Norwegian University of Life Sciences
Faculty of Veterinary Medicine and Biosciences
Department of Chemistry, Biotechnology and Food Science

Master Thesis 2014
60 credits

Comparative Analysis of Plant Regulatory Genomes

Jonas Christoffer Lindstrøm

Abstract

With the availability of a large number of plant genomes comes the task of gaining insights from them through comparative studies. One important feature in the evolution of all organisms is changes to the genomic cis-regulatory elements (CREs), usually found in non-coding genomic regions. With several genomes available, comparative studies of CREs can provide insights about the evolution of CREs.

In this project a method is developed that seeks to identify key CREs across different taxa and thus providing a platform for investigating cis-regulatory divergence. Genomic data from 25 different plant species is used. Annotations providing extensive coverage of CREs are rare, so a computational approach based on sequence motifs is used to predict CREs in regions upstream of the coding genes. Important CREs are identified as being conserved in significantly many gene families within a specified clade. The findings are compared with results from gene expression data and functional annotations from select model organisms. The method identified some conserved high prevalence CREs that were already known to have deep evolutionary roots. When divergence in important CREs were investigated between two important plant groups, no consistent pattern emerged. Although the gene expression and functional enrichment analyses provided interesting insights in themselves, they could not support the hypothesis that conserved high prevalence CREs is a good measure for motif importance.

Sammendrag

Endringer i ikke-kodende regulatoriske DNA-elementer er en viktig drivkraft i utviklingen til alle organismer. En utfordring ved den stadig økende tilgangen til plantegenomer er å utnytte disse til å gi oss kunnskap gjennom komparative analyser.

Her presenteres en metode for å identifisere viktige regulatoriske DNA-elementer på tvers av ulike taksonomiske grupper. På denne måten kan man forsøke å danne et grunnlag for å studere utviklingen av divergerte regulatoriske DNA-elementer. Det er tatt utgangspunkt i genomiske data fra 25 ulike planter. Høykvalitetsannoteringer av ikke-kodende regulaotriske elementer er sjeldne, så en EDB-tilnærming basert på seksvensmotiver er brukt for å predikere regulatoriske elementer oppstrøms av kodende gener. Viktige regulatoriske elementer er definert som å være konservert i signifikant mange genfamilier innen en klade. Funnene er så testet mot resultater fra genekspresjonsdata of funksjonelle annoteringer fra utvalgte modellorganismer. Metoden identifiserte noen konserverte høyprevalente regulatoriske elementer som allerede var kjent i hele planteriket. Det ble ikke funnet noe konsistent mønster i sammenligningen mellom to viktige grupper av planter. Geneskpresjonsdata og funksjonelle annoteringer kunne ikke støtte opp om at høyrepvalente konserverte regulatoriske elementer er er godt mål på viktigheten av disse.

Acknowledgements

The work done in this project would not be possible if not for the support from several people. First and foremost I would like to thank my supervisor, Torgeir R. Hvidsten, for all the guidance, helpful advice and computer resources he has provided. Also my co-supervisor Niklas Mähler deserves a mention in this respect.

I extend my gratitude to the Norwegian social democracy, represented by Lånekassen, for providing financial support during all my years as a student culminating with this thesis. The Norwegian University of Life Sciences, where I have been the last five years also deserves a mention for providing me with an invaluable education.

I am particularly grateful for the help given by Lars Grønvold for, among other things, double checking my results and spotting errors that otherwise would have gone unnoticed.

Last, but not least, I wish to thank my friends, family and partner for their continuous support during all this time.

Jonas Christoffer Lindstrøm

May 10th, 2014

Contents

Chapter 1

Introduction

1.1 Genes and information bearing molecules

All organisms inherit traits from its preceding generation. The most important mechanism that facilitates inheritance is the duplication of information encoded in molecules called deoxyribonucleic acids (DNA). DNA is a polymer of nucleotides in the shape of a double-helix. A nucleotide consists of a phosphate and a deoxyribose and one of four bases: Adenine, thymine, guanine and cytosine. The ordering of bases attached to the phosphate and deoxyribose backbone is the mechanism in which the inheritable information is encoded. The entirety of heritable information encoded in DNA in an organism is referred to as the *genome*.

Some regions of the genome, called *genes*, serve as basis for the structure of two other biopolymers: ribonucleic acid (RNA) and proteins, the latter trough a RNA intermediate. It is trough the expression of genes the organism's unique character, or phenotype, emerge.

The purpose of this chapter is to present necessary background knowledge that motivates the work this thesis builds upon. This first chapter has three main parts. The first part explains the biological concepts of genes, gene regulation and the evolution of gene regulation. The focus of the second part is on some of the bioinformatics approaches used to study one of the most important classes of gene regulation mechanisms presented in the first part. The chapter ends with a presentation of some of the similar studies that has been done by others, before the aim of this thesis is presented in more detail at the end.

Chapter 2 presents the specific data and methods used while the main results are presented in Chapter 3. In some cases the results presented are

only summaries, but more details will be found in the appendices. The findings are then discussed in more detail in Chapter 4, which also formulates an overall conclusion.

1.2 Gene regulation

Gene regulation plays an important role in the development and life cycle of all organisms. Correct and timely activation and repression of genes and their products ensures appropriate responses to environmental stimuli and internal signals, as well as ensuring that developmental processes proceed in the correct course.

1.2.1 DNA transcription

Among the numerous different regulatory mechanisms that has been discovered, the ones concerned with regulating the transcription of DNA to RNA are perhaps the most important (Carroll, 2008). DNA to RNA transcription is carried out by a group of proteins called RNA polymerases. In eukaryotes there are five different RNA polymerases that are referred to by roman numerals (Pol I, Pol II, Pol III etc.). RNA polymerase works by binding to DNA, breaking the hydrogen bonds between the two strands and using the DNA nucleotides as a template (by the Crick-Watson base pairing rules) for synthesizing a single stranded RNA molecule.

RNA polymerase can bind to any piece of DNA regardless of the base pair sequence, but in practice it mostly binds at certain areas called promoters (Watson et al., 2011) (pp. 383 and 590). Usually a set of proteins called transcription factors (TF) are involved in regulating RNA polymerase activity. Some transcription factors catalyze RNA polymerase binding and transcription, while others blocks, or inhibits, the binding of RNA polymerase. Transcription factors can thus increase or decrease the transcription rates.

1.2.2 Transcription factors

As transcription factors play a central role in regulating gene transcription, understanding their functional mechanisms is of importance.

Transcription factors typically have a DNA binding domain that have a high affinity to bind to certain short DNA sequences, typically 6 to 30 base pairs long. These DNA sequences are often referred to by different names, often used interchangeably, such as Transcription factor binding sites

(TFBS), *cis*-regulatory elements (CRE) or *cis*-regulatory sequences (CRS), although the latter two imply a broader class of regulatory sequences.

The binding affinity is not strictly dependent on just the nucleotide bases (i.e. the sequence), but other factors play a role as well. DNA can be packed at different densities and this is facilitated by different sets of proteins such as histones. Such DNA-protein complexes are called chromatin. Dense chromatin structures are generally unavailable for TF and polymerase binding while the opposite is true for less dense chromatin structures. Methylation of the adenine and cytosine nucleotides is another factor influencing binding affinity. The cellular concentration of specific TFs also influence the chance of a specific binding site being occupied. Furthermore, variations in the local and global three dimensional shape of the DNA and chromatin play a role as well. (Rohs et al., 2010). Variation in local structures, such as minor groove width, influence binding affinity since different parts of the nucleotide may be more readily available for protein interactions. Such local structures are also dependent on the local base pair sequence (Rohs et al., 2009).

In addition to the DNA binding domain, transcription factors can also have an activator region that interacts with other proteins such as RNA polymerase and other transcription factors. The activator region may act as a catalyst for DNA binding of other proteins. This may be done directly, or indirectly by binding to another molecule that in turn bind proteins to the DNA. Beside recruiting other proteins tha activator region can also induce conformational changes in other proteins. This is for example important in the elongation phase of transcription where the conformation of RNA polymerase is necessary to avoid premature transcription stop.

Promoters contain a number of different binding sites so that each gene is regulated by a different set of transcription factors. A group of different closely situated binding sites are called enhancers. The combination of the effects of specific TFs and their cooperative binding facilitates a complex regulatory logic. In the simplest case the binding sites regulating a specific gene may operate in a simple additive manner, whereby the transcription level effectively is the sum of the specific effects for each bound TF. The TFs may however function in a fashion resembling logical operations. A specific TF may for example be a necessary, but not sufficient, to initiate transcription. In this way a gene can be regulated so that it is only active when several conditions are met. Conversely, a specific TF may be able to block transcription regardless of what other TFs are present in the cell.

Enhancers in eucaryotes can be many tens of thousands base pairs away from the transcription start site (TSS), but many are also in closer proximity. The regulatory elements far away from the TSS relies on the tertiary DNA structure (spatial bending and global conformations) to function. Ex-

ploratory *in silico* studies of these are harder for a number of reasons. Querying the whole genome for binding sites will yield many false positives since short sequences (which may also be degenerate) will often exist by chance. This could be remedied by data on global conformations, but this have generally not been available.

1.2.3 Trans-acting coregulation

A transcription factor can play a role in regulating several genes with different functions at different locations in the genome. Some of these functions may be wildly different and bear little relationship with each other, such as aiding the development of different tissues and organs. Its roles can also be more functionally related or similar, such as in the case of a TF regulating several genes during the development of a specific organ, but have different effects at different developmental stages.

An example of a protein that assume several functions is the protein LEAFY (LFY) in Arabidopsis. LFY is known to have many functions (Wang et al., 2004). For example is one function to promote flower development. In the vegetative phase of Arabidopsis development a group of undifferentiated stem cells called the shoot apical meristem (SAM) produces leaves and branches. The differentiation of SAM into flower tissues is dependent on LFY activation of the gene APETALA1 (Wagner et al., 1999). Another function is to control the length and angle of the pedicel (the part of the stem where the flower is). LFY activity at the base of the pedicel activates the gene AS2, which in turn plays a role in promotion of apoptosis in tissues on the upper side of the stem (Yamaguchi et al., 2012).

1.2.4 *Cis*-acting coregulation

A set of genes can be co-regulated by sharing the same promoter if these are clustered in close proximity. These kinds of co-regulated clusters goes by the name of operons, and has long been known to exist in bacteria. Operons typically consists of functionally related genes. Operons, although much rarer in eucaryotes than in bacteria, has been found in maize (*Zea mays*), *Arabidopsis thaliana* and oat (*Avena* spp.) (Chu et al., 2011).

1.3 Regulatory evolution

1.3.1 The role of gene duplication

A gene duplication is a type of mutation that yields a duplicate copy of a sequence containing one or several genes. Altough one of the copies of most duplicated genes tend to accumulate mutations rendering it nonfunctional (Lynch and Conery, 2000), gene duplication events are considered to be an important source for novel functionality (neofunctionalization). Gene duplications, including both whole genome duplications and single genes, are more common in plants than in animals (Adams and Wendel, 2005, Flagel and Wendel, 2009).

Subfunctionalization is a mode of neofunctionalization that occurs when both copies of the duplicated gene contributes to the roles the ancestral gene once had. There are two evolutionary models of how subfunctionalization occurs. Subfunctionalization may occur trough neutral evolution following the Duplication-Degeneration-Complementation (DDC) model (Force et al., 1999). Under this scenario complementary mutations can occur in both gene duplications in such a way that both contributes a part of the ancestral functions. The alternative model posits a scenario where the subfunctionalization is the result of positive natural selection. One such scenario is when a gene has several functions, but some of them requires mutually exclusive variants to be optimal. This model is called 'escape from adaptive conflict' (EAC) (Hughes, 1994).

Novel functionalities related to gene duplication events can be the result of changes to regulatory networks, possibly by mutations in *cis*-regulatory elements. Another possibility is that during the gene duplication event, parts of the *cis*-regulatory elements associated with the duplicated gene are not part of the duplicated sequence (see (Flagel and Wendel, 2009) for examples). Regulatory novelties can also be the result of duplications of regulatory genes, such as transcription factors. This may be an important factor in regulatory evolution in plants, as duplicated TFs in Arabidopsis and rice has been shown to have lower rates of nonfunctionalization than other duplicated genes (Shiu et al., 2005). Expression studies has also shown evidence consistent with continued functionality subsequent of gene duplication (Duarte et al., 2006).

1.3.2 Enhancer structure and constraints

The fact that the same transcription factor can regulate different genes in different tissues at different times imposes strict constraints on the transcription factor structure, since even small changes in protein sequence can yield

widespread morphological consequences. A lot of the evolutionary variation in morphology between related species does not lie in coding sequence variation. Instead the most important variation lies in the structure of enhancers and promoters (Carroll, 2008). The enhancer structure can be described by several organizational features:

- Numbering - The presence, absence and the number of copies of individual CREs.
- Ordering - The ordering of the CREs describes their arrangement along the DNA molecule.
- Orientation - The direction of a CRE sequence relative to the other CREs, *i.e.* which strand it is present on.
- Spacing - The distance between the CREs placed on the DNA molecule.

These features may or may not play an important role, depending on the role of the specific enhancer. Many transcription factors act cooperatively, and will not work properly if the behavior of other transcription factors change. If most, or all, of the TFs which bind to the enhancer act cooperatively then the enhancer's organizational requirements will be less intolerant to mutations. The model for when this is the case is referred to as the enhanceosome model (Yáñez-Cuna et al., 2013).

The opposite case of the enhanceosome is called the billboard model, where the TFs will largely work independent of each other. The structural requirements of a billboard enhancer will therefore be slack. The TFs at work can even be substituted by other functionally similar TFs (Yáñez-Cuna et al., 2013).

The two models, the Enhanceosome model and the Billboard model, describe two extreme cases. Most enhancers will have structural requirements somewhere in between (Meireles-Filho and Stark, 2009). The strict requirements of the enhanceosomes will lead us to expect the enhancer to show collinear sequence conservation among orthologs, while billboard enhancers can show great sequence divergence while still being able to retain its functionality.

1.3.3 TFBS turnover

Enhancers, and specifically CREs, are subject to mutations in the same way the rest of the genome is. As the functional role and genomic characteristics of TFBSs differ from other functional elements in the genome (such as protein

coding genes) their evolution tend to be subject to different processes and constraints.

Turnover is the gain and loss of CREs. The principal source of novel function in coding sequences are changes to already functional sequences. New coding sequences are often the result of duplication of an already existing coding sequence. Compared to coding sequences, CREs are very short. This makes new binding sites more likely to appear as a result of point mutations in non-functional sequences than coding sequences. Conversely, binding sites are also more more prone to loss-of-function mutations than coding sequences.

1.4 Motifs and descriptions of regulatory elements

As described earlier, the sites where transcription factors tend to bind often show variations in base pair sequence and three dimensional shape. Descriptions of this variability are referred to as *motifs*. Motifs in general can be used to describe variability in any string of characters, not just biological sequences of nucleotides or amino acids. Genomic binding sites are a suitable target for motif descriptions.

A motif, in addition to be used as a binding site description, should also have the power to predict new binding site instances. Several methods for describing motifs has been developed, most of which has focused on nucleotide sequence while ignoring three dimensional shape features. There are two main classes of methods for describing sequence motifs: Probabilistic and non-probabilistic (Stormo, 2000).

1.4.1 Probabilistic motifs

The most common probabilistic representation of a motif are the position frequency matrices (PFMs). A PFM for a binding site of length k nucleotides is represented by a $4 \times k$ matrix where each column represents one position and each row contains the frequencies for one of the four nucleotides such that each column sums to one. Each column is in effect a probability distribution over the possible nucleotides given the position corresponding to the column.

Assuming that the nucleotides in a genomic sequence are independently distributed given its neighboring nucleotides, then the probability of a sequence s of length k given a position frequency matrix M is

$$P(s|M) = \prod_{i=1}^{k} \sum_{j=1}^{4} I(M_{ij}, s_i)$$

where M_{ij} is the probability of nucleotide j at position i, s_i is the observed nucleotide at position i and

$$I(M_{ij}, s_i) = \begin{cases} M_{ij} & : j = s_i \\ 0 & : j \neq s_i \end{cases}$$

PFMs can be used to predict new instances of binding sites by converting it into a position specific scoring matrix (PSSM, also called a positional weight matrix, PWMs). The PSSM is derived from the probabilistic nature of the PFM using log-likelihood ratios: The score S_{ij} for each nucleotide j at a given position i is

$$S_{ij} = log(\frac{M_{ij}}{f_j})$$

where f_j is the background frequency for nucleotide j. This is can be taken to be uniform (*i.e.* $f_j = 0.25$ for $j = A, C, G, T$) or be informed by the frequencies in the relevant genome (*i.e.* CG content).

1.4.2 Non-probabalistic motifs

A non-probabilistic approach is to use consensus sequences. A consensus sequence is a string of letters, each of which represents the consensus nucleotide at the given position. There are several methods to determine what, given the data used to construct the sequence, is the consensus at a given position (Day and McMorris, 1992). The letter at each position can also represent degenerate nucleotides allowing two, three or four possible bases at a position. Table 1.1 lists all symbols used to describe a DNA sequence.

Non-probabilistic motifs can also take the form of regular expressions. A regular expression is a description of a text pattern using a standard notation that is supported by many software systems. Regular expressions offer more flexibility than consensus sequences. For example can a regular expression pattern describe a group of sequences with a repetitive subsequence of varying length, which is not possible using consensus sequences.

For example, the MYB1 binding site motif from AGRIS can be described as consensus sequence `MTCCWACC`. The same motif in regular expression form can be written as `[ACM]TCC[ATW]ACC`. An equivalent, yet more complex version of the same regular expression is `[ACM]TC{2}[ATW]AC{2}`.

Symbol	Description	Mnemonic	RegExp matching set
A	Adenosine		
C	Cytosine		
G	Guanine		
T	Thymine		
B	C, T or G	Not A	CGTBSKY
D	A, T or G	Not C	AGTDRWK
H	A, T or C	Not G	ACTHMYW
K	C or T		GTK
M	A or C		ACM
N	A, C, T or G	All Nucleotides	ACGTBDHKMNRSVWY
R	A or G		AGR
S	C or G	Strong binding	CGS
V	A, C or G	Not T (or U)	ACGVMSR
W	A or T	Weak binding	ATW
Y	C or T	Pyrimidine	CTY

Table 1.1: Symbols for nucleotide represenation, including degenerate nucleotides and the corresponding set of symbols that must be encapsulated in the regular expression square bracket syntax.

1.4.3 Motifs for local DNA structure

The above methodologies for binding site characterization has focused solely on primary DNA structure (*i.e.* sequence). Methods for binding site prediction taking local three dimensional DNA structure into account has recently been developed. On such method is TFBSshape (Yang et al., 2013). This is a method based on predictions from complex molecular simulations of DNA structure. Based on these simulations local structures such as minor groove width, twist and others can be predicted from sequence data. Another method is SiteSleuth, which combines local structures as well as characterizations of DNA-amino acid electrostatic interactions as features in a Support Vector Machine (SVM) classifier (Maienschein-Cline et al., 2012).

1.4.4 *De novo* motif discovery and prediction

Identification of CREs and describing them as motifs relies on a combination of experimental techniques and computer algorithms. The usual approach is to identify a set of sequences believed to share a common CRE and then investigating the sequences using software tools. Motif discovery is a widely studied problem and given a set of sequences, there exists numerous compu-

tational approaches to identify putative regulatory sites.

There are several experimental approaches to identify a set of sequences to investigate. A set of co-expressed genes can for example be suspected to be regulated by some of the same TFs, making the associated promoter sequences are a natural candidate for further investigation. Comparative genomic studies based on gene orthology can also yield promoter sequences worthy of investigating. CHip-seq experiments is another approach, where genomic sequences where TFs are known to bind can be identified.

