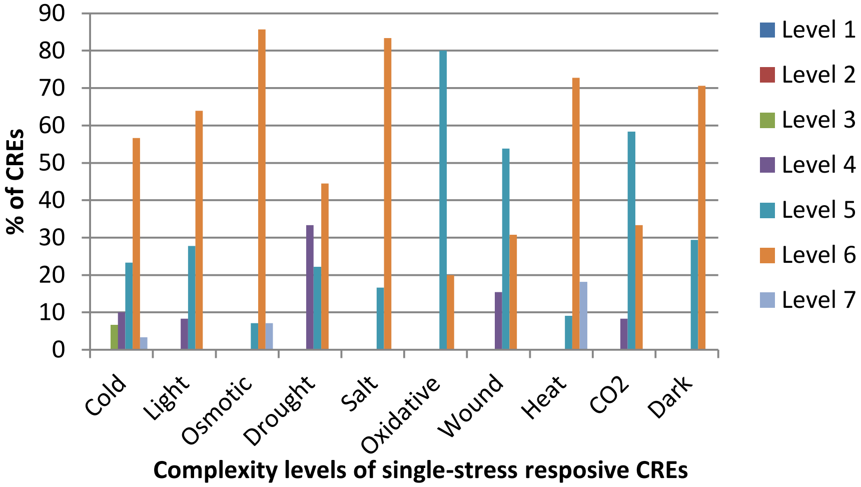
Figure 4.4 Time point dependent shared CREs between cold and drought stress. Heatmap shows strength of response (Red = CRE is strongly correlated to gene expression, Yellow= weak but significant correlation, Green= not significant). Numbers in colored boxes represents –log10 of P-value for significance of correlation (shown only for significant correlations). Out of 27 shared CREs between cold and drought stress, 25 CREs were also correlated by time point. Drought stress contained a plus time point compared to cold stress. Except of ACCACACA, all other CREs in 0.25 hr time point in CREs were also present in 0.5 hr time point in cold stress.



Cold -65K	Cold-active CREs
0.80	0.00
0.80	0.00
0.81	0.81
0.85	0.81
0.85	0.82
0.93	0.91
0.94	0.95

Figure 4.5. Distribution of complexity levels of cold stress CREs. Cold stress elements (A) showed a peak at complexity level six. There were no cold stress CREs from level one and two. This was compared to the distribution of complexity for all possible 8bp CREs (B) and 641active CREs (C). Cold CREs showed some enrichment for higher complexity (level 6 and 7) and depletion at lower complexity. D) Statistical significance was calculated by chi-squared analysis, comparing observed to expected values with the null hypothesis that cold stress CRE complexity was not different to randomly selected 8bp CREs.


A
70
60
50
40
30
20
10
0
No. of CREs
Cold
Light
Osmotic
Drought
Salt
Oxidative
Wound
Heat
CO2
Dark
Abiotic stress
Uni-stress
Multi-stress


90
80
70
60
50
40
30
20
10
0
% of CREs
Cold
Light
Osmotic
Drought
Salt
Oxidative
Wound
Heat
CO2
Dark
Complexity levels of single-stress resposive CREs
Level 1
Level 2
Level 3
Level 4
Level 5
Level 6
Level 7

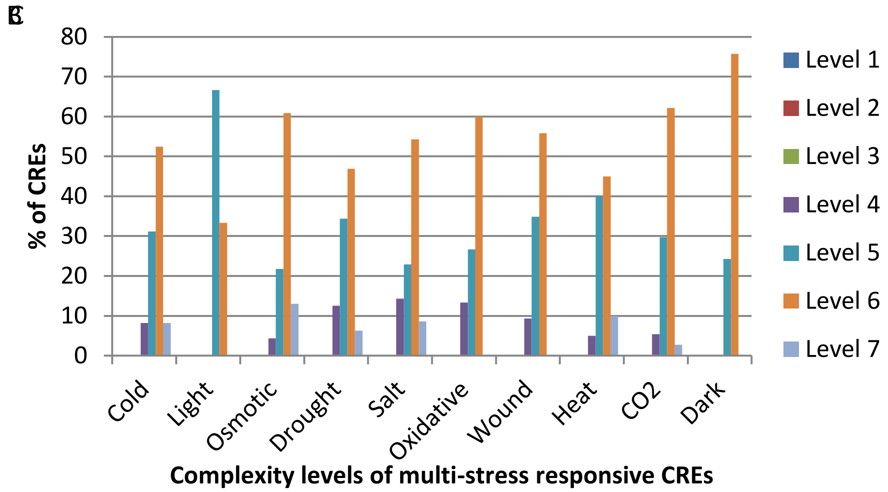


Figure 4.6. Number of CREs and percentage of complexity levels of uni-stress and multi-stress responsive CREs. A) The number of uni-stress and multi-responsive CREs in each abiotic stress. B) Percentage of CREs in complexity levels in uni-stress responsive CREs in abiotic stresses. C) Percentage of CREs in complexity levels in multi-stress responsive CREs in abiotic stresses.