Given a set of sequences, there exists numerous computational approaches to identify putative regulatory sites (Das and Dai, 2007). These can roughly be classified as either probabilistic or word based akin to the two types of motifs described above, or classified based on whether they account for phylogenetic relationships between the sequences or not. Methods incorporating phylogenetic relationships are developed for cases where the sequences are orthologous, as these kind of sequences tend to be more similar across larger stretches than just at the binding sites than for sequences where binding sites emerged *de novo* (Zambelli et al., 2013).

One popular probabilistic algorithm is MEME (Bailey and Elkan, 1994). All overlapping subsequences of length W from the sequences of interests are modeled as coming from a mixture of two sources: A binding site of length W with position specific nucleotide frequencies (a PFM) and the background (*i.e.* not a binding site) with uniform nucleotide frequencies. The parameters of the mixture model are then estimated by the Expectation-Maximization (EM) algorithm: The likelihood for each subsequence is calculated for both models (the expectation step). In the subsequent maximization step the PFM is recomputed with the subsequences weighted according to their likelihood in the E-step. These two steps are repeated until convergence.

Other algorithms tries to build a PFM using a stochastic approach called Gibbs sampling (Das and Dai, 2007). The idea is to select one W length subsequence at random from all the input sequences but one. Then the likelihood for all subsequences in the sequence not used to build the PFM is calculated. Then one of those subsequences are chosen at random, but proportional to the likelihood, to be used in the subsequent building of the PFM. The subsequence from the next sequence is then removed and the procedure repeats until convergence.

The MERCED (modeling evolution rate across species for cis-regulatory element discovery) algorithm (Ding et al., 2012b) for orthologous sequences models evolutionary distance between two species to identify conserved *cis*-regulatory k-mers. The k-mers are then clustered and a PWM is constructed from each cluster.

1.4.5 Binding site motif collections

Several databases exists that collects motifs of binding sites and other *cis*-regulatory elements. These collections tend to be part of a greater framework for analyzing gene regulation and may contain information about transcription factors, promoters *etc.* Some of them specializes in binding sites identified in plants and some are even dedicated to a single species.

The PLACE database (Higo et al., 1999) collects consensus sequence motifs from the literature. The maintenance of the database was discontinued in 2007, but it is still available. A small update happened in 2013 with a small fix in the annotation for some of the motifs. The most recent version (30.0.1) contains 469 entries.

JASPAR CORE is a curated database of non-redundant PFMs (Mathelier et al., 2013) constructed based on published experimental results. Version 5 of JASPAR contain 64 PFMs in the JASPAR CORE Plantae subdatabase.

TRANSFAC Professional (Matys et al., 2003) is a database available trough subscription. It contains annotations collected from the literature on TFBSs from a large number of eukaryotes as well as over 2900 PFMs. A public, free-of-charge version of TRANSFAC is also available, but has not been updated since 2005.

The Arabidopsis Gene Regulatory Information Server (AGRIS) (Yilmaz et al., 2011) contains information on transcription factors, promoter sequences from Arabidopsis annotated with *cis*-regulatory elements and more. It also contains a list of 99 binding site motifs as consensus sequences.

The AthaMap project has done a genome-wide survey of Arabidopsis and identified over ten million putative binding sites from 61 different TFs (Steffens et al., 2014). The binding sites comes primarily from the TRANSFAC database, but the project has expanded the database with binding site descriptions gathered from the literature (Steffens et al., 2004). The motifs are either PFM or consensus sequences. It contains 29 consensus sequences and 14 PFMs that are not from TRANSFAC.

1.5 Related research

A lot of research has been done on *cis*-regulation in general, and some of it is concerned with plant *cis*-regulation in particular. One bioinformatics approach is to use motifs from available databases or other publications to predict binding sites in a DNA sequence. This kind of research can differ in scope and focus, where some focuses on a single family of genes, while others takes a genome wide approach.

Xu et al. (2012) investigated the promoter regions of genes in Arabidopsis and Sorghum that were orthologous to a group of 10 rice TFs that had been shown to be coexpressed as a response to different kinds of stress. They found conserved clusters of CREs consisting of two to four CREs in 8 of 10 orthologous groups. They where conserved in terms of orientation, spacing and ordering. They further identified sets of binding sites that showed divergent conservation between the the ortholog groups as well as lineage specific conservation patterns within the ortholog groups.

A similar study investigated the the promoter region of genes in a sucrose transporter family in Arabidopsis and rice (Ibraheem et al., 2010). Based on a collection of motifs from three databases they found predicted a large number CREs. They also found that the predicted CREs formed gene specific clusters at different locations in the promoter region.

Another study (Ding et al., 2012a) investigated the prevalence of combinations of cis-regulatory elements in Arabidopsis and poplar. By using motifs from the PLACE database they identified over 18 thousand combinations of two to six CREs in promoter regions present in both species. One third of the identical combinations in the two species were associated with genes that were orthologous according to available annotations.

An important computational study using gene expression data and a database of known *cis*-regulatory motifs in yeast (*Saccharomyces cerevisiae*) identified regulatory elements working in pairs (Pilpel et al., 2001).

The PlantPAN project mapped CREs, tandem repeats and CpG/CpNpG islands in promoter sequences from Arabidopsis, maize and rice (Chang et al., 2008). The CRE predictions was based on consensus sequences from TRANSFAC, PLACE, AGRIS and JASPAR. These predictions are available trough a web-tool integrating Gene Ontology annotations as well as a tool for identifying enriched occurrences of combinations of CREs in a set of promoters.

Another computational approach was used by Christ et al. (2013). By utilizing the abundance of already performed gene expression experiments in Arabidopsis they identified a set of genes that were involved in root development. In the promoter regions of these genes they found three motifs that also showed significant co-occurrence.

A different approach worth mentioning, despite not involving plants, is the Broad Institute's ongoing large scale comparative project of 29 mammal genomes (Lindblad-Toh et al., 2011). The initial results focusing on identifying constrained genomic regions in humans also included and analysis of promoter regions. They found 2.7 million conserved instances of 688 putative CREs.

1.6 Aim

The principal aim of this project is to investigate the prevalence of *cis*-regulatory elements in promoter regions across species and gene families. A key to understanding the evolution and divergence of CRE prevalence lies in the conservation patterns of the CREs within families. One aim here is therefore to identify important CREs, *i.e.* regulatory elements that seems to play a role in a large number of gene families, and look for conservation and divergence patterns of such CREs.

Motifs from different databases are gathered and used to predict instances of CREs in the plant promoter regions. Information about gene families are subsequently used to infer phylogenetic conservation. Monte Carlo techniques are then used to identify CREs conserved in significantly many families.

To assess whether this approach gives biological meaningful answers the results needs to be compared to other kinds of data. Results from gene expression studies and functional gene annotations provides well established and suitable platforms to identify biological similarities in sets of genes. Data from select model organisms are used.

While similar research has focused on specific gene families or genes with specific functions from a limited number of species this project instead aims to investigate a large number of genes from a greater number of species. This makes it possible to look at conservation and divergence at different levels.

Chapter 2

Methods and materials

2.1 Genomic data and gene families

Version 2.5 of PLAZA (Bel et al., 2012) were used as the source for the promoter sequences and gene family information. Version 2.5 contains genomic information on 25 plant species. See figure 2.1 and Table 3.1.

The 2kb region upstream of the start codon, as described in the annotations from PLAZA, were extracted from the genomic sequences. Throughout, these will be referred to as *promoter sequences*, or just *promoters*. The choice of using the translation start codon instead of the transcription start site as reference point was mostly due to annotational convenience (annotations from PLAZA did not contain information on transcriptional start sites), but it has the benefit that it will include the 3' untranslated region. It also has precedence in the literature (see for example (Ibraheem et al., 2010)).

The PLAZA gene families are the result of a clustering of the protein sequences of the coding genes using the TRIBE-MCL algorithm (Enright et al., 2002). The algorithm creates a similarity matrix based on pairwise alignments of all sequences using BLAST (Altschul et al., 1990). The similarities are then scaled so that each column sums to 1, and all entries are between 0 and 1, giving them a probabilistic interpretation. The resulting matrix is then interpreted as a transition matrix for a Markov Chain. The subsequent clustering then repeats a three-step procedure: (1) Simulate a two step random walk through the chain. This is the same as squaring the transition matrix using regular matrix multiplication. (2) Exponentiate each entry in the matrix by some number a. This inflates the entries as low probabilities becomes proportionally much lower than greater probabilities. (3) Each column is then scaled to be probabilities again. These steps are repeated until the procedure produces no or little change in the matrix. The

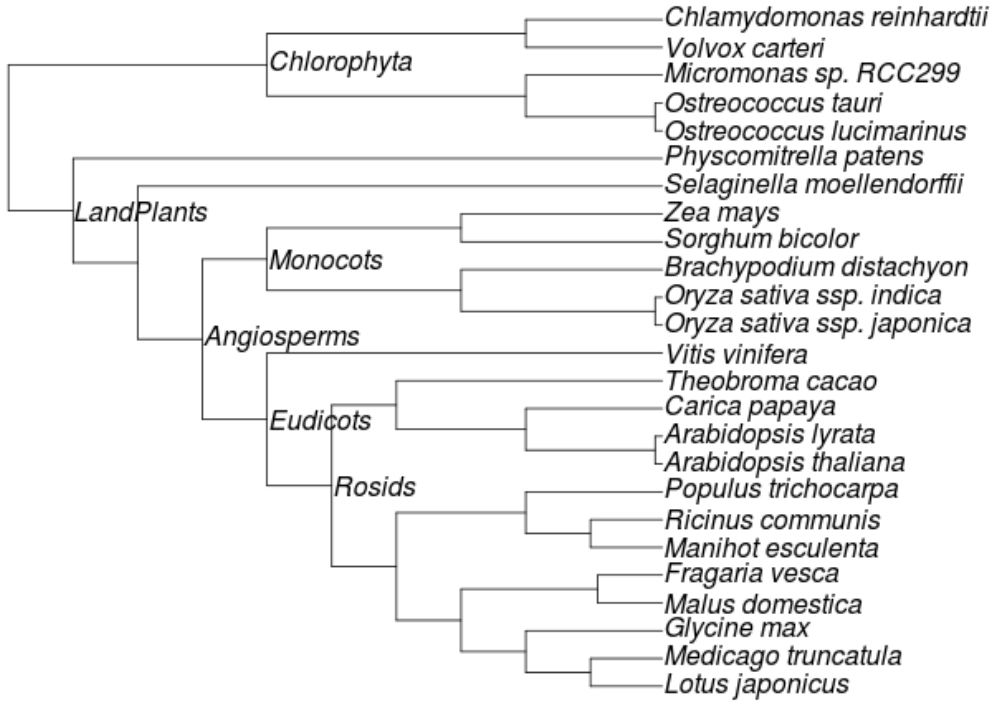


Figure 2.1: The species from PLAZA and their phylogenetic relationships. Adapted from data available from the PLAZA web site.

clusters are set to the resulting disconnected graphs.

2.2 Key species and taxa

The 25 species available on the PLAZA platform covers a wide range of clades. The following clades are covered in this investigation (also take a look at Appendix A and Figure 2.1): Eudicots, Monocots, Chlorophyta, Rosids, Fabids, Malvids, Angiosperm, Green plants and Oryza.

The clade Green Plants includes all 25 species in this study. The two main subclades represented here are Chlorophyta and land plants. The chlorophyte species are green algae that are mostly aquatic single celled or colonial species. Among the land plants are one moss (*Physcomitrella patens*) and one lycophyte (*Selaginella moellendorffii*). The main group of land plants are the Angiosperms, also called flowering plants, represented by 18 species. The two most important angiosperm divisions, monocots and eudicots, are represented by five and 13 species, respectively. The eudicots, also called true dicots, include most trees and shrubs and herbs, while the monocots

includes grasses and palms (Raven et al., 2005).

The monocots genus Oryza consists of two rice (*Oryza sativa*) subspecies *japonica* and *indica*.

Among the 13 eudicot species 12 belongs to the rosids. The Rosids are further divided into Fabids (8 species) and Malvids (4 species).

2.3 Motif data and consolidation

Motifs were gathered from three different databases: JASPAR CORE plantae, PLACE (v.30.0.1) and the binding sites list from ActisDB available from The Arabidopsis Gene Regulatory Information Server (AGRIS). See section 1.4.5 for more details and references. All motifs were consolidated into a single database where each motif were represented as consensus sequences.

Motifs from JASPAR were only available as count matrices and were converted into consensus sequences using the method described in (Cavener, 1987) as implemented in Biopython: A single nucleotide is considered to be the consensus if it occurs in more than half of the instances at that position, and is more than twice as common as the second most frequent nucleotide. Degenerate nucleotides describing two possible nucleotides are considered the consensus at positions where two nucleotides make up more than three quarters of the nucleotides in that position. If none of these two conditions are met, the position has no consensus nucleotide and is described by the symbol 'N'. The consensus sequences where then made into regular expressions.

The motifs from AGRIS contained some non-standard description of ambiguous positions. These were tranformed into consensus sequences by hand. Three motifs were removed from the data set since they were duplicate motifs described under different names.

All motifs in the PLACE database were described as consensus sequences. Four motifs were removed from the data set as they contained the symbol 'U'.

In total this gave 581 motifs, with 21 from JASPAR, 464 from PLACE and 96 from AGRIS. See also Appendix B.

2.4 Prediction of *cis*-regulatory elements

To predict *cis*-regulatory elements in the promoter regions the motif consensus sequences (and its reverse complements) were converted into regular expression patterns and then queried against the promoter sequences. The degenerate positions were described as a set of possible nucleotides using the

regular expression square bracket syntax. The sets were constructed in such a way as to allow for matching possible degenerate positions in the sequences. Because of this, some considerations were needed to be given when constructing the sets as some degenerate positions describe subsets of other degenerate positions. For instance is the set of nucleotides A, T and G described by the letter D, while the set A and T is described by the letter W. The letter W would then have to match the set $\{A, T, W\}$, but not match the letter D, even though that position possibly would contain an A or a T. The letter D on the other hand, would match the set $\{A, C, T, G, W, D\}$, since W can be considered a subset of D. Table 1.1 lists all sets needed to properly construct the regular expressions.

2.5 CRE conservation in families

A CRE is considered conserved in a gene family if it is found once in the promoter sequence in all species within a given clade. A Monte Carlo approach was used to test if a CRE is conserved in significantly many families. To calculate the p-values the list of gene families were randomly shuffled and the number of random families a CRE was conserved in was compared to the observed number of families it was conserved in. The shuffling step was repeated 1000 times.

For pragmatic reasons the significance threshold was set to $p \leq 0.001$. This is the second lowest p-value possible when the number of bootstrapping iterations is 1000. For a simple Bonferroni correction (Weisstein, 2014) to be meaningful the number of bootstrapping iterations would have to be several orders of magnitude larger. This threshold is a reasonable a trade-off between strict multiple testing corrections and inaccuracies in the p-value estimation.

2.6 Validation using gene expression data

Gene expression data from rice (*ssp. japonica*) and poplar (*Populus trichocarpa*) was used to investigate whether the presence of a predicted CRE in the promoter region show different expression signatures depending on whether the CRE is found to be conserved or non-conserved and if this difference depend on the motif being conserved in significantly many families or not. Data from 711 and 463 gene expression experiments for rice and poplar, respectively, was used (Netotea et al., 2014).

For each CRE the the genes where it was predicted were divided into two groups, one consisting of the genes in the families where the CRE was

predicted to be conserved and one where the CRE was not conserved. For rice conservation was defined at both the level of Oryza and monocots. For poplar conservation was defined across fabids and rosids.

Let C_{mc} be the $n_c * n_c$ correlation matrix for genes with CRE m that are conserved, where n_c is the number of genes with conserved CRE m in the promoter region. Similarly, let C_{mu} the $n_u * n_u$ correlation matrix for genes with CRE m that are not conserved. A measure of total coexpression for a set of genes t_{mh} is

$$t_{mh} = \sum_{i=1}^{n_h} \sum_{j=1}^{n_h} c_{ij}^2$$

where c_{ij} is the correlation between genes i and j and the h subscript is either c or u, indicating the set of genes with conserved or non-conserved CRE m.

A relative measure of coexpression for a set of genes is

$$\theta_{mh} = \frac{t_{mh} - n_h}{n_h^2 - n_h}$$

To measure the extent of over-coexpression for a motif the total coexpression measure where calculated for a random set of n_c and n_u rice genes a thousand times. The proportion of the randomized coexpression measures greater or equal to the observed t_{mc} and t_{mu} respectively. This measure of over-coexpression is equivalent with a p-value.

A two-sided paired two sample t-test for the coexpression values was used to determine whether the rice genes in the gene families where the CRE is conserved is different from those from the gene families where the motif was non conserved.

A two-sided two independent samples t-test for the coexpression values is used to test the hypothesis that the difference between the CREs that were conserved and conserved in significantly many clades were greater than 0. CREs with less than 5 predicted instances in either group were filtered out. Since this procedure is performed four times (one time for each of the two levels of conservation in the two species) a Bonferroni correction (Weisstein, 2014) for testing multiple hypotheses is applied to the traditional significance threshold of 0.05, giving the corrected threshold of 0.0125.

2.6.1 Motif combinations

The same approach were also used with combinations of motifs in rice. Using only the 144 motifs that where investigated for conserved in significantly

many families, all $n * n - 1/2 = 10296$ possible combinations of two motifs were identified in rice and conservation at the level of Oryza and monocots were determined.

2.7 GO enrichment

An enrichment analysis seeks to identify annotations that are statistically overrepresented in a set of genes. The typical goal is to validate whether a procedure or method for selecting a set of genes yields a set of genes with common functionality. Gene Ontology (GO) annotations are common resource to use in an enrichment analysis.

The GO project (Ashburner et al., 2000) defines a vocabulary to describe the functionality of genes and gene products. Each definition, or *term*, in the vocabulary are also related to other terms in well defined relationships. This allows for different levels of precision in an annotation, reflecting the state of the knowledge about a gene. The vocabulary is ordered in three main categories that complements each other. *Biological processes* describes series of interacting physiological and biochemical events. *Molecular function* describes what mechanisms or roles a gene product may perform. *Cellular component* represent locations in the cell where the gene product is active.

Several tools and methods for enrichment analysis exists (Huang et al., 2009). An analysis taking an unordered list of preselected genes, without any associated covariates (e.g. a measure for differential expression) are referred to as Singular Enrichment Analysis (SEA). The next couple of paragraphs will explain the methodology in more general terms, before the specifics are detailed in chapter 2.7.4

2.7.1 Hypergeometric SEA

Consider a list of genes selected from a larger population of genes. This population could be all the genes in one organism. All these genes are annotated with one or more GO terms. The relationship between the selected genes and the given GO term could be illustrated with a 2×2 table.

	Has GO term	Has not GO term	Total
Selected genes	n_{11}	n_{12}	$n_{1.}$
Rest of the genes	n_{21}	n_{22}	$n_{2.}$
Total	$n_{.1}$	$n_{.2}$	$n_{..}$

What is of interest here is to test if there is a positive association between the selected genes and the GO term. Specifically, the question is whether

the observed value of n_{11} in the table is greater than on would expect with a random selection of genes (*i.e.* by chance), given that the marginal totals are not changed. The hypotheses to test are

$$H_0 : n_{11} \leq n_{1.} \times n_{.1}$$
$$H_1 : n_{11} > n_{1.} \times n_{.1}$$

This situation can be suitably modeled by the hypergeometric probability distribution (Ewens and Grant, 2005). Let N be the total number of genes, m be the number of genes that has been selected, n be the number of genes annotated with the GO term in question. The probability of having y genes annotated with the GO term among the selected genes is

$$P(y) = \frac{\binom{n}{y}\binom{N-n}{m-y}}{\binom{N}{m}} = \frac{\binom{n_{.1}}{n_{11}}\binom{n_{.2}}{n_{12}}}{\binom{n_{..}}{n_{1.}}}$$

Suppose y_0 of the m selected genes is annotated with the GO term. The p-value for the test is

$$\sum_{y=y_0}^{m} P(y)$$

2.7.2 Correction for non-independent terms

Given a set of genes it is often of interest to figure out which GO terms are enriched, not if just one specific term is enriched. In other words the testing procedure is done a large number of times with a different term each time, perhaps using all terms in the available annotation. This is a problem since a set of GO terms are not necessarily independent of each other, but may be explicitly dependent.

The *elim* algorithm alleviates this problem by taking into account the relationship structure of the GO terms that is tested (Alexa et al., 2006). The algorithm first models the relationships between the terms as a directed acyclic graph, with each term as a node and the edges between them the relationships between the terms. The terms, or nodes, farthest from the one of the root nodes (the term for one of the three main categories) are tested for enrichment first. If any of those terms are deemed significantly enriched, then the terms in the path closer to the root node are removed from the annotations for that set of genes. The testing procedure then repeatedly proceeds to the nodes closer to the root and do the annotation elimination as needed.

The algorithm ensures that the most specific (and most biologically interesting) terms are tested first and independent of each other. By removing gene annotations for less specific terms the enrichment tests on those terms are not dependent on the significance of the more specific terms.

2.7.3 Correction for testing multiple hypotheses

When a large number of tests is done, then all p-values from the individual tests should not be directly compared to the significance treshold α, as this will result in a large number of false positives. Instead the p-values can be used to compute the False Discovery Rate (FDR) and the accompanying q-values. Instead of the classical significance threshold α, the threshold α_{FDR} is used and is interpreted as the expected proportion of false positives among the tests.

In this case the q-values should be interpreted with some care as the correction procedure assumes that the tests are statistically independent. In this case that assumption does not hold since the GO terms have clearly defined connections.

The `qvalue` package in R (Dabney et al.) was used to compute the FDR.

2.7.4 Procedure

The *elim* algorithm implemented in the R package `topGO` (Alexa and Rahnenfuhrer, 2010) was used to investigate whether the presence of a binding site in the promoter region of rice (*ssp. japonica*) genes show different enrichments levels in annotated GO terms depending on whether the motif is found to be conserved or non-conserved. GO annotations from PLAZA was used. The elim algorithm was run with default parameter settings.

Significantly enriched GO-terms was identified with false discovery rate of 0.1. Conservation at the level of Oryza and monocots were investigated.

The enriched GO terms may me different in the conserved and nonconserved sets of genes. A measure of overlap, V, between the two sets is defined as

$$V(A, B) = \frac{|A \cap B|}{min(|A|, |B|)}$$

where A and B are the set of significantly enriched terms for the genes where the CRE was conserved and nonconserved, respectively. This is a number between 0 and 1, and is the proportion of the smaller set that is found in the other set.

For each motif, the specificity of the associated enriched terms can be measured by the maximum number of relations (or edges) between the term and the root term. This can be computed by finding the parent terms for the current term and recursively finding the next set af parents until all routes to the root node has been traversed. The GO database and associated functions as as implemented in version 2.14.0 of the `GO.db` package (Carlson, 2014) was used. A low number of ancestors imply a more general functionality, while a greater number imply a more specific function. The overall specificity of all enriched terms for a set of genes are combined into a single number, λ, using the median function.

To assess whether the motifs that are conserved in significantly many families have more enriched terms than the conserved but in nonsignificant many families the he Wilcoxon rank sum test for two independent samples was used.

2.8 Comparative analyses

Two clades can be compared by identifying the motifs that are conserved in significantly many families in either one of the clades. If a motif is conserved in a clade of interest (i.e. one of the clades in the comparison), but non-conserved in any of the subclades, it is not considered conserved.

Chapter 3

Results

3.1 Binding site identification

A total of 419,806,966 binding site instances was predicted across the 25 species, yielding an average 16,792,278.64 of instances per species. Table 3.1 summarizes the specieswise prevalence of binding sites. At least one *cis*-regulatory element were predicted in all promoter sequence in all species.

Species	Total sites found, incl duplicates	Average number of sites found per promoter, incl duplicates	Average number of sites found per promoter, excl duplicates	Average number of genes per motif	Number of genes
A. thaliana	18,150,509	540.16	121.00	8035.15	33,602
C. papaya	13,106,840	466.90	109.63	6306.16	28,072
C. reinhardtii	7,422,764	440.76	107.40	3848.26	16,841
O. lucimarinus	3,874,596	496.42	101.74	1792.53	7,805
O. sativa	29,167,041	503.97	123.14	14397.14	57,874
P. patens	18,148,010	502.20	112.93	8414.58	36,137
P. trichocarpa	21,809,727	525.27	116.36	9820.08	41,521
S. bicolor	12,472,213	359.57	101.74	7352.11	34,686
V. vinifera	14,198,232	532.89	116.59	6405.08	26,644
A. lyrata	17,339,001	530.73	119.77	8135.23	32,670
B. distachyon	13,220,307	495.55	122.80	6754.73	26,678
G. max	25,576,820	549.93	118.70	11108.29	46,509
M. domestica	39,394,103	413.67	102.11	19564.57	95,230
M. esculenta	14,181,529	460.44	107.33	6774.38	30,800
M. sp. RCC299	4,533,576	441.18	102.57	2352.76	10,276
O. sativa	28,920,456	486.63	120.34	14361.07	59,430
O. tauri	3,903,083	480.91	104.80	1911.44	8,116
R. communis	14,075,634	450.84	100.15	6528.02	31,221
S. moellendorffii	10,198,317	457.63	111.17	5161.52	22,285
V. carteri	6,852,405	440.84	112.71	3719.57	15,544
Z. mays	12,578,986	317.68	94.86	7696.98	39,597
L. japonicus	19,151,928	274.99	69.87	9992.54	69,647
M. truncatula	30,988,913	538.12	117.96	13750.66	57,587
F. vesca	17,262,212	495.91	115.61	8213.09	34,809
T. cacao	23,279,764	503.14	111.02	10526.18	46,269

Table 3.1: The second column from the left is the total number of binding sites found in all promoter sequences, including duplicates of the same binding site in the same promoter. The third column is the average number of unique binding sites found in each promoter. The fourth column is the average number of genes each motif was found in.

3.2 Motifs conserved in significantly many families

Due to limited computational resources, testing if a CRE is conserved in significantly many families was performed on only 144 of the 581 motifs available. 29 of these motifs were from AGRIS, 5 from JASPAR and 110 from PLACE. Table 3.2 summarizes the findings, while the full results are found in the appendix.

Clade	Number of significant motifs ($p \leq 0.001$)
Eudicots	33
Monocots	67
Chlorophyta	18
Rosids	29
Fabids	26
Malvids	57
Angiosperm	30
Green Plants	4
Oryza	128

Table 3.2: Summary of the number of motifs found to be conserved in significantly many families by clade.

3.3 Differences in annotational enrichment between conserved and non-conserved binding sites

73.9% (of $N = 142$) of the CREs had more significant enriched GO terms in the genes in families where the CRE was conserved across Oryza than in the genes in families where it was not conserved. The genes where the CRE was conserved had on average 32.98 more significantly enriched terms.

No significant differences in the number of annotated GO terms between the genes where the CRE was conserved and conserved in significantly many families was found ($W = 949.5, p = 0.5093$, Wilcoxon rank sum test)

54.89% (of $N = 136$) of the CREs had more significant enriched GO terms in the genes in families where the CRE was conserved across monocots than in the genes in families where it was not conserved. The genes where the CRE was conserved had on average 16.74 more significantly enriched terms.

No significant differences in the number of annotated GO terms between the genes where the motif was conserved and conserved in significantly many families was found ($W = 893, p = 0.4784$, Wilcoxon rank sum test)

The number of significant terms for each CRE, along with other statistics, is found in Table 3.3.

	Oryza					Monocots				
	Conserved		Nonconserved			Conserved		Nonconserved		
Motif name	N	λ	N	λ	V	N	λ	N	λ	V
ABFs binding site motif	82	7.0	5	8.0	0.00	50	6.5	5	7.0	0.20
ATHB2 binding site motif	60	7.5	3	6.0	0.00	29	6.0	1	6.0	0.00
CArG promoter motif	112	7.0	0	-	-	96	7.0	0	-	-
CArG2 motif in AP3	0	-	1	5.0	-	0	-	1	5.0	-
CCA1 motif1 BS in CAB1	40	6.0	33	7.0	0.00	10	7.0	26	7.0	0.00
DREB1AND2 BS in rd29a	37	6.0	1	8.0	0.00	0	-	1	8.0	-
EIL2 BS in ERF1	0	-	80	7.0	-	0	-	80	7.0	-
ACE promoter motif	16	5.5	0	-	-	-	-	-	-	-
ERE promoter motif	2	7.5	0	-	-	-	-	-	-	-
GCC-box promoter motif	127	7.0	13	8.0	0.00	131	7.0	43	8.0	0.00
HSEs binding site motif	86	7.0	54	7.0	0.04	51	6.0	55	7.0	0.04
L1-box promoter motif	105	7.0	6	7.5	0.17	94	7.0	15	8.0	0.07
LS5 promoter motif	21	6.0	26	6.0	0.05	0	-	25	6.0	-
MYB1 binding site motif	115	7.0	6	8.0	0.17	86	7.0	3	8.0	0.33
MYB2 binding site motif	56	6.0	0	-	-	17	7.0	0	-	-
MYB3 binding site motif	81	7.0	72	7.5	0.04	46	6.0	44	7.0	0.05
Nonamer promoter motif	55	6.0	0	-	-	20	5.0	0	-	-
AG BS in SUP	45	7.0	0	-	-	14	6.5	0	-	-
SBP-box promoter motif	73	6.0	99	7.0	0.04	42	6.0	69	7.0	0.05
T-box promoter motif	121	7.0	3	8.0	0.00	122	7.0	14	7.0	0.07
TELO-box promoter motif	81	7.0	110	7.0	0.09	45	6.0	98	7.0	0.02
Z-box promoter motif	80	7.0	239	7.0	0.06	41	6.0	254	7.0	0.10
AG BS in SPL/NOZ	0	-	3	8.0	-	0	-	3	8.0	-
AGL2 binding site motif	50	6.0	0	-	-	10	4.5	0	-	-
SORLREP1	66	7.0	0	-	-	24	6.0	0	-	-
SORLIP3	54	6.0	0	-	-	19	7.0	0	-	-
SORLIP5	107	7.0	7	8.0	0.00	81	7.0	12	7.0	0.08
MA0127.1	79	7.0	1	8.0	0.00	37	6.0	0	-	-
MA0096.1	101	7.0	1	8.0	0.00	103	7.0	0	-	-
MA0120.1	116	7.0	0	-	-	112	7.0	0	-	-
MA0129.1	111	7.0	150	8.0	0.06	99	7.0	278	7.0	0.04
MA0001.1	57	6.0	1	5.0	1.00	35	5.0	1	5.0	1.00
-10PEHVPSBD	138	7.0	16	8.0	0.06	121	7.0	51	8.0	0.02
ARE1	89	7.0	0	-	-	57	7.0	2	11.0	0.50
BOX1PSGS2	27	6.0	0	-	-	-	-	-	-	-
BP5OSWX	94	7.0	0	-	-	72	7.0	0	-	-
CAREOSREP1	131	7.0	0	-	-	125	7.0	0	-	-
CATATGGMSAUR	128	7.0	0	-	-	109	7.0	0	-	-
CEREGLUBOX3PSLEGA	65	7.0	34	7.0	0.06	26	6.0	34	7.5	0.00
CGACGOSAMY3	140	7.0	1	8.0	0.00	129	7.0	4	7.5	0.00
COREOS	15	7.0	0	-	-	-	-	-	-	-
CPRFPCCHS	34	6.0	11	7.0	0.09	5	6.0	8	7.0	0.40
DRE1COREZMRAB17	95	7.0	119	8.0	0.05	78	7.0	173	7.0	0.04
AACACOREOSGLUB1	122	7.0	8	7.0	0.12	122	7.0	14	7.0	0.14
E2FANTRNR	65	7.0	0	-	-	35	6.0	0	-	-
E2FAT	54	6.5	0	-	-	28	6.0	0	-	-
E2FBNTRNR	72	6.0	74	7.0	0.04	40	6.0	58	7.0	0.03
EMHVCHORD	90	7.0	0	-	-	52	7.0	1	5.0	1.00
EREGCC	2	7.5	1	10.0	0.00	0	-	1	10.0	-

Continues on next page

	Oryza					Monocots				
	Conserved		Nonconserved			Conserved		Nonconserved		
Motif name	N	λ	N	λ	V	N	λ	N	λ	V
AAGACGTAGATACL12	33	6.0	4	8.0	0.00	0	-	4	8.0	-
ERELEE4	96	7.0	0	-	-	119	7.0	0	-	-
EVENINGAT	84	7.0	279	7.0	0.05	47	6.0	328	7.0	0.11
GAGAGMGSA1	57	7.0	0	-	-	16	6.0	0	-	-
ARECOREZMGAPC4	17	7.0	6	8.0	0.00	0	-	6	8.0	-
GBOXLERBCS	75	7.0	6	6.5	0.17	35	6.0	4	5.5	0.25
GBOXRELOSAMY3	8	5.0	0	-	-	-	-	-	-	-
ARR1AT	181	7.0	0	-	-	180	7.0	0	-	-
GCBP2ZMGAPC4	54	7.0	311	7.0	0.09	19	6.0	312	7.0	0.11
GCCCORE	127	7.0	0	-	-	131	7.0	0	-	-
GLUTEBOX2OSGT2	7	7.0	0	-	-	-	-	-	-	-
GLUTECOREOS	9	7.0	67	7.0	0.00	0	-	63	7.0	-
HBOXPVCHS15	35	7.0	0	-	-	13	6.0	0	-	-
HDZIPIIIAT	6	6.5	0	-	-	-	-	-	-	-
HEXAMERATH4	117	7.0	0	-	-	118	7.0	0	-	-
HSRENTHSR203J	22	6.5	0	-	-	-	-	-	-	-
HY5AT	8	7.0	0	-	-	-	-	-	-	-
IBOXCORE	172	7.0	0	-	-	149	7.0	0	-	-
IBOXLSCMCUCUMISIN	6	6.0	199	7.0	0.00	0	-	199	7.0	-
INRNTPSADB	166	7.0	15	7.0	0.00	150	7.0	22	8.0	0.05
L1BOXATPDF1	105	7.0	207	8.0	0.06	94	7.0	343	7.0	0.04
ABRE2HVA22	24	7.0	122	8.0	0.00	0	-	102	8.0	-
ABRE3HVA1	14	6.0	15	8.0	0.00	0	-	13	8.0	-
MRNA3ENDTAH3	50	6.0	0	-	-	14	5.0	0	-	-
MYB1LEPR	98	7.0	0	-	-	88	7.0	0	-	-
MYB26PS	80	7.0	0	-	-	51	6.0	0	-	-
ABRE3OSRAB16	12	6.5	0	-	-	-	-	-	-	-
MYBGAHV	104	7.0	0	-	-	106	7.0	0	-	-
MYCATERD1	139	7.0	0	-	-	128	7.0	0	-	-
MYCATRD22	139	7.0	0	-	-	128	7.0	0	-	-
ABREA2HVA1	57	6.0	0	-	-	-	-	-	-	-
NTBBF1ARROLB	125	7.0	0	-	-	123	7.0	3	5.0	0.00
O2F1BE2S1	21	6.0	0	-	-	-	-	-	-	-
O2F2BE2S1	29	7.0	0	-	-	2	6.0	0	-	-
O2F3BE2S1	31	6.0	0	-	-	-	-	-	-	-
OCETYPEIIIINTHISTONE	4	9.5	12	6.0	0.25	0	-	12	6.0	-
OCSGMHSP26A	0	-	25	6.0	-	0	-	25	6.0	-
OSE1ROOTNODULE	138	7.0	0	-	-	132	7.0	0	-	-
OSE2ROOTNODULE	190	7.0	17	8.0	0.06	165	7.0	31	7.0	0.03
P1BS	113	7.0	31	8.0	0.03	112	7.0	79	8.0	0.03
PALINDROMICCBOXGM	50	7.0	0	-	-	21	6.0	0	-	-
PE2FNTRNR1A	56	6.5	0	-	-	19	5.0	0	-	-
PIATGAPB	60	7.0	7	6.0	0.14	22	6.0	4	6.5	0.50
ABREBNNAPA	21	7.0	0	-	-	-	-	-	-	-
PRECONSCRHSP70A	158	7.0	14	8.0	0.00	124	7.0	23	8.0	0.00
QELEMENTZMZM13	122	7.0	2	8.5	0.00	111	7.0	1	10.0	0.00
RAV1AAT	185	7.0	0	-	-	170	8.0	0	-	-
RAV1BAT	121	7.0	72	8.0	0.07	120	7.0	135	8.0	0.03
RBCSBOX2PS	0	-	4	3.5	-	0	-	4	3.5	-
RBCSBOX3PS	12	6.5	3	5.0	0.00	-	-	-	-	-
20NTNTNOS	0	-	110	7.0	-	0	-	110	7.0	-
ABRECE1HVA22	41	6.0	12	5.5	0.08	15	6.0	13	6.0	0.00
RE1ASPHYA3	17	6.0	88	7.0	0.24	0	-	73	7.0	-
RHERPATEXPA7	154	7.0	13	7.0	0.00	139	7.0	31	7.0	0.03
ABRECE3HVA1	6	9.0	0	-	-	-	-	-	-	-
RYREPEAT4	14	7.0	4	11.0	0.00	0	-	4	11.0	-
RYREPEATLEGUMINBOX	121	7.0	0	-	-	120	7.0	0	-	-
RYREPEATVFLEB4	89	7.0	0	-	-	76	6.5	0	-	-

Continues on next page

	Oryza					Monocots				
	Conserved		Nonconserved			Conserved		Nonconserved		
Motif name	N	λ	N	λ	V	N	λ	N	λ	V
SB1NPABC1	0	-	4	9.5	-	0	-	4	9.5	-
SBOXATRBCS	94	7.0	0	-	-	65	7.0	0	-	-
SGBFGMGMAUX28	34	7.0	202	7.0	0.15	0	-	137	7.0	-
SITEIIAOSPCNA	50	6.5	0	-	-	8	7.5	0	-	-
SORLIP4AT	44	6.5	201	7.0	0.11	17	6.0	132	7.0	0.06
SORLREP4AT	66	7.0	1	8.0	0.00	17	7.0	0	-	-
ABREMOTIFAOSOSEM	69	6.0	0	-	-	32	6.0	0	-	-
TATABOX1	52	6.5	0	-	-	25	5.0	0	-	-
ATHB5ATCORE	73	7.0	328	7.0	0.07	39	6.0	405	7.0	0.08
TATAPVTRNALEU	109	7.0	0	-	-	103	7.0	0	-	-
TATCCACHVAL21	102	7.0	156	8.0	0.08	84	7.0	181	8.0	0.05
TE2F2NTPCNA	50	7.0	0	-	-	20	5.5	0	-	-
TGA1ANTPR1A	0	-	7	5.0	-	0	-	7	5.0	-
TRANSTART	8	7.0	0	-	-	-	-	-	-	-
UP1ATMSD	101	7.0	6	7.0	0.00	69	7.0	3	7.0	0.33
WRKY71OS	181	7.0	0	-	-	177	7.0	0	-	-
WUSATAg	92	7.0	1	12.0	0.00	95	7.0	1	12.0	0.00
XYLAT	100	7.0	0	-	-	90	7.0	0	-	-
ABREZMRAB28	77	7.0	2	5.5	0.00	32	6.0	0	-	-
ACGTROOT1	40	7.0	149	7.0	0.07	0	-	98	7.0	-
ACGTSEED3	26	6.0	14	6.0	0.00	0	-	13	6.0	-
AGATCONSENSUS	50	6.0	229	7.0	0.10	12	6.5	132	7.0	0.17
AGL1ATCONSENSUS	22	8.0	13	6.0	0.00	0	-	10	5.5	-
AGL2ATCONSENSUS	50	6.0	16	6.5	0.19	10	4.5	10	7.5	0.00
AGL3ATCONSENSUS	15	6.0	40	6.0	0.00	0	-	37	6.0	-
ALF1NTPARC	0	-	3	10.0	-	0	-	3	10.0	-
AMMORESIVDCRNIA1	94	7.0	0	-	-	80	7.0	0	-	-
ANAERO3CONSENSUS	111	7.0	2	7.5	0.00	104	7.0	1	7.0	0.00
ANAERO5CONSENSUS	47	6.0	17	8.0	0.06	22	5.5	13	7.0	0.00

Table 3.3: Results from enrichment analysis, for conservation across Oryza and monocots. The number of significant GO terms for each motif is found in the columns N, while the generality score for the enriched terms are found in the λ columns. The W columns is the overlapping score between the enriched terms in the conserved genes and the nonconserved genes.