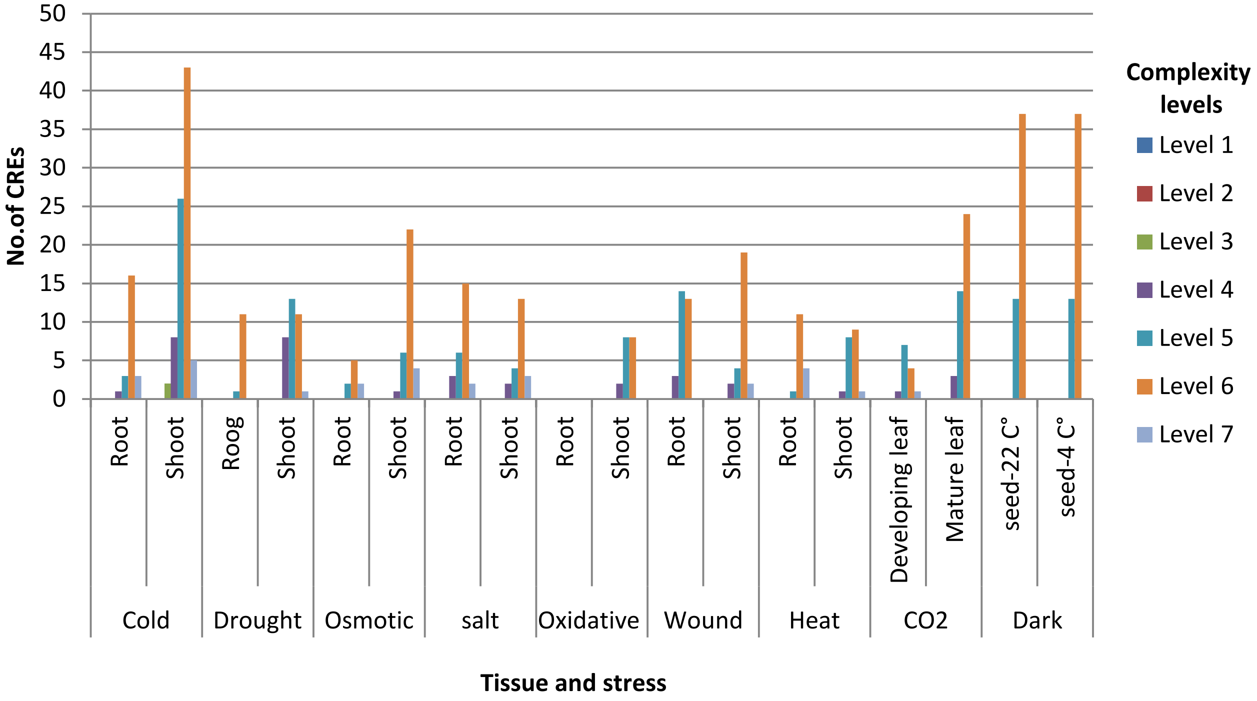
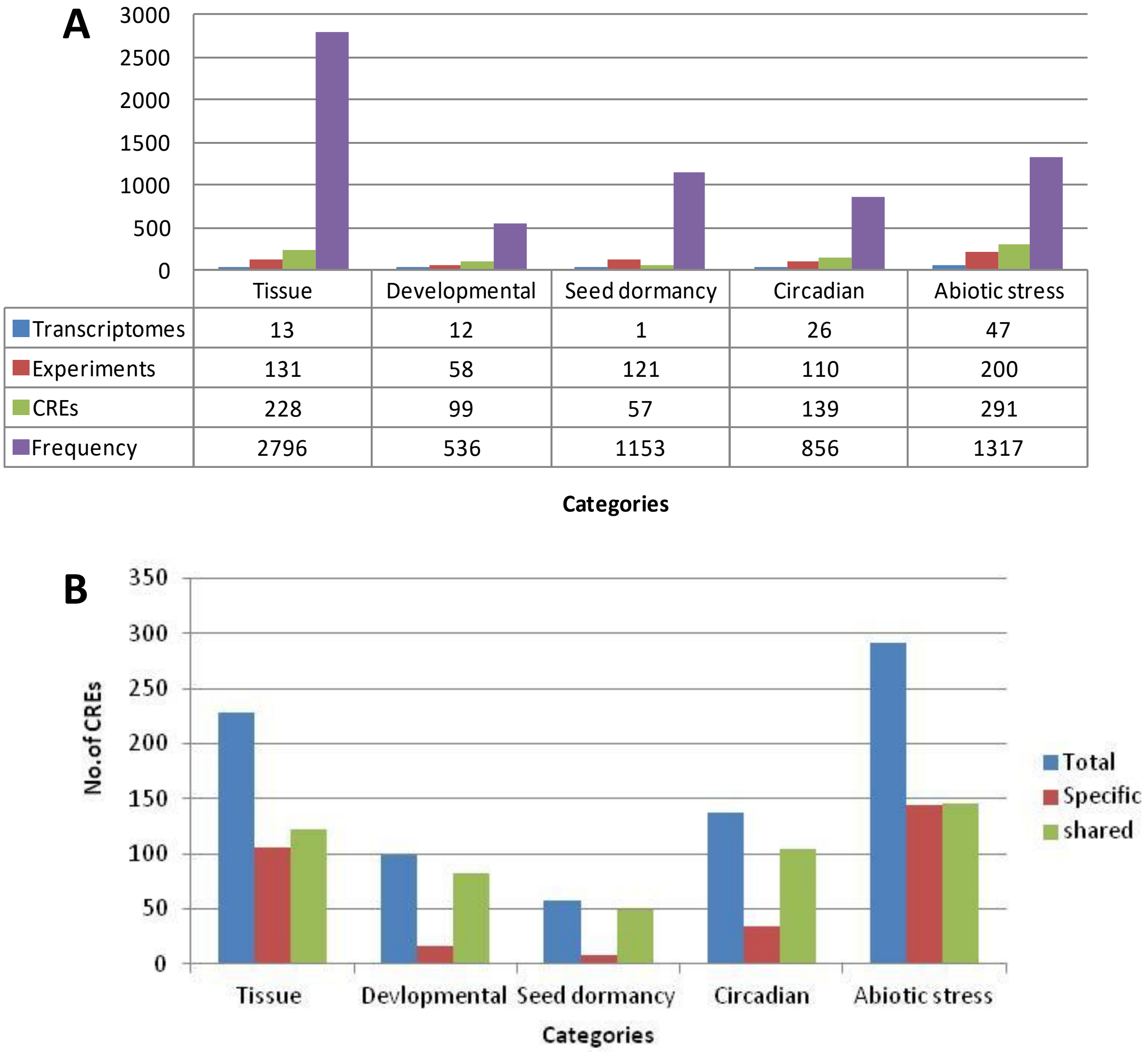