3.4 Gene expression validation

The coexpression values for rice and poplar are found in table 3.5 and 3.6, respectively. As seen in Table 3.4 there were no difference in expression score between the conserved and non-conserved CREs.

Due to limited computational resources only 98 CREs in rice and 105 CREs in populus were considered.

In rice, when considered across monocots, no difference between the conserved CREs and the CREs that were found to be conserved in significantly many families were detected ($p = 0.056$). This test was not performed on conservation across Oryza, as the filtering of few predicted CRE instances

removed all instances not conserved in significantly many families.

In populus, when conserved across fabids, no difference between the conserved CREs and the CREs that were found to be conserved in significantly many families were detected ($p = 0.28$). No significant difference when conservation is considered across rosids was found as well ($p = 0.035$)

Species	Clade	Mean difference (95% CI)	t	p-value
Rice	Oryza	-0.0179 (-0.0256, -0.0102)	-4.693	3.1×10^5
	Monocots	-0.0151 (-0.0220, -0.00826)	-4.461	7×10^5
Poplar	Fabids	0.00353 (-0.0128 0.0199)	0.438	0.664
	Rosids	0.0111 (-0.00984, 0.0321)	1.0764	0.289

Table 3.4: Results paired t-test for the difference in coexpression scores conserved vs. nonconserved. Only pairs where a motif was found in at least 6 instances in either cateogry included.

	Oryza						Monocots					
	conserved			nonconserved			conserved			nonconserved		
Motif name	θ_c	p-value	N	θ_u	p-value	N	θ_c	p-value	N	θ_u	p-value	N
ATHB2 binding site motif	0.075	0.945	12	0.108	0.745	55	0.044	0.828	3	0.107	0.806	64
CArG promoter motif	0.105	0.984	310	0.119	0.347	208	0.105	0.963	179	0.113	0.744	339
DREB1AND2 BS in rd29a	0.144	0.055	30	0.110	0.904	304	-	-	0	0.112	0.828	334
EIL2 BS in ERF1	-	-	0	0.114	1.0	13717	-	-	0	0.114	1.0	13717
ACE promoter motif	0.054	0.752	3	0.113	0.619	56	-	-	0	0.112	0.631	59
ERE promoter motif	-	-	0	0.122	0.131	232	-	-	0	0.122	0.158	232
L1-box promoter motif	0.108	0.99	641	0.131	0.001	447	0.105	0.996	373	0.125	0.003	715
MYB1 binding site motif	0.106	1.0	770	0.131	0.002	466	0.106	0.998	478	0.123	0.023	758
Nonamer promoter motif	0.105	0.781	37	0.107	0.893	142	0.139	0.213	9	0.107	0.899	170
AG BS in SUP	0.062	0.69	3	0.178	0.123	5	-	-	0	0.137	0.25	8
PII promoter motif	-	-	0	0.115	0.569	268	-	-	0	0.115	0.586	268
AG BS in SPL/NOZ	-	-	0	0.112	0.855	421	-	-	0	0.112	0.855	421
SORLREP1	0.112	0.743	162	0.106	0.776	44	0.105	0.667	18	0.111	0.815	188
SORLIP3	0.117	0.441	100	0.132	0.135	37	0.099	0.632	8	0.122	0.24	129
MA0127.1	0.115	0.668	434	0.129	0.054	147	0.112	0.706	121	0.120	0.124	460
MA0096.1	0.106	0.693	26	0.179	0.064	7	0.117	0.446	14	0.120	0.392	19
MA0120.1	0.121	0.21	191	0.135	0.007	129	0.105	0.928	120	0.143	0.0	200
CAREOSREP1	0.101	0.783	19	0.000	1.0	1	0.091	0.837	14	0.086	0.721	6
CATATGGMSAUR	0.097	0.934	40	0.074	0.949	11	0.092	0.928	27	0.092	0.915	24
COREOS	-	-	0	0.110	0.584	20	-	-	0	0.110	0.567	20
E2FANTRNR	0.277	0.018	4	0.175	0.163	4	-	-	0	0.213	0.026	8
E2FBNTRNR	0.103	1.0	1822	0.123	0.0	3886	0.104	1.0	902	0.119	0.006	4806
ELRE2PCPAL1	-	-	0	0.119	0.299	337	-	-	0	0.119	0.278	337

Continues on next page

	Oryza						Monocots					
	conserved			nonconserved			conserved			nonconserved		
Motif name	θ_c	p-value	N	θ_u	p-value	N	θ_c	p-value	N	θ_u	p-value	N
EREGCC	-	-	0	0.113	0.711	191	-	-	0	0.113	0.701	191
AAGACGTAGATACL12	0.112	0.556	23	0.121	0.118	343	-	-	0	0.120	0.175	366
ERELEE4	0.111	0.527	15	0.060	0.919	6	0.112	0.511	10	0.107	0.606	11
GAGAGMGSA1	0.360	0.071	2	0.188	0.078	6	-	-	0	0.181	0.035	8
GCCCORE	0.111	0.843	265	0.171	0.0	58	0.108	0.932	214	0.155	0.0	109
HDZIPIIIAT	-	-	0	0.102	0.728	20	-	-	0	0.102	0.753	20
HSRENTHSR203J	0.003	0.901	2	0.127	0.227	37	-	-	0	0.126	0.225	39
HY5AT	-	-	0	0.119	0.385	38	-	-	0	0.119	0.442	38
IBOXCORE	0.115	0.582	120	0.102	0.551	5	0.110	0.752	103	0.158	0.022	22
MRNA3ENDTAH3	0.067	0.512	2	0.123	0.303	25	0.000	1.0	1	0.119	0.401	26
MYB1LEPR	0.105	0.788	47	0.124	0.288	24	0.110	0.63	26	0.113	0.547	45
MYB26PS	0.118	0.407	70	0.126	0.105	119	0.115	0.5	35	0.125	0.09	154
ABRE3OSRAB16	0.125	0.389	3	0.117	0.399	353	-	-	0	0.117	0.415	356
ABREA2HVA1	0.094	0.862	21	0.113	0.622	124	-	-	0	0.110	0.813	145
NTBBF1ARROLB	0.115	0.613	344	0.121	0.326	59	0.112	0.801	274	0.127	0.099	129
O2F1BE2S1	0.162	0.249	3	0.112	0.655	82	-	-	0	0.113	0.612	85
O2F2BE2S1	0.086	0.444	2	0.136	0.195	16	-	-	0	0.129	0.238	18
O2F3BE2S1	0.000	1.0	1	0.103	0.603	9	-	-	0	0.128	0.334	10
PE2FNTRNR1A	0.042	0.848	3	0.057	0.831	4	-	-	0	0.080	0.8	7
ABREBNNAPA	0.048	0.899	4	0.119	0.384	66	-	-	0	0.118	0.442	70
QELEMENTZMZM13	0.109	0.816	110	0.158	0.023	26	0.106	0.865	95	0.162	0.003	41
RAV1AAT	0.103	0.87	52	0.184	0.094	6	0.110	0.637	48	0.100	0.667	10
RBCSBOX3PS	0.134	0.339	3	0.116	0.483	465	-	-	0	0.116	0.465	468

Continues on next page

	Oryza						Monocots					
	conserved			nonconserved			conserved			nonconserved		
Motif name	θ_c	p-value	N	θ_u	p-value	N	θ_c	p-value	N	θ_u	p-value	N
RYREPEAT4	-	-	0	0.114	0.583	161	-	-	0	0.114	0.62	161
RYREPEATLEGUMINBOX	0.104	0.804	36	0.122	0.401	5	0.096	0.881	26	0.146	0.135	15
RYREPEATVFLEB4	0.097	0.82	17	0.140	0.077	31	0.080	0.855	9	0.139	0.062	39
SITEIIAOSPCNA	0.114	0.515	19	0.117	0.481	205	0.100	0.586	6	0.117	0.444	218
SORLREP4AT	0.092	0.961	37	0.130	0.048	114	0.072	0.957	11	0.121	0.232	140
TATABOX1	0.033	0.892	3	0.223	0.014	8	-	-	0	0.176	0.033	11
TATAPVTRNALEU	0.115	0.566	122	0.146	0.009	59	0.106	0.837	70	0.135	0.015	111
WRKY71OS	0.114	0.579	137	0.139	0.211	10	0.118	0.418	126	0.101	0.743	21
XYLAT	0.085	1.0	116	0.116	0.521	114	0.096	0.978	75	0.102	0.982	155
ABREZMRAB28	0.109	0.796	126	0.117	0.417	328	0.136	0.1	36	0.115	0.587	418
ACGTSEED3	0.094	1.0	179	0.114	0.953	4028	-	-	0	0.113	0.992	4207
ANAERO3CONSENSUS	0.123	0.292	48	0.124	0.355	16	0.123	0.324	35	0.122	0.337	29
ARE1	0.087	0.993	47	0.127	0.136	75	0.086	0.98	28	0.118	0.38	94
ARR1AT	0.123	0.236	59	0.507	0.019	2	0.122	0.317	52	0.141	0.211	9
ABFs binding site motif	0.107	0.996	473	0.123	0.159	156	0.091	1.0	121	0.113	0.732	508
ATHB2 binding site motif	0.075	0.945	12	0.108	0.745	55	0.044	0.828	3	0.107	0.806	64
CArG promoter motif	0.105	0.984	310	0.119	0.347	208	0.105	0.963	179	0.113	0.744	339
CArG2 motif in AP3	-	-	0	0.113	0.854	845	-	-	0	0.113	0.874	845
DREB1AND2 BS in rd29a	0.144	0.055	30	0.110	0.904	304	-	-	0	0.112	0.828	334
EIL2 BS in ERF1	-	-	0	0.114	1.0	13717	-	-	0	0.114	1.0	13717
ACE promoter motif	0.054	0.752	3	0.113	0.619	56	-	-	0	0.112	0.631	59
ERE promoter motif	-	-	0	0.122	0.131	232	-	-	0	0.122	0.158	232
L1-box promoter motif	0.108	0.99	641	0.131	0.001	447	0.105	0.996	373	0.125	0.003	715

Continues on next page

	Oryza						Monocots					
	conserved			nonconserved			conserved			nonconserved		
Motif name	θ_c	p-value	N	θ_u	p-value	N	θ_c	p-value	N	θ_u	p-value	N
MYB1 binding site motif	0.106	1.0	770	0.131	0.002	466	0.106	0.998	478	0.123	0.023	758
Nonamer promoter motif	0.105	0.781	37	0.107	0.893	142	0.139	0.213	9	0.107	0.899	170
AG BS in SUP	0.062	0.69	3	0.178	0.123	5	-	-	0	0.137	0.25	8
PII promoter motif	-	-	0	0.115	0.569	268	-	-	0	0.115	0.586	268
AG BS in SPL/NOZ	-	-	0	0.112	0.855	421	-	-	0	0.112	0.855	421
SORLREP1	0.112	0.743	162	0.106	0.776	44	0.105	0.667	18	0.111	0.815	188
SORLIP3	0.117	0.441	100	0.132	0.135	37	0.099	0.632	8	0.122	0.24	129
MA0127.1	0.115	0.668	434	0.129	0.054	147	0.112	0.706	121	0.120	0.124	460
MA0096.1	0.106	0.693	26	0.179	0.064	7	0.117	0.446	14	0.120	0.392	19
MA0120.1	0.121	0.21	191	0.135	0.007	129	0.105	0.928	120	0.143	0.0	200
MA0001.1	0.101	0.999	363	0.117	0.402	982	0.107	0.913	139	0.112	0.931	1206
CAREOSREP1	0.101	0.783	19	0.000	1.0	1	0.091	0.837	14	0.086	0.721	6
CATATGGMSAUR	0.097	0.934	40	0.074	0.949	11	0.092	0.928	27	0.092	0.915	24
CGACGOSAMY3	0.102	1.0	1343	0.109	0.833	154	0.102	1.0	1133	0.106	0.985	364
COREOS	-	-	0	0.110	0.584	20	-	-	0	0.110	0.567	20
E2FANTRNR	0.277	0.018	4	0.175	0.163	4	-	-	0	0.213	0.026	8
E2FBNTRNR	0.103	1.0	1822	0.123	0.0	3886	0.104	1.0	902	0.119	0.006	4806
ELRE2PCPAL1	-	-	0	0.119	0.299	337	-	-	0	0.119	0.278	337
EMHVCHORD	0.106	0.993	388	0.123	0.06	479	0.107	0.921	175	0.117	0.399	692
EREGCC	-	-	0	0.113	0.711	191	-	-	0	0.113	0.701	191
AAGACGTAGATACL12	0.112	0.556	23	0.121	0.118	343	-	-	0	0.120	0.175	366
ERELEE4	0.111	0.527	15	0.060	0.919	6	0.112	0.511	10	0.107	0.606	11
GAGAGMGSA1	0.360	0.071	2	0.188	0.078	6	-	-	0	0.181	0.035	8

Continues on next page

	Oryza						Monocots					
	conserved			nonconserved			conserved			nonconserved		
Motif name	θ_c	p-value	N	θ_u	p-value	N	θ_c	p-value	N	θ_u	p-value	N
GBOXLERBCS	0.101	1.0	418	0.120	0.101	847	0.100	0.986	122	0.115	0.731	1143
GCCCORE	0.111	0.843	265	0.171	0.0	58	0.108	0.932	214	0.155	0.0	109
HDZIPIIIAT	-	-	0	0.102	0.728	20	-	-	0	0.102	0.753	20
HSRENTHSR203J	0.003	0.901	2	0.127	0.227	37	-	-	0	0.126	0.225	39
HY5AT	-	-	0	0.119	0.385	38	-	-	0	0.119	0.442	38
IBOXCORE	0.115	0.582	120	0.102	0.551	5	0.110	0.752	103	0.158	0.022	22
MRNA3ENDTAH3	0.067	0.512	2	0.123	0.303	25	0.000	1.0	1	0.119	0.401	26
MYB1LEPR	0.105	0.788	47	0.124	0.288	24	0.110	0.63	26	0.113	0.547	45
MYB26PS	0.118	0.407	70	0.126	0.105	119	0.115	0.5	35	0.125	0.09	154
ABRE3OSRAB16	0.125	0.389	3	0.117	0.399	353	-	-	0	0.117	0.415	356
ABREA2HVA1	0.094	0.862	21	0.113	0.622	124	-	-	0	0.110	0.813	145
NTBBF1ARROLB	0.115	0.613	344	0.121	0.326	59	0.112	0.801	274	0.127	0.099	129
O2F1BE2S1	0.162	0.249	3	0.112	0.655	82	-	-	0	0.113	0.612	85
O2F2BE2S1	0.086	0.444	2	0.136	0.195	16	-	-	0	0.129	0.238	18
O2F3BE2S1	0.000	1.0	1	0.103	0.603	9	-	-	0	0.128	0.334	10
PE2FNTRNR1A	0.042	0.848	3	0.057	0.831	4	-	-	0	0.080	0.8	7
ABREBNNAPA	0.048	0.899	4	0.119	0.384	66	-	-	0	0.118	0.442	70
QELEMENTZMZM13	0.109	0.816	110	0.158	0.023	26	0.106	0.865	95	0.162	0.003	41
RAV1AAT	0.103	0.87	52	0.184	0.094	6	0.110	0.637	48	0.100	0.667	10
RBCSBOX3PS	0.134	0.339	3	0.116	0.483	465	-	-	0	0.116	0.465	468
RYREPEAT4	-	-	0	0.114	0.583	161	-	-	0	0.114	0.62	161
RYREPEATLEGUMINBOX	0.104	0.804	36	0.122	0.401	5	0.096	0.881	26	0.146	0.135	15
RYREPEATVFLEB4	0.097	0.82	17	0.140	0.077	31	0.080	0.855	9	0.139	0.062	39

Continues on next page

	Oryza						Monocots					
	conserved			nonconserved			conserved			nonconserved		
Motif name	θ_c	p-value	N	θ_u	p-value	N	θ_c	p-value	N	θ_u	p-value	N
SITEIIAOSPCNA	0.114	0.515	19	0.117	0.481	205	0.100	0.586	6	0.117	0.444	218
SORLREP4AT	0.092	0.961	37	0.130	0.048	114	0.072	0.957	11	0.121	0.232	140
TATABOX1	0.033	0.892	3	0.223	0.014	8	-	-	0	0.176	0.033	11
TATAPVTRNALEU	0.115	0.566	122	0.146	0.009	59	0.106	0.837	70	0.135	0.015	111
UP1ATMSD	0.114	0.738	513	0.121	0.122	409	0.114	0.655	267	0.119	0.21	655
WRKY71OS	0.114	0.579	137	0.139	0.211	10	0.118	0.418	126	0.101	0.743	21
WUSATAg	0.096	1.0	579	0.111	0.869	387	0.099	1.0	359	0.105	0.998	607
XYLAT	0.085	1.0	116	0.116	0.521	114	0.096	0.978	75	0.102	0.982	155
ABREZMRAB28	0.109	0.796	126	0.117	0.417	328	0.136	0.1	36	0.115	0.587	418
ACGTSEED3	0.094	1.0	179	0.114	0.953	4028	-	-	0	0.113	0.992	4207
ANAERO3CONSENSUS	0.123	0.292	48	0.124	0.355	16	0.123	0.324	35	0.122	0.337	29
ANAERO5CONSENSUS	0.117	0.423	241	0.150	0.0	1196	0.113	0.586	66	0.145	0.0	1371
ARE1	0.087	0.993	47	0.127	0.136	75	0.086	0.98	28	0.118	0.38	94
ARECOREZMGAPC4	0.120	0.381	21	0.116	0.557	954	-	-	0	0.116	0.561	975
ARR1AT	0.123	0.236	59	0.507	0.019	2	0.122	0.317	52	0.141	0.211	9

Table 3.5: The coexpression values for the genes constituting the families where the motifs are conserved and non-conserved. Results for motifs that where found in less than five promoters in all groups not shown.

Binding site	Fabids						Rosids					
	conserved			nonconserved			conserved			nonconserved		
	θ_c	p-value	N	θ_u	p-value	N	θ_c	p-value	N	θ_u	p-value	N
ABFs binding site motif	0.129	0.866	32	0.130	1.0	366	0.118	0.94	27	0.130	1.0	371
ATHB2 binding site motif	0.184	0.238	9	0.166	0.236	123	0.277	0.051	7	0.163	0.328	125
CArG promoter motif	0.152	0.717	468	0.146	0.991	863	0.156	0.564	417	0.145	0.994	914
CArG2 motif in AP3	-	-	0	0.150	0.994	2105	-	-	0	0.150	0.993	2105
DREB1AND2 BS in rd29a	-	-	0	0.160	0.349	188	-	-	0	0.160	0.376	188
EIL2 BS in ERF1	-	-	0	0.153	1.0	20181	-	-	0	0.153	1.0	20181
ACE promoter motif	-	-	0	0.139	0.858	62	-	-	0	0.139	0.851	62
ERE promoter motif	-	-	0	0.175	0.179	46	-	-	0	0.175	0.175	46
GCC-box promoter motif	0.153	0.771	583	0.155	0.729	1470	0.151	0.85	513	0.155	0.749	1540
L1-box promoter motif	0.155	0.696	971	0.165	0.024	1452	0.153	0.854	875	0.166	0.006	1548
MYB1 binding site motif	0.151	0.847	454	0.157	0.485	910	0.151	0.836	388	0.157	0.556	976
Nonamer promoter motif	-	-	0	0.157	0.519	465	-	-	0	0.157	0.513	465
AG BS in SUP	-	-	0	0.139	0.668	17	-	-	0	0.139	0.657	17
PII promoter motif	-	-	0	0.153	0.564	63	-	-	0	0.153	0.526	63
T-box promoter motif	0.147	0.999	1838	0.181	0.0	1022	0.148	0.997	1750	0.176	0.0	1110
AG BS in SPL/NOZ	-	-	0	0.150	0.898	483	-	-	0	0.150	0.889	483
SORLREP1	0.167	0.372	19	0.148	0.896	328	0.203	0.17	9	0.148	0.886	338
SORLIP3	0.293	0.008	14	0.143	0.963	332	0.394	0.007	7	0.145	0.955	339
SORLIP5	0.163	0.145	738	0.154	0.773	1509	0.163	0.148	647	0.155	0.773	1600
MA0127.1	0.155	0.574	535	0.158	0.469	1414	0.151	0.825	481	0.159	0.36	1468
MA0120.1	0.158	0.455	17	0.199	0.05	34	0.138	0.664	16	0.205	0.045	35
CEREGLUBOX3PSLEGA	0.164	0.234	242	0.161	0.124	1950	0.172	0.081	187	0.160	0.139	2005
CGACGOSAMY3	0.147	0.945	455	0.166	0.074	533	0.150	0.886	435	0.163	0.134	553

Continues on next page

Binding site	Fabids						Rosids					
	conserved			nonconserved			conserved			nonconserved		
	θ_c	p-value	N	θ_u	p-value	N	θ_c	p-value	N	θ_u	p-value	N
COREOS	-	-	0	0.154	0.485	41	-	-	0	0.154	0.49	41
CPRFPCCHS	-	-	0	0.152	0.849	643	-	-	0	0.152	0.866	643
E2FANTRNR	-	-	0	0.088	0.997	19	-	-	0	0.088	0.998	19
E2FBNTRNR	0.203	0.0	460	0.153	0.996	9317	0.210	0.0	345	0.154	0.993	9432
ELRE2PCPAL1	-	-	0	0.151	0.884	817	-	-	0	0.151	0.889	817
EMHVCHORD	0.150	0.798	232	0.163	0.116	799	0.141	0.939	190	0.164	0.069	841
EREGCC	-	-	0	0.153	0.707	288	-	-	0	0.153	0.718	288
AAGACGTAGATACL12	-	-	0	0.166	0.214	160	-	-	0	0.166	0.21	160
ERELEE4	0.100	0.974	19	0.148	0.505	9	0.101	0.98	18	0.145	0.56	10
GAGAGMGSA1	-	-	0	0.114	0.932	20	-	-	0	0.114	0.937	20
GBOXLERBCS	0.160	0.386	114	0.151	0.92	1239	0.163	0.311	100	0.151	0.947	1253
GCCCORE	0.147	0.792	115	0.131	1.0	292	0.144	0.825	98	0.131	1.0	309
HDZIPIIIAT	-	-	0	0.133	0.884	52	-	-	0	0.133	0.909	52
HSRENTHSR203J	-	-	0	0.158	0.452	52	-	-	0	0.158	0.472	52
HY5AT	-	-	0	0.143	0.921	181	-	-	0	0.143	0.91	181
IBOXCORE	0.138	0.942	105	0.127	0.921	35	0.138	0.937	105	0.127	0.899	35
MRNA3ENDTAH3	-	-	0	0.136	0.771	29	-	-	0	0.136	0.765	29
MYB1LEPR	0.135	0.658	13	0.166	0.353	13	0.141	0.609	12	0.166	0.341	14
MYB26PS	0.151	0.381	3	0.102	0.924	12	0.151	0.408	3	0.102	0.931	12
ABRE3OSRAB16	-	-	0	0.161	0.283	396	-	-	0	0.161	0.295	396
ABREA2HVA1	-	-	0	0.171	0.255	38	-	-	0	0.171	0.251	38
NTBBF1ARROLB	0.131	1.0	267	0.181	0.029	151	0.133	0.999	254	0.172	0.09	164
O2F1BE2S1	-	-	0	0.145	0.829	120	-	-	0	0.145	0.81	120

Continues on next page

Binding site	Fabids						Rosids					
	conserved			nonconserved			conserved			nonconserved		
	θ_c	p-value	N	θ_u	p-value	N	θ_c	p-value	N	θ_u	p-value	N
O2F2BE2S1	-	-	0	0.150	0.594	49	-	-	0	0.150	0.599	49
O2F3BE2S1	-	-	0	0.175	0.228	31	-	-	0	0.175	0.225	31
PE2FNTRNR1A	-	-	0	0.124	0.747	10	-	-	0	0.124	0.736	10
PIATGAPB	0.212	0.0	122	0.157	0.525	2554	0.222	0.001	90	0.157	0.47	2586
ABREBNNAPA	-	-	0	0.134	0.841	35	-	-	0	0.134	0.841	35
QELEMENTZMZM13	0.149	0.716	126	0.183	0.033	122	0.146	0.794	116	0.183	0.045	132
RAV1AAT	0.191	0.08	38	0.223	0.068	15	0.191	0.09	38	0.223	0.064	15
RBCSBOX3PS	-	-	0	0.160	0.22	1182	-	-	0	0.160	0.225	1182
RYREPEAT4	-	-	0	0.131	0.973	96	-	-	0	0.131	0.968	96
RYREPEATLEGUMINBOX	0.122	0.719	8	0.164	0.404	4	0.117	0.66	5	0.166	0.365	7
RYREPEATVFLEB4	0.306	0.031	6	0.137	0.809	36	0.306	0.027	6	0.137	0.796	36
SITEIIAOSPCNA	-	-	0	0.134	0.992	191	-	-	0	0.134	0.995	191
SORLREP4AT	0.168	0.329	24	0.154	0.65	240	0.150	0.494	14	0.155	0.585	250
TATABOX1	-	-	0	0.147	0.522	7	-	-	0	0.147	0.485	7
TATAPVTRNALEU	0.151	0.681	144	0.172	0.097	181	0.154	0.573	137	0.169	0.132	188
TGA1ANTPR1A	-	-	0	0.158	0.46	2860	-	-	0	0.158	0.456	2860
UP1ATMSD	0.150	0.763	188	0.153	0.813	829	0.154	0.604	164	0.152	0.856	853
WRKY71OS	0.145	0.871	183	0.191	0.05	58	0.146	0.843	178	0.184	0.07	63
WUSATAg	0.142	1.0	843	0.160	0.267	1043	0.143	0.998	784	0.158	0.417	1102
XYLAT	0.142	0.903	158	0.180	0.019	240	0.148	0.822	148	0.174	0.039	250
ABREZMRAB28	0.180	0.183	30	0.161	0.236	575	0.180	0.18	30	0.161	0.25	575
ACGTSEED3	-	-	0	0.151	1.0	5871	-	-	0	0.151	1.0	5871
AGL2ATCONSENSUS	0.191	0.04	76	0.151	0.994	2582	0.190	0.04	66	0.151	0.99	2592

Continues on next page

Binding site	Fabids						Rosids					
	conserved			nonconserved			conserved			nonconserved		
	θ_c	p-value	N	θ_u	p-value	N	θ_c	p-value	N	θ_u	p-value	N
ANAERO3CONSENSUS	0.190	0.068	51	0.162	0.343	68	0.206	0.027	43	0.158	0.449	76
ANAERO5CONSENSUS	0.204	0.001	124	0.153	0.936	2954	0.221	0.001	62	0.154	0.918	3016
ARE1	0.092	0.719	4	0.144	0.743	62	0.109	0.564	3	0.145	0.749	63
ARECOREZMGAPC4	-	-	0	0.142	0.993	502	-	-	0	0.142	0.995	502
ARR1AT	0.142	0.738	39	0.229	0.059	14	0.142	0.756	39	0.229	0.056	14

Table 3.6: The coexpression values for the genes constituting the families where the motifs are conserved and non-conserved. Results for motifs that where found in less than five promoters in all groups not shown.

3.5 Motif combinations

Of the 10296 possible combinations of two motifs 6694 (65.0 %) were found in at least one rice promoter sequence. Table 3.7 summarizes the results of conserved CRE combinations. Due to limited computational resources only 9132 CRE combinations were considered.

Clade	Conserved	%
Monocots	2634	25.6
Oryza	6287	61.1

Table 3.7: The number of motif combinations that where conserved in at least one family, and the percentage of the total amount of combinations this constitutes.

		Significance	
Motif 1	Motif 2	Oryza	Monocots
CArG promoter motif	IBOXCORE	Nonconserved	Nonconserved
CCA1 motif1 BS in CAB1	PRECONSCRHSP70A	Nonconserved	-
CCA1 motif1 BS in CAB1	ANAERO5CONSENSUS	Conserved	Nonconserved
DREB1AND2 BS in rd29a	MYB3 binding site motif	-	Conserved
DREB1AND2 BS in rd29a	AGATCONSENSUS	-	Conserved
ERE promoter motif	Z-box promoter motif	-	Conserved
ERE promoter motif	RE1ASPHYA3	-	Conserved
ERE promoter motif	20NTNTNOS	-	Conserved
HSEs binding site motif	ANAERO5CONSENSUS	Conserved	Nonconserved
L1-box promoter motif	MA0129.1	Nonconserved	-
L1-box promoter motif	ANAERO3CONSENSUS	-	Nonconserved
L1-box promoter motif	ANAERO5CONSENSUS	-	Nonconserved
LS5 promoter motif	HDZIPIIIAT	Nonconserved	-
LS5 promoter motif	ANAERO5CONSENSUS	Both	Nonconserved
MYB3 binding site motif	ANAERO5CONSENSUS	Both	Nonconserved
SBP-box promoter motif	AACACOREOSGLUB1	Nonconserved	-
SBP-box promoter motif	ANAERO5CONSENSUS	Conserved	Nonconserved
T-box promoter motif	AG BS in SPL/NOZ	-	Conserved
Z-box promoter motif	PIATGAPB	-	Nonconserved
Z-box promoter motif	RBCSBOX2PS	Nonconserved	-
Z-box promoter motif	ANAERO5CONSENSUS	Both	Both
AG BS in SPL/NOZ	TGA1ANTPR1A	-	Conserved
AG BS in SPL/NOZ	XYLAT	Conserved	-
SORLIP5	SGBFGMGMAUX28	-	Nonconserved
SORLIP5	SORLIP4AT	-	Nonconserved
SORLIP5	ANAERO5CONSENSUS	Conserved	-
MA0096.1	AACACOREOSGLUB1	Conserved	-
MA0120.1	ANAERO5CONSENSUS	-	Nonconserved
CEREGLUBOX3PSLEGA	ANAERO5CONSENSUS	Both	Nonconserved
CPRFPCCHS	PRECONSCRHSP70A	Conserved	Nonconserved
CPRFPCCHS	ANAERO5CONSENSUS	Both	Nonconserved
E2FBNTRNR	ANAERO5CONSENSUS	Conserved	Nonconserved
GCBP2ZMGAPC4	ANAERO5CONSENSUS	Both	-
GLUTECOREOS	ANAERO5CONSENSUS	Both	Nonconserved
IBOXLSCMCUCUMISIN	ANAERO5CONSENSUS	Both	-
ABRE2HVA22	AGL2ATCONSENSUS	Nonconserved	-
ABRE2HVA22	ANAERO5CONSENSUS	Both	Nonconserved

Continues on next page

		Significance	
Motif 1	Motif 2	Oryza	Monocots
ABRE3HVA1	ANAERO5CONSENSUS	Conserved	Nonconserved
ABRE3OSRAB16	ANAERO5CONSENSUS	-	Nonconserved
OCETYPEIIINTHISTONE	ANAERO5CONSENSUS	Both	Nonconserved
OCSGMHSP26A	ANAERO5CONSENSUS	Both	Nonconserved
OSE2ROOTNODULE	PIATGAPB	Nonconserved	-
OSE2ROOTNODULE	TATAPVTRNALEU	-	Conserved
P1BS	ANAERO5CONSENSUS	Both	Nonconserved
PIATGAPB	SGBFGMGMAUX28	Nonconserved	-
PIATGAPB	AGATCONSENSUS	Nonconserved	-
PIATGAPB	ANAERO5CONSENSUS	Nonconserved	Nonconserved
PRECONSCRHSP70A	ANAERO5CONSENSUS	Conserved	Nonconserved
RAV1BAT	ANAERO5CONSENSUS	Both	Nonconserved
RBCSBOX2PS	ANAERO5CONSENSUS	-	Nonconserved
ABRECE1HVA22	ANAERO5CONSENSUS	-	Nonconserved
RHERPATEXPA7	ANAERO5CONSENSUS	Both	-
TATCCACHVAL21	ANAERO5CONSENSUS	Both	-
UP1ATMSD	ANAERO5CONSENSUS	Conserved	Nonconserved
ACGTSEED3	ANAERO5CONSENSUS	Both	Nonconserved
AGATCONSENSUS	ANAERO5CONSENSUS	Both	Nonconserved
ANAERO5CONSENSUS	ARECOREZMGAPC4	Conserved	Nonconserved
ANAERO5CONSENSUS	ATHB5ATCORE	Both	-

Table 3.8: Combinations of motifs that are significantly over-coexpressed, with a false discovery rate of 0.1. The two significance *columns indicates whether the motif was significantly ($p \leq 0.001$) over-coexpressed in the conserved or nonconserved only, or both, across the clades Oryza and monocots. Dashes indicate that the combination was not significantly over-coexpressed. The labeling of the two motifs in a combination is arbitrary and does not imply any ordering. Results for motifs that where found in less than five promoters in all groups not shown.*

3.6 Divergence in CREs between monocots and eudicots

34 CREs were found to be conserved in significantly many families among the monocots that were not found to be conserved in eudicots. These are listed in Table 3.9. No CREs were found to be conserved in significantly many families among the eudicots only.

Name
ABFs binding site motif
ATHB2 binding site motif
GCC-box promoter motif
HSEs binding site motif
MYB3 binding site motif
SBP-box promoter motif
TELO-box promoter motif
Z-box promoter motif
MA0096.1
MA0129.1
SORLIP5
BP5OSWX
DRE1COREZMRAB17
E2FAT
E2FBNTRNR
EMHVCHORD
EVENINGAT
GAGAGMGSA1
GBOXLERBCS
GCCCORE
HEXAMERATH4
L1BOXATPDF1
MYB26PS
RAV1BAT
SBOXATRBCS
SORLREP4AT
ABREMOTIFAOSOSEM
TATABOX1
TATCCACHVAL21
UP1ATMSD
AMMORESIVDCRNIA1
ANAERO3CONSENSUS
ARE1
ATHB5ATCORE

Table 3.9: List of motifs that where found to be conserved in significantly many families in monocots, but not in eudicots.

Chapter 4

Discussion

4.1 CRE prevalence and conservation

The predicted regulatory elements show a very high prevalence across all species, with more than 100 different *cis*-regulatory elements predicted in the promoter for each gene on average for all but two species. Furthermore, the average number of CREs found, including CREs found more than once, tend to be 4-5 times greater than when just counting up the unique instances. This suggests that it is not uncommon for the same CRE to be present more than once in a promoter.

The predictions should be considered with caution. For many CREs the experimental evidence come from only a single species and thus it's functionality in other species may be dubious, especially if they are distantly related. Furthermore, several of the CREs are similar to each other and may be functionally equivalent.

The length of the region upstream of the translation start site (2000 bp) referred to as the *promoter region* were longer than in many similar studies. This will most likely result in more spurious findings and false positives.

Ding et al. (2012b) found 66,530 instances of 317 *cis*-regulatory elements in *C. reinhardtii*. This is a much lower number than the 7.4 million instances found here, but these numbers should be compared with care. Ding et al. (2012b) used only 8,742 genes that had an ortholog in *V. carteri*, which is about half of the genes used here. In addition they only looked at the 1 kb upstream of the translation start site, half of the length used here. They also only identified CREs 8 bp long.

It is informative to take a closer look at some of the CREs that were identified as being conserved in significantly many families. This serves several

purposes, first as an attempt to assess whether the results are plausible given the already existing knowledge about their role in gene regulation and their prevalence in across species. Secondly it may provide new insights about regulatory evolution.

4.1.1 Important CREs across all plants

As seen in Table 3.2 only four CREs were conserved in significantly many families across all 25 species (*Green Plants*). The four CREs are ARR1AT, WRKY71OS, OSE2ROOTNODULE and RAV1AAT, all from PLACE. The small number of CREs makes it feasible to delve into the existing litterature.

The annotations for the ARR1AT CRE identifies the CRE as a binding site for the *A. thaliana* ARR1 transcription factor (Sakai et al., 2000), and other studies has found it in rice as well (Ross et al., 2004). ARR1 belongs to a group of regulators called B-type RR (for response regulator) involved in cytokinin signal transduction pathways (Ishida et al., 2008). This group of regulators is also well characterized in rice (Ito and Kurata, 2006) and comparative investigations has also found B-type RR's in moss and alge species (Pils and Heyl, 2009). The ARR1AT CRE were also found to be conserved in significantly many families in all other subgroups investigated here.

The WRKY71OS CRE is described in the PLACE annotations as the *core of TGAC-containing W-box*. The W-box sequence is bound bound by members of the WRKY family of transcription factors. The full W-box consensus motif is `ttTGACY`, and the WRKY71OS motif consists of the core sequence `TGAC`, which is essential for binding of WRKY proteins (Eulgem et al., 2000). Members of WRKY family have been found in great abundance in *A. thaliana* and rice, and is also found ferns, mosses and algae (Ülker and Somssich, 2004). WRKY proteins have also been found in slime molds and other protists, indicating that the family originated before the advent of plants. The proliferation of this TF family in plants has been attributed to its role in defense against plant pathogens, although a wide range of other functions are reported as well (Rushton et al., 2010). This motif was also conserved in significantly many families in all clades.

The OSE2ROOTNODULE CRE is according to the PLACE annotations found in promoters to leghemoglobin genes. Leghemoglobins play an important role in regulating oxygen in root cells infected by nitrogen fixating bacteria (Raven et al., 2005). The amino acid sequence of the *Vicia faba* leghemoglobin gene VfLb29 (Swiss-Prot: P93848.1) where the CRE were first described (Vieweg et al., 2004) gave significant BLAST results to members of the HOM001294 PLAZA family. This family had members in all

plants. OSE2ROOTNODULE is identical to the NODCON2GM motif, but this was not one of the motifs were tested for being conserved in significantly many families. The description of this motif also suggests similar functionality. OSE2ROOTNODULE was also conserved in significantly many families in all clades.

The RAV1AAT CRE were first characterized as bound by the RAV1 transcription factor in *A. thaliana* (Kagaya et al., 1999). RAV1 is an interesting case as it has two DNA binding domains, described as AP2-like and B3-like, that binds to two distinct sequence motifs. While RAV1AAT is conserved in significantly many families in all clades, the companion motif, RAV1BAT, is not conserved in significantly many families in Green Plants ($p = 1$), but only is in clades Monocots, Chlorophyta, Malvids and Oryza ($p = 0$ for all). The PLAZA gene family where the *A. thaliana* RAV1 gene belong to (HOM000447) has members in all 20 available species of land plants, but not in chlorophyta. In an explorative study of cis-regulatory elements in the green algae *C. reinhardtii* 378 instances of the two combined motifs was found (Ding et al., 2012b). At least two predicted *C. reinhardtii* proteins are annotated in RefSeq as containing AP2 an B3 domains (accessions XM_001693601.1 and XM_001696127.1), but a co-occurrence of the two domains in a single *C. reinhardtii* protein has not been described.

Of the four CREs the OSE2ROOTNODULE CRE is perhaps the result that should be considered with greatest care. Although leghemoglobins are found in a great range of species, there does not seem to be any evidence that suggests that they should be regulated by the same motif. It is however possible that this result is a new discovery of an important *cis*-regulatory element. For the three other CREs there are much evidence to credibly imply biological meaningful roles across the plant kingdom.