Figure 4.7. Complexity levels of CREs in shoot and root tissues in abiotic stress. The distribution of all abiotic stress active CREs in selected stresses and tissues is shown. Differences in distribution of CRE complexity is tissue specific, especially in most stresses, with an increase in complexity levels 4 and 5 in shoots of cold, drought and heat stresses, and decreased level 5 complexity in wounding (compare purple and green bars). There was also a strong decrease in complexity overall in CREs that regulated genes in developing leaves vs. mature leaves exposed to high CO_2 levels.

	Tissue	Abiotic stress	Developmental	Seed dormancy	Circadian
Tissue		1.03	1.85	2.02	1.59
Abiotic stress	1.03		1.74	1.78	1.48
Developmental	1.85	1.74		4.54	2.72
Seed dormancy	2.02	1.78	4.54		2.69
Circadian	1.59	1.48	2.72	2.69	

Figure 4.8. Number of CREs involved in developmental processes and their cross talk with each other and abiotic stress. Numbers indicate the CREs that are active in two different categories (indicated by row and column titles). The intersection of the same category represent category specific CREs. Green indicates large number; red indicates smaller number of either specific or shared CREs. Tissue indicates root, shoot, pollen tube root hair ground tissue and among the others. Developmental indicate experiments regarding rosette, cotyledon, and tissue development. Abiotic stress indicates the cold, drought, and other stresses. Seed dormancy include experiments regarding dormant seed under different temperature, light, moisture conditions. Circadian indicates studying the effect of day and night on gene expression of plants. Red and blue circles indicate largest and fewest shared CREs between two categories respectively.



	Tissue	Developmental	Seed dormancy	Circadian	Abiotic stress
🟦 Transcriptomes	13	12	1	26	47
🟥 Experiments	131	58	121	110	200
🟩 CREs	228	99	57	139	291
🟪 Frequency	2796	536	1153	856	1317

Figure 4.9. Number and distribution of CREs in different biological and abiotic stress categories. A) The number of experiments, transcriptomes, and CREs in each category. A single transcriptome may contain numerous experiments. Many CREs may significantly be correlated to a single experiment at the same time. Experiments indicate number experiments in all microarrays of a specific category. B) The number of total, specific and shared CREs in each category. The word specific refers to CREs belonging to certain category without sharing it with others. Abiotic stress has the highest number of CREs compared to all other categories.