It should be noted that these four CREs are exceptionally short (4-5 bp). Only 6 of the 144 motifs investigated are 5 bp or shorter. The expected number of times a sequence of length 5 bp would occur by chance is about $0.25^5 \times 2000 = 1.95$ times in a 2000 bp sequence. This high expected incidence of chance finds could explain why such short sequence motifs are conserved in significantly many families, but the fact that two such sequence motifs were not found to be significant could mean that the significant sequence motifs are in fact real important *cis*-regulatory elements.

Biological constraints can provide an alternative explanation for why the sequence motifs are so short. Long *cis*-regulatory sequence motifs are very specific, but prone to mutations, while short sequence motifs are less specific but more resilient to mutations (Stewart et al., 2012). CREs conserved across a great range of species would therefore not be expected to be very long. The low binding specificty for transcription factors associated with short

binding sites is also not such a problem for the WRKY71OS and RAV1AAT motifs. WRKY71OS is only an extremely conserved part of a great range of longer binding sites and RAV1AAT is only one half of a binding site duo. Higher specificity for other short binding sites may also be provided by other synergistic processes such as cooperative binding.

4.2 Go terms

In many cases the genes where a *cis*-regulatory element was found were associated with significantly enriched GO terms. The genes where a CREs was conserved had more enriched terms than the genes where the CREs was nonconserved. This was found both when the CREs were conserved at the level of Oryza and monocots, although in a smaller degree when conservation in studied across the monocots.

If a greater number of enriched terms for a set of genes are taken as a sign of a better criterion for identifying a set of genes, then it seems as if taking conservation into account is a good approach. The lower ratio of enriched terms between the conserved and nonconserved CREs when conservation is considered across the monocots is interesting. Monocots are a rather diverse group of plants, and conservation across it may therefore be a considered a very strict criterion. The set of genes where the CREs are conserved did however yield more enriched terms than the nonconserved. It is unclear how much certainty should be put into the results concerning the difference between the conserved and nonconserved CREs, as no significance testing has been performed. But the results suggests that there is some functional differences between the two groups, both when considered across monocots and Oryza.

The enriched terms were generally different in the genes where the CREs were conserved and nonconserved, as indicated by the low V scores in Table 3.3. For conservation across monocots the lower 90% percentile is 0.36, while for Oryza it is 0.16.

The CREs in the genes where they are conserved differ a bit from the nonconserved also in terms of how specific the terms are. The conserved has an average V of 6.5 in monocots and 6.8 in Oryza. The nonconserved on the other hand has the respective averages 7.2 and 7.3.

In this analysis only single CREs are investigated. CREs are often not found alone, as can be seen in Table 3.7. This means that a single CRE conserved in Oryza have a greater chance of co-occurring with another conserved CRE than a CRE conserved across the monocots. A conserved duo of predicted CREs would be expected to have a greater chance of having a

functional role, and this should in turn lead to the pattern we see here with the loosest conservation criterion having most enriched terms.

The differences in specificity for the enriched GO terms is interesting. One would expect the conserved CREs, being more likely to have a function, had more specific roles. One could speculate that this can imply a recent divergence in the rice regulatory networks, but as the significance between the number are unclear, it is impossible to say. It may also be the result of bias in the annotations, as more specific terms are annotated also imply more knowledge of the associated gene.

4.3 Gene expression

For some CREs, the genes where it was predicted in the promoter appeared to show interesting expression patterns. In rice the genes where the L1-box promoter motif, MA0120.1, E2FBNTRNR, GCCCORE and ANAERO5CONSENSUS was not conserved show significant over-coexpression, while the genes where these CREs was conserved did not. MA0120.1 was significantly over-coexpressed only in the genes nonconserved across Oryza, but not across monocots. The oppsite was the case for the others, except for GCCCORE, which were found to be significantly over-coexpressed across both clades. Interestingly, none of the conserved CREs are found to be significantly over-coexpressed.

Given that the genes where a CRE is nonconserved across monocots is a superset the genes that the motif is nonconserved across Oryza, and *vice versa* for the genes where a CRE is conserved, the p-values are not independent. With this in mind, the small discrepancies between the two degrees of conservation should not be given much weight.

At both levels of conservation, the nonconserved CRE instances are generally associated with a significantly higher coexpression score than the conserved CREs.

All of these CREs are conserved in significantly many families across monocots and oryza, except for ANAERO5CONSENSUS, which is conserved in significantly many families only in oryza.

In populus the genes where the T-box promoter motif was not conserved showed significant over-coexpression, while the genes where the CRE was conserved did not. This holds across both the fabids and rosids clades. The E2FBNTRNR, PIATGAPB and ANAERO5CONSENSUS CREs showed significant over-coexpression in the genes where they were conserved, but not in the genes where they were not conserved. Again, this was the case across both the fabids and rosids clades.

There were in general no significant differences in the coexpression scores between the genes where the CREs were conserved and nonconserved, again across both the rosids and fabids level.

It is interesting to note that the nonconserved CREs in rice had a significantly greater coexpression score than the conserved, while no such difference was found in populus. This may suggest that rice has some novel changes in its regulatory networks. One could only speculate why this is, but one possible explanation could be that rice is an extremely important food plant that was domesticated about 9000 years ago, and has since then been subject to strong artificial selection, that has even resulted in the two distinct subspecies (Molina et al., 2011). While populus also is a cultivated plant, it has most likely not been subject to the same artificial selection pressures as rice.

Of the five CREs that showed any significant over-coexpression in rice, whether in the conserved or nonconserved gene sets, all but ANAERO5CONSENSUS are conserved in significantly many families across both monocots and Oryza. ANAERO5CONSENSUS was conserved in significantly many families only across Oryza.

Of the four CREs that showed any sort of significant over-coexpression in populus, all but the T-box promoter motif are not conserved in significantly many families, across fabids or rosids. The T-box promoter motif is conserved in significantly many families across both fabids and rosids.

Since the coexpression scores were not calculated for all the CREs that were investigated for conservation in significantly many families, the tests for differences in coexpression values between the conserved CREs and the CREs conserved in significantly many families are unreliable and it is wise to not speak of them.

Taken together, these results does not seem to correspond well with the results identifying the CREs that are conserved in significantly many families. Some caveats should be noted, however. For one, the gene expression score used here is not based well established measures such as the Expression Coherence Score (Pilpel et al., 2001).

It may also be that the one-sample significance test for over-coexpression used here is unsuitable for the main purpose of this investigation. Possibly a better approach would be to compare the coexpression of the conserved and nonconserved CRE instances directly as two independent samples. For rice the paired two-sample t-tests suggests that this could have yielded more interesting results. Eyeballing Table 3.5 it is not hard to find CREs that possibly could yield significant results. For example, the MYB1 binding site motif, EMHVCHORD and RYREPEATVFLEB4 CREs have rather large dif-

ferences in the coexpression scores between the conserved and nonconserved gene sets.

4.3.1 CRE combinations in rice

A couple of things are immediately clear when looking at the results in Table 3.8. The first is that a lot more CREs are found to be associated with significant over-coexpression when considered in pairs than when considered alone. In total 54 CREs is on that list, compared to 5 when only considered alone.

The second is that a some of the motifs appear several times. The ANAERO5CONSENSUS motif is by far the most frequent motif, with 31 instances on the list. In all combinations where it was significantly over-coexpressed this was only so in the genes where it was not conserved across the monocots. The opposite was true when conservation was considered across Oryza, where the the combinations were either significant only where it was conserved, or both conserved and nonconserved. This suggests that ANAERO5CONSENSUS have some important functionality in the Oryza lineage. The finding that ANAERO5CONSENSUS is conserved in significantly many families in the Oryza clade, but not in the monocot clade (see Appendix A) fits well with this result.

With the methodology used here, it is hard to say for sure whether the over-coexpressed CRE combinations are found here because of synergistic effects, or if only one of the CREs are sufficient and any over-coexpression found in combination is spurious. This may very well be the case with ANAERO5CONSENSUS, where the genes where it was associated with showed significant over-coexpression.

Of the 54 CREs listed in Table 3.8, 25 are conserved in significantly many families across both Oryza and the monocots, while six are neither. 23 CREs are conserved in significantly many families across Oryza, but not across the monocots. None are conserved in significantly many families across the monocts only.

4.4 Monocot-eudicot divergence

It is not feasible to discuss the functionality of every single CRE that where found to be conserved in significantly many families in monocots, but not in eudicots, in depth. It is nevertheless appropriate to do a cursory look at the annotations from the source databases in addition to compare it to the results from the GO enrichment and gene expression analyses.

Eight of the motifs are from AGRIS, meaning that they can be assumed to be functional in the eudicot *A. thaliana*. For the two motifs from JASPAR, MA0096.1 and MA0129.1 are annotated as comming from the eudicots Snappdragon (*Antirrhinum majus*) and *Nicotina sp.*, respectively.

In light of the discussion in Chapter 4.1.1, one notable motif presence in table 3.9 is RAV1BAT, while RAV1AAT is is absent. One would expect both motifs to not having diverged, at least not just one of them. Interestingly, as seen in Appendix A, RAV1BAT is only conserved in significantly many families in the malvids subclade of eudicots, which is where we find *A. thaliana*, where the motif duo where first identified.

The E2FBNTRNR motif is according to the PLACE annotations first identified in the eudicot tobacco (*Nicotiana tabacum*). In the gene expression analysis this motif had significant over-coexpression in the where it was unconserved in rice, as well as significantly over-coexpressed where it was conserved in poplar.

GCCCORE is according to the PLACE annotations found in the eudicots *A. thaliana* and tomato (*Lycopersicon esculentum*). It was also found to be significantly over-coexpressed in the genes where it was not conserved in rice.

For the rest of the CREs, 12 had annotations indicating they were discovered in eudicot species, while four indicated they were discovered in monocot species. One were discovered in both monocots and eudicots, and another one was discovered in Chlorophyta. One also had indeterminate annotations.

4.5 Conclusion

Using high prevalence conservation of *cis*-regulatory elements as a criterion for identifying important CREs throughout the natural history of plants does not seem to be a useful method. Although CREs that in some cases are known to have deep evolutionary roots where found to be conserved in significantly many families across all green plants, investigating divergence of CREs in two important lineages gives inconsistent results.

The high prevalence conserved CREs does not seem to give more insights than the ordinary conservation criterion when used in analyzing gene expression and annotational data. When CREs are considered in pairs, some more agreement seems to appear, but it is unclear how much weight these results should be given.

Bibliography

K. L. Adams and J. F. Wendel. Polyploidy and genome evolution in plants. *Current opinion in plant biology*, 8(2):135–141, Apr. 2005. ISSN 1369-5266. doi: 10.1016/j.pbi.2005.01.001. PMID: 15752992.

A. Alexa and J. Rahnenfuhrer. *topGO: topGO: Enrichment analysis for Gene Ontology*, 2010. R package version 2.14.0.

A. Alexa, J. Rahnenführer, and T. Lengauer. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, 22(13):1600–1607, Jan. 2006. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/btl140. PMID: 16606683.

S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, Oct. 1990. ISSN 0022-2836. doi: 10.1016/S0022-2836(05)80360-2. PMID: 2231712.

M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, May 2000. ISSN 1061-4036. doi: 10.1038/75556. PMID: 10802651 PMCID: PMC3037419.

T. L. Bailey and C. Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, 2:28–36, 1994. ISSN 1553-0833. PMID: 7584402.

M. V. Bel, S. Proost, E. Wischnitzki, S. Movahedi, C. Scheerlinck, Y. V. d. Peer, and K. Vandepoele. Dissecting plant genomes with the PLAZA comparative genomics platform. *Plant Physiology*, 158(2):590–600, Feb.

2012. ISSN 0032-0889, 1532-2548. doi: 10.1104/pp.111.189514. PMID: 22198273.

M. Carlson. *GO.db: A set of annotation maps describing the entire Gene Ontology*, 2014.

S. B. Carroll. Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. *Cell*, 134(1):25–36, July 2008. ISSN 1097-4172. doi: 10.1016/j.cell.2008.06.030. PMID: 18614008.

D. R. Cavener. Comparison of the consensus sequence flanking translational start sites in drosophila and vertebrates. *Nucleic acids research*, 15(4): 1353–1361, Feb. 1987. ISSN 0305-1048. PMID: 3822832.

W.-C. Chang, T.-Y. Lee, H.-D. Huang, H.-Y. Huang, and R.-L. Pan. PlantPAN: plant promoter analysis navigator, for identifying combinatorial cis-regulatory elements with distance constraint in plant gene groups. *BMC Genomics*, 9:561, Nov. 2008. ISSN 1471-2164. doi: 10.1186/1471-2164-9-561. PMID: 19036138 PMCID: PMC2633311.

A. Christ, I. Maegele, N. Ha, H. H. Nguyen, M. D. Crespi, and A. Maizel. In silico identification and in vivo validation of a set of evolutionary conserved plant root-specific cis-regulatory elements. *Mechanisms of development*, 130(1):70–81, Jan. 2013. ISSN 1872-6356. doi: 10.1016/j.mod.2012.03.002. PMID: 22504372.

H. Y. Chu, E. Wegel, and A. Osbourn. From hormones to secondary metabolism: the emergence of metabolic gene clusters in plants. *The Plant Journal*, 66(1):66–79, 2011. ISSN 1365-313X. doi: 10.1111/j.1365-313X.2011.04503.x.

A. Dabney, J. D. Storey, and with assistance from Gregory R. Warnes. *qvalue: Q-value estimation for false discovery rate control*. R package version 1.36.0.

M. K. Das and H.-K. Dai. A survey of DNA motif finding algorithms. *BMC Bioinformatics*, 8(Suppl 7):S21, Nov. 2007. ISSN 1471-2105. doi: 10.1186/1471-2105-8-S7-S21. PMID: 18047721.

W. H. Day and F. R. McMorris. Critical comparison of consensus methods for molecular sequences. *Nucleic Acids Research*, 20(5):1093–1099, Mar. 1992. ISSN 0305-1048. PMID: 1549472 PMCID: PMC312096.

J. Ding, H. Hu, and X. Li. Thousands of cis-regulatory sequence combinations are shared by arabidopsis and poplar. *Plant Physiology*, 158(1):145–155, Jan. 2012a. ISSN 0032-0889, 1532-2548. doi: 10.1104/pp.111.186080. PMID: 22058225.

J. Ding, X. Li, and H. Hu. Systematic prediction of cis-regulatory elements in the chlamydomonas reinhardtii genome using comparative genomics. *Plant Physiology*, 160(2):613–623, Oct. 2012b. ISSN 0032-0889, 1532-2548. doi: 10.1104/pp.112.200840. PMID: 22915576.

J. M. Duarte, L. Cui, P. K. Wall, Q. Zhang, X. Zhang, J. Leebens-Mack, H. Ma, N. Altman, and C. W. dePamphilis. Expression pattern shifts following duplication indicative of subfunctionalization and neofunctionalization in regulatory genes of arabidopsis. *Molecular Biology and Evolution*, 23(2):469–478, Jan. 2006. ISSN 0737-4038, 1537-1719. doi: 10.1093/molbev/msj051. PMID: 16280546.

A. J. Enright, S. Van Dongen, and C. A. Ouzounis. An efficient algorithm for large-scale detection of protein families. *Nucleic acids research*, 30 (7):1575–1584, Apr. 2002. ISSN 1362-4962. PMID: 11917018 PMCID: PMC101833.

T. Eulgem, P. J. Rushton, S. Robatzek, and I. E. Somssich. The WRKY superfamily of plant transcription factors. *Trends in Plant Science*, 5(5): 199–206, May 2000. ISSN 1360-1385. doi: 10.1016/S1360-1385(00)01600-9.

W. J. Ewens and G. R. Grant. *Statistical Methods in Bioinformatics*. Springer, 2005.

L. E. Flagel and J. F. Wendel. Gene duplication and evolutionary novelty in plants. *New Phytologist*, 183(3):557–564, 2009. ISSN 1469-8137. doi: 10.1111/j.1469-8137.2009.02923.x.

A. Force, M. Lynch, F. B. Pickett, A. Amores, Y. L. Yan, and J. Postlethwait. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics*, 151(4):1531–1545, Apr. 1999. ISSN 0016-6731. PMID: 10101175 PMCID: PMC1460548.

K. Higo, Y. Ugawa, M. Iwamoto, and T. Korenaga. Plant cis-acting regulatory DNA elements (PLACE) database: 1999. *Nucleic acids research*, 27(1):297–300, Jan. 1999. ISSN 0305-1048. PMID: 9847208 PMCID: PMC148163.

D. W. Huang, B. T. Sherman, and R. A. Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1):1–13, Jan. 2009. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkn923. PMID: 19033363.

A. L. Hughes. The evolution of functionally novel proteins after gene duplication. *Proceedings. Biological sciences / The Royal Society*, 256(1346): 119–124, May 1994. ISSN 0962-8452. doi: 10.1098/rspb.1994.0058. PMID: 8029240.

O. Ibraheem, C. E. J. Botha, and G. Bradley. In silico analysis of cis-acting regulatory elements in 5 regulatory regions of sucrose transporter gene families in rice (oryza sativa japonica) and arabidopsis thaliana. *Computational Biology and Chemistry*, 34(5–6):268–283, Dec. 2010. ISSN 1476-9271. doi: 10.1016/j.compbiolchem.2010.09.003. URL `http://www.sciencedirect.com/science/article/pii/S1476927110000721`.

K. Ishida, T. Yamashino, A. Yokoyama, and T. Mizuno. Three type-b response regulators, ARR1, ARR10 and ARR12, play essential but redundant roles in cytokinin signal transduction throughout the life cycle of arabidopsis thaliana. *Plant & cell physiology*, 49(1):47–57, Jan. 2008. ISSN 0032-0781. doi: 10.1093/pcp/pcm165. PMID: 18037673.

Y. Ito and N. Kurata. Identification and characterization of cytokinin-signalling gene families in rice. *Gene*, 382:57–65, Nov. 2006. ISSN 0378-1119. doi: 10.1016/j.gene.2006.06.020.

Y. Kagaya, K. Ohmiya, and T. Hattori. RAV1, a novel DNA-binding protein, binds to bipartite recognition sequence through two distinct DNA-binding domains uniquely found in higher plants. *Nucleic Acids Research*, 27(2):470–478, Jan. 1999. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/27.2.470. PMID: 9862967.

K. Lindblad-Toh, M. Garber, O. Zuk, M. F. Lin, B. J. Parker, S. Washietl, P. Kheradpour, J. Ernst, G. Jordan, E. Mauceli, L. D. Ward, C. B. Lowe, A. K. Holloway, M. Clamp, S. Gnerre, J. Alföldi, K. Beal, J. Chang, H. Clawson, J. Cuff, F. Di Palma, S. Fitzgerald, P. Flicek, M. Guttman, M. J. Hubisz, D. B. Jaffe, I. Jungreis, W. J. Kent, D. Kostka, M. Lara, A. L. Martins, T. Massingham, I. Moltke, B. J. Raney, M. D. Rasmussen, J. Robinson, A. Stark, A. J. Vilella, J. Wen, X. Xie, M. C. Zody, Broad Institute Sequencing Platform and Whole Genome Assembly Team, K. C. Worley, C. L. Kovar, D. M. Muzny, R. A. Gibbs, Baylor College of Medicine Human Genome Sequencing Center Sequencing Team, W. C. Warren, E. R.

Mardis, G. M. Weinstock, R. K. Wilson, Genome Institute at Washington University, E. Birney, E. H. Margulies, J. Herrero, E. D. Green, D. Haussler, A. Siepel, N. Goldman, K. S. Pollard, J. S. Pedersen, E. S. Lander, and M. Kellis. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, 478(7370):476–482, Oct. 2011. ISSN 0028-0836. doi: 10.1038/nature10530.

M. Lynch and J. S. Conery. The evolutionary fate and consequences of duplicate genes. *Science*, 290(5494):1151–1155, Oct. 2000. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.290.5494.1151. PMID: 11073452.

M. Maienschein-Cline, A. R. Dinner, W. S. Hlavacek, and F. Mu. Improved predictions of transcription factor binding sites using physicochemical features of DNA. *Nucleic Acids Research*, 40(22):e175–e175, Jan. 2012. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gks771. PMID: 22923524.

A. Mathelier, X. Zhao, A. W. Zhang, F. Parcy, R. Worsley-Hunt, D. J. Arenillas, S. Buchman, C.-y. Chen, A. Chou, H. Ienasescu, J. Lim, C. Shyr, G. Tan, M. Zhou, B. Lenhard, A. Sandelin, and W. W. Wasserman. JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Research*, page gkt997, Nov. 2013. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkt997. PMID: 24194598.

V. Matys, E. Fricke, R. Geffers, E. Gössling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis, D.-U. Kloos, S. Land, B. Lewicki-Potapov, H. Michael, R. Münch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele, and E. Wingender. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic acids research*, 31(1):374–378, Jan. 2003. ISSN 1362-4962. PMID: 12520026 PMCID: PMC165555.

A. C. A. Meireles-Filho and A. Stark. Comparative genomics of gene regulation-conservation and divergence of cis-regulatory information. *Current opinion in genetics & development*, 19(6):565–570, Dec. 2009. ISSN 1879-0380. doi: 10.1016/j.gde.2009.10.006. PMID: 19913403.

J. Molina, M. Sikora, N. Garud, J. M. Flowers, S. Rubinstein, A. Reynolds, P. Huang, S. Jackson, B. A. Schaal, C. D. Bustamante, A. R. Boyko, and M. D. Purugganan. Molecular evidence for a single evolutionary origin of domesticated rice. *Proceedings of the National Academy of Sciences*, 108(20):8351–8356, May 2011. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1104686108. PMID: 21536870.

S. Netotea, D. Sundell, N. R. Street, and T. R. Hvidsten. ComPlEx: conservation and divergence of co-expression networks in a. thaliana, populus and o. sativa. *BMC Genomics*, 15(1):106, Feb. 2014. ISSN 1471-2164. doi: 10.1186/1471-2164-15-106. PMID: 24498971.

Y. Pilpel, P. Sudarsanam, and G. M. Church. Identifying regulatory networks by combinatorial analysis of promoter elements. *Nature Genetics*, 29(2): 153–159, Oct. 2001. ISSN 1061-4036. doi: 10.1038/ng724.

B. Pils and A. Heyl. Unraveling the evolution of cytokinin signaling. *Plant physiology*, 151(2):782–791, Oct. 2009. ISSN 0032-0889. doi: 10.1104/pp.109.139188. PMID: 19675156 PMCID: PMC2754637.

P. H. Raven, R. F. Evert, and S. E. Eichhorn. *Biology of Plants*. W. H. Freeman, Jan. 2005. ISBN 9780716710073.

R. Rohs, S. M. West, P. Liu, and B. Honig. Nuance in the double-helix and its role in protein-DNA recognition. *Current opinion in structural biology*, 19 (2):171–177, Apr. 2009. ISSN 0959-440X. doi: 10.1016/j.sbi.2009.03.002. PMID: 19362815 PMCID: PMC2701566.

R. Rohs, X. Jin, S. M. West, R. Joshi, B. Honig, and R. S. Mann. Origins of specificity in protein-DNA recognition. *Annual Review of Biochemistry*, 79:233–269, 2010. ISSN 0066-4154. doi: 10.1146/annurev-biochem-060408-091030. PMID: 20334529 PMCID: PMC3285485.

E. J. H. Ross, J. M. Stone, C. G. Elowsky, R. Arredondo-Peter, R. V. Klucas, and G. Sarath. Activation of the oryza sativa non-symbiotic haemoglobin-2 promoter by the cytokinin-regulated transcription factor, ARR1. *Journal of Experimental Botany*, 55(403):1721–1731, Jan. 2004. ISSN 0022-0957, 1460-2431. doi: 10.1093/jxb/erh211. PMID: 15258171.

P. J. Rushton, I. E. Somssich, P. Ringler, and Q. J. Shen. WRKY transcription factors. *Trends in Plant Science*, 15(5):247–258, May 2010. ISSN 1360-1385. doi: 10.1016/j.tplants.2010.02.006.

H. Sakai, T. Aoyama, and A. Oka. Arabidopsis ARR1 and ARR2 response regulators operate as transcriptional activators. *The Plant journal: for cell and molecular biology*, 24(6):703–711, Dec. 2000. ISSN 0960-7412. PMID: 11135105.

S.-H. Shiu, M.-C. Shih, and W.-H. Li. Transcription factor families have much higher expansion rates in plants than in animals. *Plant Physiology*, 139(1):

18–26, Jan. 2005. ISSN 0032-0889, 1532-2548. doi: 10.1104/pp.105.065110. PMID: 16166257.

N. O. Steffens, C. Galuschka, M. Schindler, L. Bülow, and R. Hehl. AthaMap: an online resource for in silico transcription factor binding sites in the arabidopsis thaliana genome. *Nucleic Acids Research*, 32(suppl 1):D368–D372, Jan. 2004. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkh017. PMID: 14681436.

N. O. Steffens, C. Galuschka, M. Schindler, L. Bülow, and R. Hehl. Athamap website documentation, February 2014. URL `http://www.athamap.de/documentation`$_m$*atrix*$_b$*ased.php*.

A. J. Stewart, S. Hannenhalli, and J. B. Plotkin. Why transcription factor binding sites are ten nucleotides long. *Genetics*, 192(3):973–985, Nov. 2012. ISSN 0016-6731, 1943-2631. doi: 10.1534/genetics.112.143370. PMID: 22887818.

G. D. Stormo. DNA binding sites: representation and discovery. *Bioinformatics*, 16(1):16–23, Jan. 2000. ISSN 1367-4803, 1460-2059. doi: 10.1093/bioinformatics/16.1.16. PMID: 10812473.

M. F. Vieweg, M. Frühling, H.-J. Quandt, U. Heim, H. Bäumlein, A. Pühler, H. Küster, and M. P. Andreas. The promoter of the vicia faba l. leghemoglobin gene VfLb29 is specifically activated in the infected cells of root nodules and in the arbuscule-containing cells of mycorrhizal roots from different legume and nonlegume plants. *Molecular plant-microbe interactions: MPMI*, 17(1):62–69, Jan. 2004. ISSN 0894-0282. doi: 10.1094/MPMI.2004.17.1.62. PMID: 14714869.

D. Wagner, R. W. Sablowski, and E. M. Meyerowitz. Transcriptional activation of APETALA1 by LEAFY. *Science (New York, N.Y.)*, 285(5427): 582–584, July 1999. ISSN 0036-8075. PMID: 10417387.

L.-L. Wang, H.-M. Liang, J.-L. Pang, and M.-Y. Zhu. Regulation network and biological roles of LEAFY in arabidopsis thaliana in floral development. *Yi chuan = Hereditas / Zhongguo yi chuan xue hui bian ji*, 26(1): 137–142, Jan. 2004. ISSN 0253-9772. PMID: 15626683.

J. Watson, T. Baker, and S. Bell. *Molecular Biology of the Gene*. Benjamin Cummings, 2011.

E. W. Weisstein. Bonferroni correction – from wolfram MathWorld, 2014. URL `http://mathworld.wolfram.com/BonferroniCorrection.html`.

The Bonferroni correction is a multiple-comparison correction used when several dependent or independent statistical tests are being performed simultaneously (since while a given alpha value alpha may be appropriate for each individual comparison, it is not for the set of all comparisons). In order to avoid a lot of spurious positives, the alpha value needs to be lowered to account for the number of comparisons being performed. The simplest and most conservative approach is the Bonferroni...

F. Xu, M.-R. Park, A. Kitazumi, V. Herath, B. Mohanty, S. J. Yun, and B. G. d. l. Reyes. Cis-regulatory signatures of orthologous stress-associated bZIP transcription factors from rice, sorghum and arabidopsis based on phylogenetic footprints. *BMC Genomics*, 13(1):497, Sept. 2012. ISSN 1471-2164. doi: 10.1186/1471-2164-13-497. PMID: 22992304.

N. Yamaguchi, A. Yamaguchi, M. Abe, D. Wagner, and Y. Komeda. LEAFY controls arabidopsis pedicel length and orientation by affecting adaxial-abaxial cell fate. *The Plant journal: for cell and molecular biology*, 69(5):844–856, Mar. 2012. ISSN 1365-313X. doi: 10.1111/j.1365-313X.2011.04836.x. PMID: 22050454.

L. Yang, T. Zhou, I. Dror, A. Mathelier, W. W. Wasserman, R. Gordân, and R. Rohs. TFBSshape: a motif database for DNA shape features of transcription factor binding sites. *Nucleic Acids Research*, page gkt1087, Nov. 2013. ISSN 0305-1048, 1362-4962. doi: 10.1093/nar/gkt1087. PMID: 24214955.

A. Yilmaz, M. K. Mejia-Guerra, K. Kurz, X. Liang, L. Welch, and E. Grotewold. AGRIS: the arabidopsis gene regulatory information server, an update. *Nucleic acids research*, 39(Database issue):D1118–1122, Jan. 2011. ISSN 1362-4962. doi: 10.1093/nar/gkq1120. PMID: 21059685 PMCID: PMC3013708.

J. O. Yáñez-Cuna, E. Z. Kvon, and A. Stark. Deciphering the transcriptional cis-regulatory code. *Trends in Genetics*, 29(1):11–22, Jan. 2013. ISSN 0168-9525. doi: 10.1016/j.tig.2012.09.007.

F. Zambelli, G. Pesole, and G. Pavesi. Motif discovery and transcription factor binding sites before and after the next-generation sequencing era. *Briefings in Bioinformatics*, 14(2):225–237, Jan. 2013. ISSN 1467-5463, 1477-4054. doi: 10.1093/bib/bbs016. PMID: 22517426.

B. Ülker and I. E. Somssich. WRKY transcription factors: from DNA binding towards biological function. *Current Opinion in Plant Biology*, 7(5):491–498, Oct. 2004. ISSN 1369-5266. doi: 10.1016/j.pbi.2004.07.012.

Appendices

Appendix A

Motif conservation: All results

Motif	Eudicots	Monocots	Chlorophyta	Rosids	Fabids	Malvids	Angiosperm	GreenPlants	Oryza
ABFs binding site motif	21 (0.021)	110 (0.000)	10 (0.998)	22 (0.037)	28 (0.057)	98 (0.000)	15 (0.293)	1 (1.000)	1233 (0.000)
ATHB2 binding site motif	8 (0.587)	21 (0.000)	0 (1.000)	8 (0.754)	17 (0.312)	22 (0.986)	5 (0.320)	0 (1.000)	347 (0.000)
CArG promoter motif	228 (0.000)	453 (0.000)	1 (1.000)	232 (0.001)	329 (0.002)	537 (0.000)	174 (0.002)	1 (1.000)	2986 (0.000)
CArG2 motif in AP3	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)
CCA1 motif1 BS in CAB1	2 (0.399)	3 (0.652)	0 (1.000)	2 (0.475)	4 (0.317)	10 (0.401)	2 (0.092)	0 (1.000)	237 (0.000)
DREB1AND2 BS in rd29a	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	1 (0.558)	0 (1.000)	0 (1.000)	129 (0.000)
EIL1 BS in ERF1	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)
EIL2 BS in ERF1	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)
ACE promoter motif	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	35 (0.000)
ERE promoter motif	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	1 (0.008)
GCC-box promoter motif	142 (0.931)	2799 (0.000)	1810 (0.000)	156 (0.947)	202 (1.000)	451 (0.000)	140 (0.962)	66 (1.000)	8484 (0.000)
HSEs binding site motif	66 (0.262)	97 (0.000)	0 (1.000)	70 (0.302)	92 (0.317)	192 (0.000)	43 (0.352)	0 (1.000)	1122 (0.000)
L1-box promoter motif	351 (0.000)	470 (0.000)	2 (0.952)	378 (0.000)	519 (0.000)	863 (0.000)	225 (0.000)	1 (0.992)	3272 (0.000)
LS5 promoter motif	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	39 (0.000)
MYB1 binding site motif	185 (0.001)	566 (0.000)	38 (1.000)	193 (0.003)	272 (0.015)	400 (0.000)	158 (0.021)	19 (1.000)	3573 (0.000)
MYB2 binding site motif	10 (0.101)	17 (0.075)	0 (1.000)	11 (0.237)	14 (0.655)	42 (0.003)	5 (0.357)	0 (1.000)	431 (0.000)
MYB3 binding site motif	59 (0.408)	109 (0.000)	0 (1.000)	71 (0.449)	101 (0.446)	158 (0.146)	35 (0.920)	0 (1.000)	1163 (0.000)
Nonamer promoter motif	0 (1.000)	14 (0.021)	1 (0.678)	0 (1.000)	0 (1.000)	3 (0.080)	0 (1.000)	0 (1.000)	326 (0.000)
AG BS in SUP	2 (0.160)	4 (0.192)	0 (1.000)	2 (0.188)	3 (0.265)	9 (0.143)	0 (1.000)	0 (1.000)	180 (0.000)
PII promoter motif	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)
SBP-box promoter motif	16 (0.011)	72 (0.000)	3 (0.920)	18 (0.016)	28 (0.005)	55 (0.000)	15 (0.004)	0 (1.000)	818 (0.000)
T-box promoter motif	1632 (0.000)	2798 (0.000)	363 (0.001)	1724 (0.000)	2145 (0.000)	3605 (0.000)	1272 (0.000)	185 (1.000)	9094 (0.000)
TELO-box promoter motif	88 (0.012)	82 (0.000)	0 (1.000)	93 (0.005)	126 (0.000)	295 (0.000)	47 (0.016)	0 (1.000)	1072 (0.000)
Z-box promoter motif	14 (0.265)	63 (0.000)	1 (0.985)	17 (0.361)	29 (0.019)	54 (0.407)	12 (0.202)	0 (1.000)	886 (0.000)
AG BS in SPL/NOZ	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)
AGL2 binding site motif	4 (0.621)	5 (0.550)	0 (1.000)	4 (0.708)	6 (0.733)	21 (0.521)	2 (0.476)	0 (1.000)	302 (0.000)
SORLREP1	4 (0.509)	18 (0.025)	0 (1.000)	4 (0.736)	13 (0.038)	14 (0.888)	4 (0.181)	0 (1.000)	490 (0.000)
SORLIP3	4 (0.258)	12 (0.006)	0 (1.000)	4 (0.374)	10 (0.151)	13 (0.059)	3 (0.144)	0 (1.000)	298 (0.000)
SORLIP5	190 (0.187)	593 (0.000)	38 (1.000)	209 (0.108)	298 (0.178)	449 (0.000)	163 (0.214)	23 (1.000)	3131 (0.000)
MA0127.1	127 (0.000)	95 (0.000)	0 (1.000)	132 (0.000)	174 (0.065)	282 (0.000)	59 (0.239)	0 (1.000)	1237 (0.000)
MA0096.1	147 (0.024)	638 (0.000)	224 (0.953)	162 (0.125)	226 (0.014)	438 (0.000)	137 (0.024)	41 (1.000)	3812 (0.000)
MA0120.1	216 (0.000)	806 (0.000)	59 (1.000)	230 (0.000)	291 (0.004)	712 (0.000)	199 (0.000)	29 (1.000)	4093 (0.000)
MA0129.1	277 (0.020)	1030 (0.000)	409 (0.000)	311 (0.102)	407 (0.350)	833 (0.000)	251 (0.021)	67 (1.000)	5141 (0.000)
MA0001.1	17 (0.140)	40 (0.005)	0 (1.000)	18 (0.172)	30 (0.062)	46 (0.230)	14 (0.025)	0 (1.000)	570 (0.000)
-10PEHVPSBD	2089 (0.000)	2824 (0.000)	84 (1.000)	2189 (0.000)	2702 (0.000)	4191 (0.000)	1467 (0.000)	64 (1.000)	9169 (0.000)
BOX1PSGS2	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	4 (0.049)	0 (1.000)	0 (1.000)	71 (0.000)
BOX1PVCHS15	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	4 (0.000)
BP5OSWX	66 (0.089)	345 (0.000)	47 (1.000)	71 (0.294)	97 (0.258)	182 (0.001)	61 (0.172)	12 (1.000)	2278 (0.000)
CAREOSREP1	1165 (0.000)	2397 (0.000)	527 (0.000)	1240 (0.000)	1608 (0.000)	2540 (0.000)	976 (0.000)	182 (1.000)	8190 (0.000)

Continues on next page

Motif	Eudicots	Monocots	Chlorophyta	Rosids	Fabids	Malvids	Angiosperm	GreenPlants	Oryza
CARG3ATAP3	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)
CATATGGMSAUR	705 (0.000)	1586 (0.000)	18 (1.000)	742 (0.000)	983 (0.000)	1579 (0.000)	583 (0.000)	13 (1.000)	6961 (0.000)
CEREGLUBOX3PSLEGA	24 (0.093)	25 (0.059)	0 (1.000)	27 (0.100)	37 (0.252)	63 (0.220)	14 (0.018)	0 (1.000)	650 (0.000)
CGACGOSAMY3	554 (0.000)	3881 (0.000)	2131 (0.000)	603 (0.036)	733 (0.756)	1740 (0.000)	544 (0.000)	255 (1.000)	10198 (0.000)
COREOS	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	17 (0.000)
CPRFPCCHS	0 (1.000)	2 (0.156)	0 (1.000)	0 (1.000)	0 (1.000)	1 (0.083)	0 (1.000)	0 (1.000)	168 (0.000)
DRE1COREZMRAB17	52 (0.992)	362 (0.000)	37 (1.000)	57 (1.000)	75 (0.995)	178 (0.997)	49 (0.994)	12 (1.000)	2236 (0.000)
AACACOREOSGLUB1	827 (0.000)	1489 (0.000)	43 (1.000)	886 (0.000)	1151 (0.000)	1929 (0.000)	640 (0.000)	33 (1.000)	5987 (0.000)
E2FANTRNR	7 (0.257)	44 (0.003)	15 (1.000)	7 (0.470)	7 (0.927)	31 (0.191)	7 (0.098)	0 (1.000)	632 (0.000)
E2FAT	2 (0.156)	38 (0.000)	6 (1.000)	3 (0.036)	3 (0.173)	18 (0.001)	2 (0.138)	0 (1.000)	634 (0.000)
E2FBNTRNR	5 (0.110)	71 (0.000)	13 (1.000)	5 (0.318)	8 (0.413)	19 (0.075)	5 (0.090)	0 (1.000)	854 (0.000)
AACAOSGLUB1	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)
ELRE2PCPAL1	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)
EMHVCHORD	73 (0.158)	126 (0.000)	0 (1.000)	83 (0.044)	118 (0.010)	205 (0.000)	46 (0.820)	0 (1.000)	1603 (0.000)
EREGCC	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	1 (0.005)
AAGACGTAGATACL12	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	2 (0.296)	0 (1.000)	0 (1.000)	109 (0.000)
ERELEE4	1045 (0.000)	1016 (0.000)	6 (1.000)	1088 (0.000)	1435 (0.000)	2241 (0.000)	608 (0.000)	5 (1.000)	5438 (0.000)
EVENINGAT	82 (0.389)	108 (0.000)	0 (1.000)	85 (0.465)	121 (0.086)	253 (0.000)	54 (0.123)	0 (1.000)	1344 (0.000)
GAGAGMGSA1	10 (0.668)	20 (0.000)	0 (1.000)	11 (0.682)	23 (0.704)	38 (0.009)	7 (0.042)	0 (1.000)	632 (0.000)
GBOXLERBCS	13 (0.017)	44 (0.000)	4 (0.474)	13 (0.046)	17 (0.053)	71 (0.000)	10 (0.022)	0 (1.000)	900 (0.000)
GBOXRELOSAMY3	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	15 (0.000)
GCBP2ZMGAPC4	0 (1.000)	15 (0.003)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	459 (0.000)
GCCCORE	142 (0.940)	2799 (0.000)	1810 (0.000)	156 (0.950)	202 (1.000)	451 (0.000)	140 (0.963)	66 (1.000)	8484 (0.000)
GLUTEBOX2OSGT2	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	13 (0.000)
GLUTECOREOS	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	15 (0.000)
HBOXPVCHS15	0 (1.000)	6 (0.037)	0 (1.000)	0 (1.000)	1 (0.550)	1 (0.781)	0 (1.000)	0 (1.000)	213 (0.000)
HDZIPIIIAT	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	12 (0.000)
HEXAMERATH4	92 (0.998)	1817 (0.000)	1160 (0.000)	104 (1.000)	137 (1.000)	350 (0.033)	90 (0.999)	47 (1.000)	6754 (0.000)
HSRENTHSR203J	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	1 (0.331)	0 (1.000)	0 (1.000)	53 (0.000)
HY5AT	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	17 (0.000)
IBOXCORE	3493 (0.000)	5237 (0.000)	1067 (0.000)	3614 (0.000)	3895 (0.000)	5769 (0.000)	2847 (0.000)	512 (0.004)	12108 (0.000)
IBOXLSCMCUCUMISIN	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	3 (0.000)
INRNTPSADB	3382 (0.000)	4931 (0.000)	755 (0.000)	3507 (0.000)	3836 (0.000)	5635 (0.000)	2687 (0.000)	410 (0.326)	11760 (0.000)
L1BOXATPDF1	351 (0.012)	470 (0.000)	2 (0.974)	378 (0.021)	519 (0.033)	863 (0.000)	225 (0.095)	1 (0.994)	3272 (0.000)
ABRE2HVA22	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	28 (0.000)
ABRE3HVA1	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	39 (0.000)
MRNA3ENDTAH3	4 (0.387)	11 (0.006)	0 (1.000)	4 (0.441)	10 (0.148)	24 (0.000)	2 (0.504)	0 (1.000)	341 (0.000)
MYB1LEPR	345 (0.000)	666 (0.000)	2 (1.000)	384 (0.000)	531 (0.000)	804 (0.000)	259 (0.000)	1 (1.000)	3695 (0.000)
MYB26PS	41 (0.525)	99 (0.000)	1 (0.231)	42 (0.693)	65 (0.529)	129 (0.000)	30 (0.458)	0 (1.000)	1023 (0.000)