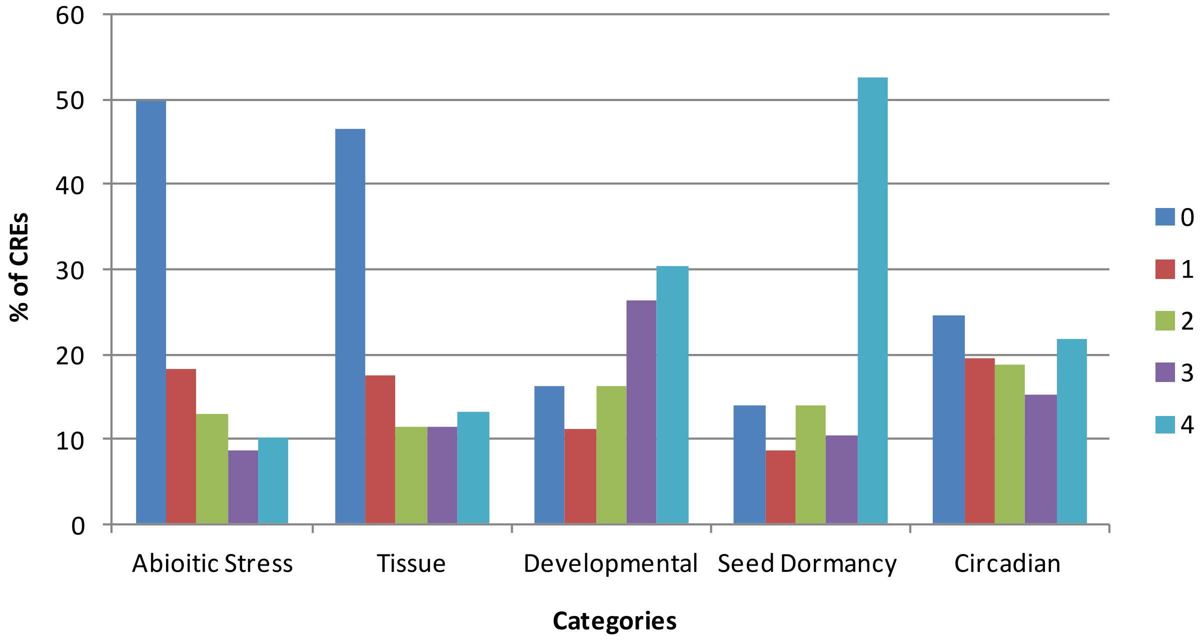


Figure 4.10 Percentage of CREs shared elements between various developmental categories and abiotic stress. Numbers represents shared CREs among different categories. Zero refers to percentage of CREs in certain category shared with none of the rest of four categories. Number 1 refers to the percentage of CREs in certain category shared with one of other categories, and so on. Number 4 refers to the percentage of CREs in certain categories shared with rest of other four categories.

CHAPTER 5

CONCLUSIONS AND RECOMENDATIONS

Transcriptional gene expression regulation is a key factor shaping biological status of an organism. Studying such gene expression regulation in a genome wide scale will provide scientists a comprehensive understanding regarding biological functions of tissue and organs in organisms. Nowadays availability of differential gene expression data is considered a potential to pursue this task. In the current study, differential gene expression data collected from databases and published articles are used in combination with a word enumeration tool to identify transcription factor binding sites in 500 bp promoter region of Arabidopsis and rice. We searched for 65,536 potential CREs of 8 bp length in the (-500bp upstream) promoter regions of coregulated genes in differential gene expression data. Of these, a list of only 641 and 856 CREs in *Arabidopsis thaliana* and rice (*Oryza sativa*) respectively were significantly correlated with gene induction and suppression. Among those CREs many of the previously common well known and characterized CREs like ABRE, EE and DRE were observed. In Arabidopsis, ABRE, EE showed high significant correlation with gene expression as well as broad response. However, DRE showed frequently less strength and narrow response breadth. In rice the aforementioned CREs showed lower significance correlation with gene expression and extremely narrow response breadth. Despite the specific length of CREs used in our study, longer elements were constructed due to sequence overlap. CREs up to 12 bp was found to exist in 500 bp regions of Arabidopsis genes.

The strength (correlation to differential expression) and breadth (number of different stimuli) of response were characterized for all CREs. EE and a novel element (AAACCCTA) showed highest strength in both Arabidopsis and rice, respectively. Regarding breadth of response, AAACCCTA, and GCGGCGGA showed broadest response through significantly correlation to 197 and 58 transcriptomes in both Arabidopsis and rice, respectively. In contrast, 291 and 279 CREs showed least breadth response through significantly correlating to only one transcriptome in Arabidopsis and rice respectively. CREs were also investigated for their complexity. In Arabidopsis, 314 out of 641 CREs were of level 6 complexity, which means having three repeats of the same nucleotide, and 25 CREs were of level 7 complexity, which is highest level of complexity through possessing 2 repeats of each of the four types of nucleotides. In rice the majority of CREs, 263 out of 856 were of level 5 complexity having four repeats of the same nucleotide, while only 22 CREs were of level 7 complexity.

Sequence and expression pattern similarity were also studied. Homonyms were referred to CREs with one letter mismatch (highest sequence similarity), while synonyms were referred to CREs having similar expression pattern regardless of their sequence similarities. We found that not all homonyms are synonyms, and some synonyms are also homonyms. Such discovery enables us understanding the evolutionary baseline of gene expression networks and degeneracy nature of CREs. It leads to a conclusion that degenerate CREs are not the same elements, but different elements with different functions.

Abiotic stress CREs were extensively investigated in Arabidopsis. Shared CREs among abiotic stresses and stress specific CREs were identified. Cold and wound stress showed highest number of shared elements, 31 CREs. Light stress showed least number of CREs shared with all other stress categories, 1 CRE, except for cold and drought stress, which shared 3 CREs and 0 CRE with dark stress.

REFERENCES

Amasino R (2004) Vernalization, competence, and the epigenetic memory of winter. The Plant cell 16 (10):2553-2559

Baker SS, Wilhelm KS, Thomashow MF (1994) The 5′-region of Arabidopsis thaliana cor15a has cis-acting elements that confer cold-, drought- and ABA-regulated gene expression. Plant molecular biology 24 (5):701-713.

Bernard V, Brunaud V, Lecharny A (2010) TC-motifs at the TATA-box expected position in plant genes: a novel class of motifs involved in the transcription regulation. BMC genomics 11:166-166.

Chen K, Rajewsky N (2007) The evolution of gene regulation by transcription factors and microRNAs. Nature Reviews Genetics 8:93.

Choi H-i, Hong J-h, Ha J-o, Kang J-y, Kim SY (2000) ABFs, a Family of ABA-responsive Element Binding Factors. Journal of Biological Chemistry 275 (3):1723-1730. doi:10.1074/jbc.275.3.1723

Chow C-N, Zheng H-Q, Wu N-Y, Chien C-H, Huang H-D, Lee T-Y, Chiang-Hsieh Y-F, Hou P-F, Yang T-Y, Chang W-C (2016) PlantPAN 2.0: an update of plant promoter analysis navigator for reconstructing transcriptional regulatory networks in plants. Nucleic acids research 44 (Database issue):D1154-D1160.

Cserhati M (2015) Motif content comparison between monocot and dicot species. Genomics Data 3 (0):128-136.

D. TJ, J. GT, G. HD (2003) Multiple Sequence Alignment Using ClustalW and ClustalX. Current Protocols in Bioinformatics 00 (1):2.3.1-2.3.22.

de Hoon MJL, Imoto S, Nolan J, Miyano S (2004) Open source clustering software. Bioinformatics 20 (9):1453-1454.

Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA (2003) DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biology 4 (9):R60-R60

Dey B, Thukral S, Krishnan S, Chakrobarty M, Gupta S, Manghani C, Rani V (2012) DNA–protein interactions: methods for detection and analysis. Molecular and Cellular Biochemistry 365 (1):279-299.

Dong MA, Farré EM, Thomashow MF (2011) Circadian clock-associated 1 and late elongated hypocotyl regulate expression of the C-repeat binding factor (CBF) pathway in Arabidopsis. Proceedings of the National Academy of Sciences 108 (17):7241-7246

Dubouzet JG, Sakuma Y, Ito Y, Kasuga M, Dubouzet EG, Miura S, Seki M, Shinozaki K, Yamaguchi-Shinozaki K (2003) OsDREB genes in rice, Oryza sativa L., encode transcription activators that function in drought-, high-salt- and cold-responsive gene expression. The Plant Journal 33 (4):751-763.

Earley K, Lawrence RJ, Pontes O, Reuther R, Enciso AJ, Silva M, Neves N, Gross M, Viegas W, Pikaard CS (2006) Erasure of histone acetylation by Arabidopsis HDA6 mediates large-scale gene silencing in nucleolar dominance. Genes & development 20 (10):1283-1293

Finkelstein R (2013) Abscisic Acid Synthesis and Response. The Arabidopsis Book / American Society of Plant Biologists 11:0166.

Fitzek E (2012) Reconstruction of melecular regulatory network in *Arabidopsis thaliana*.

Force A, Lynch M, Pickett FB, Amores A, Yan Y-l, Postlethwait J (1999) Preservation of Duplicate Genes by Complementary, Degenerative Mutations. Genetics 151 (4):1531

Foster R, Izawa T, Chua NH (1994) Plant bZIP proteins gather at ACGT elements. The FASEB Journal 8 (2):192-200.

Frey A, Godin B, Bonnet M, Sotta B, Marion-Poll A (2004) Maternal synthesis of abscisic acid controls seed development and yield in Nicotiana plumbaginifolia. Planta 218 (6):958-964.

Fujita Y, Fujita M, Satoh R, Maruyama K, Parvez MM, Seki M, Hiratsu K, Ohme-Takagi M, Shinozaki K, Yamaguchi-Shinozaki K (2005) AREB1 Is a Transcription Activator of Novel ABRE-Dependent ABA Signaling That Enhances Drought Stress Tolerance in &lt;em&gt;Arabidopsis&lt;/em&gt;. The Plant cell 17 (12):3470

Gao Z, Liu L, Ruan J (2017) Logo2PWM: a tool to convert sequence logo to position weight matrix. BMC genomics 18 (6):709

Garner MM, Revzin A (1981) A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system. Nucleic acids research 9 (13):3047-3060

Geisler M, Kleczkowski LA, Karpinski S (2006) A universal algorithm for genome-wide in silicio identification of biologically significant gene promoter putative cis-regulatory-elements; identification of new elements for reactive oxygen species and sucrose signaling in Arabidopsis. The Plant Journal 45 (3):384-398.

Grace ML, Chandrasekharan MB, Hall TC, Crowe AJ (2004) Sequence and Spacing of TATA Box Elements Are Critical for Accurate Initiation from the β-Phaseolin Promoter. Journal of Biological Chemistry 279 (9):8102-8110.

Grotewold E, Springer N (2009) The Plant Genome: Decoding the Transcriptional Hardwiring. Annual Plant Reviews 35:196-227

Gu Z, Steinmetz LM, Gu X, Scharfe C, Davis RW, Li W-H (2003) Role of duplicate genes in genetic robustness against null mutations. Nature 421 (6918):63

Guiltinan MJ, Marcotte WR, Quatrano RS (1990) A plant leucine zipper protein that recognizes an abscisic acid response element. Science 250 (4978):267-271

Harmer SL, Hogenesch JB, Straume M, Chang H-S, Han B, Zhu T, Wang X, Kreps JA, Kay SA (2000a) Orchestrated Transcription of Key Pathways in Arabidopsis by the Circadian Clock. Science 290 (5499):2110-2113.