Continues on next page

Motif	Eudicots	Monocots	Chlorophyta	Rosids	Fabids	Malvids	Angiosperm	GreenPlants	Oryza
ABRE3OSRAB16	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	24 (0.000)
MYBGAHV	1163 (0.000)	1296 (0.000)	11 (1.000)	1237 (0.000)	1548 (0.000)	2613 (0.000)	736 (0.000)	9 (1.000)	6307 (0.000)
MYCATERD1	1133 (0.000)	3156 (0.000)	288 (0.000)	1198 (0.000)	1584 (0.000)	2411 (0.000)	1006 (0.000)	121 (1.000)	9632 (0.000)
MYCATRD22	1133 (0.000)	3156 (0.000)	288 (0.000)	1198 (0.000)	1584 (0.000)	2411 (0.000)	1006 (0.000)	121 (1.000)	9632 (0.000)
ABREA2HVA1	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	1 (0.014)	0 (1.000)	0 (1.000)	370 (0.000)
NTBBF1ARROLB	1836 (0.000)	2384 (0.000)	96 (1.000)	1934 (0.000)	2403 (0.000)	3828 (0.000)	1283 (0.000)	77 (1.000)	8857 (0.000)
O2F1BE2S1	0 (1.000)	1 (0.262)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	81 (0.000)
O2F2BE2S1	0 (1.000)	1 (0.383)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	105 (0.000)
O2F3BE2S1	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	49 (0.000)
OCETYPEIIINTHISTONE	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	2 (0.000)
OCSGMHSP26A	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)
OSE1ROOTNODULE	2597 (0.000)	3476 (0.000)	209 (0.998)	2735 (0.000)	3120 (0.000)	5092 (0.000)	1855 (0.000)	137 (1.000)	10275 (0.000)
OSE2ROOTNODULE	3492 (0.000)	5660 (0.000)	1865 (0.000)	3614 (0.000)	3895 (0.000)	5768 (0.000)	2999 (0.000)	870 (0.000)	12324 (0.000)
P1BS	485 (0.001)	1141 (0.000)	33 (1.000)	520 (0.019)	689 (0.054)	1160 (0.000)	414 (0.001)	27 (1.000)	5504 (0.000)
PALINDROMICCBOXGM	1 (0.674)	15 (0.254)	0 (1.000)	1 (0.781)	1 (0.951)	13 (0.091)	1 (0.568)	0 (1.000)	444 (0.000)
PE2FNTRNR1A	0 (1.000)	20 (0.007)	13 (1.000)	0 (1.000)	1 (0.247)	1 (0.554)	0 (1.000)	0 (1.000)	373 (0.000)
PIATGAPB	4 (0.365)	25 (0.014)	1 (0.891)	4 (0.578)	9 (0.137)	17 (0.578)	4 (0.168)	0 (1.000)	523 (0.000)
ABREBNNAPA	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	37 (0.000)
PRECONSCRHSP70A	1197 (0.000)	4231 (0.000)	2137 (0.000)	1295 (0.000)	1563 (0.000)	2990 (0.000)	1114 (0.000)	438 (1.000)	10905 (0.000)
QELEMENTZMZM13	698 (0.000)	1764 (0.000)	197 (0.609)	732 (0.000)	971 (0.003)	1557 (0.000)	608 (0.000)	78 (1.000)	6435 (0.000)
RAV1AAT	3451 (0.000)	5738 (0.000)	1866 (0.000)	3583 (0.000)	3891 (0.000)	5700 (0.000)	3012 (0.000)	854 (0.000)	12295 (0.000)
RAV1BAT	438 (0.102)	2118 (0.000)	419 (0.000)	458 (0.390)	626 (0.967)	1124 (0.000)	413 (0.163)	95 (1.000)	7119 (0.000)
RBCSBOX2PS	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	1 (0.001)
RBCSBOX3PS	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	23 (0.000)
ABRECE1HVA22	0 (1.000)	9 (0.537)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	280 (0.000)
RE1ASPHYA3	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	26 (0.000)
RHERPATEXPA7	1335 (0.000)	4309 (0.000)	2144 (0.000)	1476 (0.000)	1810 (0.000)	3227 (0.000)	1223 (0.000)	483 (1.000)	10673 (0.000)
ABRECE3HVA1	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	4 (0.000)
RYREPEAT4	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	13 (0.000)
RYREPEATLEGUMINBOX	623 (0.000)	2022 (0.000)	129 (0.682)	658 (0.000)	921 (0.107)	1233 (0.000)	573 (0.000)	52 (1.000)	7353 (0.000)
RYREPEATVFLEB4	55 (0.000)	384 (0.000)	0 (1.000)	57 (0.000)	84 (0.001)	113 (0.000)	54 (0.000)	0 (1.000)	2304 (0.000)
SB1NPABC1	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)
SBOXATRBCS	44 (0.026)	200 (0.000)	19 (1.000)	48 (0.004)	78 (0.002)	97 (0.000)	39 (0.120)	5 (1.000)	1721 (0.000)
SGBFGMGMAUX28	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	7 (0.000)	0 (1.000)	0 (1.000)	102 (0.000)
SITEIIAOSPCNA	0 (1.000)	4 (0.372)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	197 (0.000)
SORLIP4AT	2 (0.150)	11 (0.003)	0 (1.000)	2 (0.188)	5 (0.025)	6 (0.385)	1 (0.443)	0 (1.000)	284 (0.000)
SORLREP4AT	10 (0.065)	23 (0.001)	0 (1.000)	10 (0.159)	18 (0.214)	27 (0.027)	6 (0.135)	0 (1.000)	567 (0.000)
ABREMOTIFAOSOSEM	6 (0.273)	56 (0.000)	2 (0.996)	11 (0.008)	16 (0.018)	22 (0.463)	6 (0.150)	0 (1.000)	651 (0.000)
SRENTTTO1	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	1 (0.001)

Continues on next page

Motif	Eudicots	Monocots	Chlorophyta	Rosids	Fabids	Malvids	Angiosperm	GreenPlants	Oryza
TATABOX1	3 (0.445)	22 (0.000)	0 (1.000)	5 (0.164)	9 (0.140)	15 (0.128)	1 (0.799)	0 (1.000)	291 (0.000)
TATAPVTRNALEU	624 (0.000)	627 (0.000)	0 (1.000)	650 (0.041)	808 (0.948)	1534 (0.000)	370 (0.000)	0 (1.000)	4523 (0.000)
TATCCACHVAL21	178 (0.685)	434 (0.000)	25 (1.000)	190 (0.797)	268 (0.620)	446 (0.027)	145 (0.718)	14 (1.000)	2925 (0.000)
TE2F2NTPCNA	2 (0.508)	26 (0.005)	3 (1.000)	2 (0.666)	5 (0.285)	18 (0.001)	1 (0.800)	0 (1.000)	420 (0.000)
TGA1ANTPR1A	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)
TRANSTART	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	7 (0.000)
UP1ATMSD	76 (0.993)	302 (0.000)	0 (1.000)	79 (0.998)	97 (1.000)	314 (0.000)	66 (0.994)	0 (1.000)	2792 (0.000)
WRKY71OS	3657 (0.000)	6049 (0.000)	2252 (0.000)	3781 (0.000)	4038 (0.000)	5834 (0.000)	3284 (0.000)	1068 (0.000)	12471 (0.000)
WUSATAg	470 (0.000)	480 (0.000)	2 (0.998)	492 (0.000)	631 (0.000)	1062 (0.000)	272 (0.000)	2 (0.996)	3563 (0.000)
XYLAT	299 (0.000)	363 (0.000)	4 (0.983)	311 (0.000)	424 (0.000)	782 (0.000)	192 (0.000)	3 (0.998)	2541 (0.000)
ABREZMRAB28	6 (0.139)	38 (0.010)	1 (0.969)	7 (0.094)	7 (0.425)	47 (0.000)	4 (0.389)	0 (1.000)	982 (0.000)
ACGTROOT1	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	3 (0.009)	0 (1.000)	0 (1.000)	181 (0.000)
ACGTSEED3	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	85 (0.000)
AGATCONSENSUS	4 (0.762)	6 (0.152)	0 (1.000)	4 (0.827)	9 (0.321)	28 (0.152)	2 (0.469)	0 (1.000)	259 (0.000)
AGL1ATCONSENSUS	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	62 (0.000)
AGL2ATCONSENSUS	4 (0.648)	5 (0.540)	0 (1.000)	4 (0.721)	6 (0.714)	21 (0.507)	2 (0.485)	0 (1.000)	302 (0.000)
AGL3ATCONSENSUS	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	23 (0.000)
ALF1NTPARC	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)
AMMORESIVDCRNIA1	77 (0.981)	257 (0.000)	46 (1.000)	84 (0.999)	115 (0.997)	259 (0.042)	66 (0.991)	13 (1.000)	2003 (0.000)
ANAERO3CONSENSUS	356 (0.059)	859 (0.000)	101 (1.000)	380 (0.183)	555 (0.000)	839 (0.000)	297 (0.018)	47 (1.000)	4482 (0.000)
ANAERO5CONSENSUS	3 (0.811)	15 (0.302)	0 (1.000)	3 (0.883)	9 (0.491)	14 (0.491)	2 (0.848)	0 (1.000)	352 (0.000)
ARE1	14 (0.819)	180 (0.000)	40 (1.000)	14 (0.938)	22 (0.994)	39 (0.926)	14 (0.753)	3 (1.000)	1459 (0.000)
ARECOREZMGAPC4	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	33 (0.000)
ARR1AT	3702 (0.000)	6069 (0.000)	2237 (0.000)	3824 (0.000)	4086 (0.000)	5844 (0.000)	3328 (0.000)	1071 (0.000)	12473 (0.000)
20NTNTNOS	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)
ASF1ATNOS	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)	0 (1.000)
ATHB5ATCORE	47 (0.607)	88 (0.000)	0 (1.000)	49 (0.791)	77 (0.522)	123 (0.017)	28 (0.568)	0 (1.000)	883 (0.000)

This table shows how many families each motif was conserved in in each clade. The p-values are in parentheses.

Appendix B

Binding site motifs

This is a list of the 144 motifs that were analyzed with respect to conservation in significantly many families.

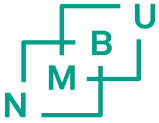
Motif name	Consensus motif
ABFs binding site motif	CACGTGGC
ACE promoter motif	GACACGTAGA
AG BS in SUP	CCATTTTTGG
AGL2 binding site motif	NNWNCCAWWWWTRGWWAN
ATHB2 binding site motif	CAATSATTG
CArG promoter motif	CCWWWWWGG
CArG2 motif in AP3	CTTACCTTTCATGGATTA
CCA1 motif1 BS in CAB1	AAACAATCTA
DREB1AND2 BS in rd29a	TACCGACAT
EIL1 BS in ERF1	TTCAAGGGGGCATGTATCTTGAA
EIL2 BS in ERF1	TTCAAGGGGGCATGTATCTTGAA
ERE promoter motif	TAAGAGCCGCC
GCC-box promoter motif	GCCGCC
HSEs binding site motif	AGAANNTTCT
L1-box promoter motif	TAAATGYA
LS5 promoter motif	ACGTCATAGA
MYB1 binding site motif	MTCCWACC
MYB2 binding site motif	TAACTSGTT
MYB3 binding site motif	TAACTAAC
Nonamer promoter motif	AGATCGACG
PII promoter motif	TTGGTTTTGATCAAAACCAA
SBP-box promoter motif	TNCGTACAA
T-box promoter motif	ACTTTG
TELO-box promoter motif	AAACCCTAA
Z-box promoter motif	ATACGTGT
AG BS in SPL/NOZ	AAAACAGAATAGGAAA
SORLREP1	TTWTACTAGT
SORLIP3	CTCAAGTGA
SORLIP5	GAGTGAG
MA0127.1	ANTTCTTATK
MA0120.1	TTKYYNYTNBCG
MA0129.1	HACGTCA
MA0001.1	CCAWAWATAG
MA0096.1	MTGACGT
-10PEHVPSBD	TATTCT

Continues on next page

Motif name	Consensus motif
20NTNTNOS	TGAGCTAAGCACATACGTCA
AACACOREOSGLUB1	AACAAAC
AACAOSGLUB1	CAACAAACTATATC
AAGACGTAGATACL12	AAGACGTAG
ABRE2HVA22	CGCACGTGTC
ABRE3HVA1	GCAACGTGTC
ABRE3OSRAB16	GTACGTGGCGC
ABREA2HVA1	CCTACGTGGC
ABREBNNAPA	CGCCACGTGTCC
ABRECE1HVA22	TGCCACCGG
ABRECE3HVA1	ACGCGTGTCCTC
ABREMOTIFAOSOSEM	TACGTGTC
ABREZMRAB28	CCACGTGG
ACGTROOT1	GCCACGTGGC
ACGTSEED3	GTACGTGGCG
AGATCONSENSUS	TTWCCWWWWNNGGWW
AGL1ATCONSENSUS	NTTDCCWWWWNNGGWAAN
AGL2ATCONSENSUS	NNWNCCAWWWWTRGWWAN
AGL3ATCONSENSUS	TTWCYAWWWWTRGWAA
ALF1NTPARC	TTACGCAAGCAATGACA
AMMORESIVDCRNIA1	CGAACTT
ANAERO3CONSENSUS	TCATCAC
ANAERO5CONSENSUS	TTCCCTGTT
ARE1	RGTGACNNNGC
ARECOREZMGAPC4	AGCAACGGTC
ARR1AT	NGATT
ASF1ATNOS	TGAGCTAAGCACATACGTCAG
ATHB5ATCORE	CAATNATTG
BOX1PSGS2	ATAGAAATCAA
BOX1PVCHS15	TAAAAGTTAAAAAC
BP5OSWX	CAACGTG
CAREOSREP1	CAACTC
CARG3ATAP3	CTTTCCATTTTTAGTAAC
CATATGGMSAUR	CATATG
CEREGLUBOX3PSLEGA	TGTAAAAGT
CGACGOSAMY3	CGACG
COREOS	AAKAATWYRTAWATAAAAMTTTTATWTA
CPRFPCCHS	CCACGTGGCC
DRE1COREZMRAB17	ACCGAGA
E2FANTRNR	TTTCCCGC
E2FAT	TYTCCCGCC
E2FBNTRNR	GCGGCAAA
ELRE2PCPAL1	ATTCTCACCTACCA
EMHVCHORD	TGTAAAGT
EREGCC	TAAGAGCCGCC
ERELEE4	AWTTCAAA
EVENINGAT	AAAATATCT
GAGAGMGSA1	GAGAGAGAGAGAGAGAGA
GBOXLERBCS	MCACGTGGC
GBOXRELOSAMY3	CTACGTGGCCA
GCBP2ZMGAPC4	GTGGGCCCG
GCCCORE	GCCGCC
GLUTEBOX2OSGT2	TCCGTGTACCA
GLUTECOREOS	CTTTCGTGTAC
HBOXPVCHS15	CCTACCNNNNNNNCTNNNNA
HDZIPIIIAT	GTAATSATTAC
HEXAMERATH4	CCGTCG
HSRENTHSR203J	CAAAATTTTGTA
HY5AT	TGACACGTGGCA
IBOXCORE	GATAA

Continues on next page

Motif name	Consensus motif
IBOXLSCMCUCUMISIN	AGATATGATAAAA
INRNTPSADB	YTCANTYY
L1BOXATPDF1	TAAATGYA
MRNA3ENDTAH3	AATGGAAATG
MYB1LEPR	GTTAGTT
MYB26PS	GTTAGGTT
MYBGAHV	TAACAAA
MYCATERD1	CATGTG
MYCATRD22	CACATG
NTBBF1ARROLB	ACTTTA
O2F1BE2S1	TCCACGTCGA
O2F2BE2S1	GCCACCTCAT
O2F3BE2S1	TCCACGTACT
OCETYPEIIINTHISTONE	GATCCGCGNNNNNNNNNNNNNNACCAATCS
OCSGMHSP26A	TGATGTAAGAGATTACGTAA
OSE1ROOTNODULE	AAAGAT
OSE2ROOTNODULE	CTCTT
P1BS	GNATATNC
PALINDROMICCBOXGM	TGACGTCA
PE2FNTRNR1A	ATTCGCGC
PIATGAPB	GTGATCAC
PRECONSCRHSP70A	SCGAYNRNNNNNNNNNNNNNNHD
QELEMENTZMZM13	AGGTCA
RAV1AAT	CAACA
RAV1BAT	CACCTG
RBCSBOX2PS	GTGTGGTTAATATG
RBCSBOX3PS	ATCATTTTCACT
RE1ASPHYA3	CATGGGCGCGG
RHERPATEXPA7	KCACGW
RYREPEAT4	TCCATGCATGCAC
RYREPEATLEGUMINBOX	CATGCAY
RYREPEATVFLEB4	CATGCATG
SB1NPABC1	CACTAACACAAAGTAA
SBOXATRBCS	CACCTCCA
SGBFGMGMAUX28	TCCACGTGTC
SITEIIAOSPCNA	TGGGCCCGT
SORLIP4AT	GTATGATGG
SORLREP4AT	CTCCTAATT
SRENTTTO1	TGGTAGGTGAGAT
TATABOX1	CTATAAATAC
TATAPVTRNALEU	TTTATATA
TATCCACHVAL21	TATCCAC
TE2F2NTPCNA	ATTCCCGC
TGA1ANTPR1A	CGTCATCGAGATGACG
TRANSTART	TAAACAATGGCT
UP1ATMSD	GGCCCAWWW
WRKY71OS	TGAC
WUSATAg	TTAATGG
XYLAT	ACAAAGAA



Norwegian University
of Life Sciences

Postboks 5003
NO-1432 Ås, Norway
+47 67 23 00 00
www.nmbu.no