Harmer SL, Hogenesch JB, Straume M, Chang H-S, Han B, Zhu T, Wang X, Kreps JA, Kay SA (2000b) Orchestrated transcription of key pathways in Arabidopsis by the circadian clock. Science 290 (5499):2110-2113

Hattori T, Totsuka M, Hobo T, Kagaya Y, Yamamoto-Toyoda A (2002) Experimentally Determined Sequence Requirement of ACGT-Containing Abscisic Acid Response Element. Plant and Cell Physiology 43 (1):136-140.

Hobo T, Asada M, Kowyama Y, Hattori T (1999) ACGT-containing abscisic acid response element (ABRE) and coupling element 3 (CE3) are functionally equivalent. The Plant Journal 19 (6):679-689.

Ito Y, Katsura K, Maruyama K, Taji T, Kobayashi M, Seki M, Shinozaki K, Yamaguchi-Shinozaki K (2006) Functional Analysis of Rice DREB1/CBF-type Transcription Factors Involved in Cold-responsive Gene Expression in Transgenic Rice. Plant and Cell Physiology 47 (1):141-153.

Izawa T, Foster R, Chua N-H (1993) Plant bZIP Protein DNA Binding Specificity. Journal of Molecular Biology 230 (4):1131-1144.

Izawa T, Shimamoto K (1996) Becoming a model plant: The importance of rice to plant science. Trends in Plant Science 1 (3):95-99.

Jiao Y, Lau OS, Deng XW (2007) Light-regulated transcriptional networks in higher plants. Nature Reviews Genetics 8 (3):217

Jiang, C., Iu, B. & Singh, J. Requirement of a CCGAC cis-acting element for cold induction of the BN115 gene from winter Brassica napus Plant Mol Biol (1996) 30: 679.

Joachim K, Dion W, Jakub H, Dierk W, Stefan W, Oliver B, Cecilia DA, Erich BB, Jörg K, Klaus H (2007) The AtGenExpress global stress expression data set: protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses. The Plant Journal 50 (2):347-363.

Jones-Rhoades MW, Bartel DP, Bartel B (2006) MicroRNAs AND THEIR REGULATORY ROLES IN PLANTS. Annual Review of Plant Biology 57 (1):19-53.

Kamp BP, Bøgh JA, Montserrat P (1997) Regulatory elements in vivo in the promoter of the abscisic acid responsive gene rab17 from maize. The Plant Journal 11 (6):1285-1295.

Karssen C, Brinkhorst-Van der Swan D, Breekland A, Koornneef M (1983) Induction of dormancy during seed development by endogenous abscisic acid: studies on abscisic acid deficient genotypes of Arabidopsis thaliana (L.) Heynh. Planta 157 (2):158-165

Khraiwesh B, Zhu J-K, Zhu J (2012) Role of miRNAs and siRNAs in biotic and abiotic stress responses of plants. Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms 1819 (2):137-148.

Kucera B, Cohn MA, Leubner-Metzger G (2005) Plant hormone interactions during seed dormancy release and germination. Seed Science Research 15 (4):281-307

Kyonoshin M, Yoh S, Mie K, Yusuke I, Motoaki S, Hideki G, Yukihisa S, Shigeo Y, Kazuo S, Kazuko YS (2004) Identification of cold-inducible downstream genes of the Arabidopsis DREB1A/CBF3 transcriptional factor using two microarray systems. The Plant Journal 38 (6):982-993.

Lai AG, Doherty CJ, Mueller-Roeber B, Kay SA, Schippers JH, Dijkwel PP (2012) CIRCADIAN CLOCK-ASSOCIATED 1 regulates ROS homeostasis and oxidative stress responses. Proceedings of the National Academy of Sciences 109 (42):17129-17134

Lawrence C, Altschul S, Boguski M, Liu J, Neuwald A, Wootton J (1993) Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science 262 (5131):208-214.

Lichtenberg J, Yilmaz A, Welch JD, Kurz K, Liang X, Drews F, Ecker K, Lee SS, Geisler M, Grotewold E (2009) The word landscape of the non-coding segments of the Arabidopsis thaliana genome. BMC genomics 10 (1):463

Liu H-H, Tian X, Li Y-J, Wu C-A, Zheng C-C (2008) Microarray-based analysis of stress-regulated microRNAs in Arabidopsis thaliana. Rna 14 (5):836-843.

Liu Q, Kasuga M, Sakuma Y, Abe H, Miura S, Yamaguchi-Shinozaki K, Shinozaki K (1998) Two Transcription Factors, DREB1 and DREB2, with an EREBP/AP2 DNA Binding Domain Separate Two Cellular Signal Transduction Pathways in Drought- and Low-Temperature-Responsive Gene Expression, Respectively, in Arabidopsis. The Plant Cell Online 10 (8):1391-1406.

Maruyama K, Takeda M, Kidokoro S, Yamada K, Sakuma Y, Urano K, Fujita M, Yoshiwara K, Matsukura S, Morishita Y, Sasaki R, Suzuki H, Saito K, Shibata D, Shinozaki K, Yamaguchi-Shinozaki K (2009) Metabolic Pathways Involved in Cold Acclimation

Identified by Integrated Analysis of Metabolites and Transcripts Regulated by DREB1A and DREB2A. Plant physiology 150 (4):1972.

Meinke DW, Cherry JM, Dean C, Rounsley SD, Koornneef M (1998) Arabidopsis thaliana: a model plant for genome analysis. Science 282 (5389):662-682.

Menkens AE, Schindler U, Cashmore AR (1995) The G-box: a ubiquitous regulatory DNA element in plants bound by the GBF family of bZIP proteins. Trends in Biochemical Sciences 20 (12):506-510.

Mikkelsen MD, Thomashow MF (2009) A role for circadian evening elements in cold-regulated gene expression in Arabidopsis. The Plant Journal 60 (2):328-339.

Mittler R (2006) Abiotic stress, the field environment and stress combination. Trends in Plant Science 11 (1):15-19.

Molina C, Grotewold E (2005) Genome wide analysis of Arabidopsis core promoters. BMC genomics 6.

Moreau Y, Smet FD, Thijs G, Marchal K, Moor BD (2002) Functional bioinformatics of microarray data: from expression to regulation. Proceedings of the IEEE 90 (11):1722-1743.

Nakashima K, Shinwari ZK, Sakuma Y, Seki M, Miura S, Shinozaki K, Yamaguchi-Shinozaki K (2000) Organization and expression of two Arabidopsis DREB2 genes encoding DRE-binding proteins involved in dehydration- and high-salinity-responsive gene expression. Plant molecular biology 42 (4):657-665.

Narusaka Y, Nakashima K, Shinwari ZK, Sakuma Y, Furihata T, Abe H, Narusaka M, Shinozaki K, Yamaguchi-Shinozaki K (2003) Interaction between two cis-acting elements, ABRE and DRE, in ABA-dependent expression of Arabidopsis rd29A gene in response to dehydration and high-salinity stresses. The Plant Journal 34 (2):137-148.

Nouri M-Z, Moumeni A, Komatsu S (2015) Abiotic Stresses: Insight into Gene Regulation and Protein Expression in Photosynthetic Pathways of Plants. International Journal of Molecular Sciences 16 (9):20392-20416.

Oeda K, Salinas J, Chua NH (1991) A tobacco bZip transcription activator (TAF-1) binds to a G-box-like motif conserved in plant genes. The EMBO Journal 10 (7):1793-1802.

Raz V, Bergervoet J, Koornneef M (2001) Sequential steps for developmental arrest in Arabidopsis seeds. Development 128 (2):243-252

Robin S, Robin S, Rodolphe F, Schbath S (2005) DNA, words and models: statistics of exceptional words. Cambridge University Press.

Rock CD, Quatrano RS (1995) The Role of Hormones During Seed Development. In: Davies PJ (ed) Plant Hormones: Physiology, Biochemistry and Molecular Biology. Springer Netherlands, Dordrecht, pp 671-697.

Rossi V, Locatelli S, Varotto S, Donn G, Pirona R, Henderson DA, Hartings H, Motto M (2007) Maize histone deacetylase hda101 is involved in plant development, gene transcription, and sequence-specific modulation of histone modification of genes and repeats. The Plant cell 19 (4):1145-1162

Sakuma Y, Liu Q, Dubouzet JG, Abe H, Shinozaki K, Yamaguchi-Shinozaki K (2002) DNA-Binding Specificity of the ERF/AP2 Domain of Arabidopsis DREBs, Transcription Factors Involved in Dehydration- and Cold-Inducible Gene Expression. Biochemical and Biophysical Research Communications 290 (3):998-1009.

Sandelin A, Alkema W, Engström P, Wasserman WW, Lenhard B (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic acids research 32 (Database issue):D91-D94.

Schaffer R, Ramsay N, Samach A, Corden S, Putterill J, Carré IA, Coupland G (1998) The late elongated hypocotyl mutation of Arabidopsis disrupts circadian rhythms and the photoperiodic control of flowering. Cell 93 (7):1219-1229

Shinshi H, Usami S, Ohme-Takagi M (1995) Identification of an ethylene-responsive region in the promoter of a tobacco class I chitinase gene. Plant molecular biology 27 (5):923-932

Singh KB (1998) Transcriptional Regulation in Plants: The Importance of Combinatorial Control. Plant physiology 118 (4):1111

Steward N, Ito M, Yamaguchi Y, Koizumi N, Sano H (2002) Periodic DNA methylation in maize nucleosomes and demethylation by environmental stress. Journal of Biological Chemistry 277 (40):37741-37746

Stockinger EJ, Gilmour SJ, Thomashow MF (1997) Arabidopsis thaliana CBF1 encodes an AP2 domain-containing transcriptional activator that binds to the C-repeat/DRE, a cis-acting DNA regulatory element that stimulates transcription in response to low temperature and water deficit. Proceedings of the National Academy of Sciences 94 (3):1035-1040

Strayer C, Oyama T, Schultz TF, Raman R, Somers DE, Más P, Panda S, Kreps JA, Kay SA (2000) Cloning of the Arabidopsis clock gene TOC1, an autoregulatory response regulator homolog. Science 289 (5480):768-771

Suzuki A, Wu C-Y, Washida H, Takaiwa F (1998) Rice MYB Protein OSMYB5 Specifically Binds to the AACA Motif Conserved among Promoters of Genes for Storage Protein Glutelin. Plant and Cell Physiology 39 (5):555-559.

Thijs G, Marchal K, Lescot M, Rombauts S, Moor BD, Rouzé P, Moreau Y (2002) A Gibbs Sampling Method to Detect Overrepresented Motifs in the Upstream Regions of Coexpressed Genes. Journal of Computational Biology 9 (2):447-464.

Tokunori H, Mihoko A, Yasuo K, Tsukaho H (1999) ACGT-containing abscisic acid response element (ABRE) and coupling element 3 (CE3) are functionally equivalent. The Plant Journal 19 (6):679-689.

Trevino V, Falciani F, Barrera-Saldaña HA (2007) DNA microarrays: a powerful genomic tool for biomedical and clinical research. Molecular Medicine 13 (9-10):527-541.

Uno Y, Furihata T, Abe H, Yoshida R, Shinozaki K, Yamaguchi-Shinozaki K (2000a) Arabidopsis basic leucine zipper transcription factors involved in an abscisic acid-dependent signal transduction pathway under drought and high-salinity conditions. Proceedings of the National Academy of Sciences of the United States of America 97 (21):11632-11637

Uno Y, Furihata T, Abe H, Yoshida R, Shinozaki K, Yamaguchi-Shinozaki K (2000b) Arabidopsis basic leucine zipper transcription factors involved in an abscisic acid-dependent signal transduction pathway under drought and high-salinity conditions. Proceedings of the National Academy of Sciences 97 (21):11632-11637

Wang Z-Y, Tobin EM (1998) Constitutive expression of the CIRCADIAN CLOCK ASSOCIATED 1 (CCA1) gene disrupts circadian rhythms and suppresses its own expression. Cell 93 (7):1207-1217

Waters, A. J., Makarevitch, I. , Noshay, J. , Burghardt, L. T., Hirsch, C. N., Hirsch, C. D. and Springer, N. M. (2017), Natural variation for gene expression responses to abiotic stress in maize. Plant J, 89: 706-717. Jiang C, Iu B, Singh J (1996) Requirement of a CCGAC cis-acting element for cold induction of the BN115 gene from winter Brassica napus. Plant molecular biology 30 (3):679-684.

Wu CY, Washida H, Onodera Y, Harada K, Takaiwa F (2000) Quantitative nature of the Prolamin-box, ACGT and AACA motifs in a rice glutelin gene promoter: minimal cis-element requirements for endosperm-specific gene expression. The Plant Journal 23 (3):415-421.

Wullschleger SD, Difazio SP (2003) Emerging Use of Gene Expression Microarrays in Plant Physiology. Comparative and Functional Genomics 4 (2):216-224.

Yadav SK (2010) Cold stress tolerance mechanisms in plants. A review. Agronomy for Sustainable Development 30 (3):515-527.

Yamaguchi-Shinozaki K, Shinozaki K (1993) Characterization of the expression of a desiccation-responsive rd29 gene of Arabidopsis thaliana and analysis of its promoter in transgenic plants. Molecular and General Genetics MGG 236 (2-3):331-340

Yamaguchi-Shinozaki K, Shinozaki K (1994) A novel cis-acting element in an Arabidopsis gene is involved in responsiveness to drought, low-temperature, or high-salt stress. The Plant cell 6 (2):251-264.

Yamaguchi-Shinozaki K, Shinozaki K (2005) Organization of cis-acting regulatory elements in osmotic- and cold-stress-responsive promoters. Trends in Plant Science 10 (2):88-94.

Yan T, Yoo D, Berardini TZ, Mueller LA, Weems DC, Weng S, Cherry JM, Rhee SY (2005) PatMatch: a program for finding patterns in peptide and nucleotide sequences. Nucleic acids research 33 (suppl_2):W262-W266.

Ye R, Zhou F, Lin Y (2012) Two novel positive cis-regulatory elements involved in green tissue-specific promoter activity in rice (Oryza sativa L ssp.). Plant Cell Rep 31 (7):1159-1172.

Yoshida R, Hobo T, Ichimura K, Mizoguchi T, Takahashi F, Aronso J, Ecker JR, Shinozaki K (2002) ABA-activated SnRK2 protein kinase is required for dehydration stress signaling in Arabidopsis. Plant and Cell Physiology 43 (12):1473-1483

Yoshida T, Mogami J, Yamaguchi-Shinozaki K (2014) ABA-dependent and ABA-independent signaling in response to osmotic stress in plants. Current Opinion in Plant Biology 21:133-139.

Zhou Q, Wong WH (2007) Coupling hidden Markov models for the discovery of Cis -regulatory modules in multiple species. Ann Appl Stat 1 (1):36-65.

VITA

Graduate School
Southern Illinois University Carbondale

Belan M. Khalil

belan.khalil@siu.edu

University of Duhok, Iraq
Master of Science, May 2010

University of Salahaddin, Iraq
Bachelor of Science, 2002

Dissertation Title:
ANALYSIS OF THE CIS-REGULATORY ELEMENT LEXICON IN UPSTREAM GENE PROMOTERS OF *ARABIDOPSIS THALIANA* AND *ORYZA SATIVA*

Major Professor: Dr. J.B. Matthew Geisler