University of Potsdam
Institute of Biochemistry and Biology, Golm, Germany

# *In silico* Identification of Genes Regulated by Abscisic Acid in *Arabidopsis thaliana* (L.) Heynh.

Dissertation

A thesis submitted to the
Faculty of Mathematics and Natural Sciences
of the
University of Potsdam
for the degree of

‘doctor rerum naturalium’

(Dr. rer. nat.)

by

Judith Lucía Gómez-Porras

Potsdam, October 2005

*A mi jefecito,*
*quién me enseño las cosas realmente importantes*
*que hay que saber en mi carrera.*
*Lo demás, está en los libros.*

# TABLE OF CONTENTS

## List of abbreviations and Symbols

| | |
|---|---|
| ABA | Abscisic acid |
| *aba* | ABA-deficient |
| *abi* | ABA-insensitive |
| ABRE | ABA-responsive element |
| As1 | Activation sequence 1 |
| bp | Base pair |
| bZIP | Leucine zipper transcription factor |
| ca | circa |
| Ca$^{2+}$ | Calcium |
| cADPR | Cyclic ADP-ribose |
| CaM | Calcium-calmodulin |
| cDNA | copy DNA |
| CDPK | Ca$^{2+}$-dependent protein kinases |
| CE | Coupling element |
| cv | cultivar |
| DEPC | Diethyl pyrocarbonate |
| DGPP | Diacyilglycerol phosphate |
| DNA | Deoxyribonucleic acid |
| dNTP | deoxy nucleotides |
| DRE | Dehydration responsive element |
| DTT | Dithiothreitol |
| EGTA | Ethylenglycol-(β-aminoethylether)-N,N,N',N'-tetraacetate |
| EST | Expressed sequence tags |
| FDR | False discovery rate |
| *g* | gravity constant |
| GEPAS | Gene expression pattern analysis suite |
| GO | Gene ontology |
| GTP | Guanosine triphosphate |
| HMM | Hidden Markov model |
| $IP_3$ | Inositol 1,4,5-triphosphate |
| IUPAC | Nomenclature committee of the international union of biochemistry |
| kb | Kilobase |
| LEA | Late embryogenesis abundant |
| LiCl | Lithium chloride |
| MAPK | Mitogen-activated protein kinases |
| Mb | Megabase |
| $MgCl_2$ | Magnesium chloride |
| MS | Murashige and Skoog |
| NaOH | Sodim hydroxide |
| ORF | Open reading frame |
| PA | Phosphatidic acid |
| PCA | Principal component analysis |
| PCR | Polymerase chain reaction |
| $PI_{(3,5)}P_2$ | Phosphatidylinositol 3,5-bisphosphate |
| $PLA_2$ | Phospholipase $A_2$ |
| PLC | Phospholipase C |
| PLD | Phospholipase D |
| PSFM | Position specific frequency matrices |
| RNA | Ribonucleic acid |
| RNA Pol | RNA polymerase |
| RT-PCR | Real-time PCR |
| SD | Standard deviation |
| TAF | TBP-associated factor |
| TBP | TATA-box binding protein |
| TF | Transcription factor |
| TFBS | Transcription factor binding site |
| TIGR | Institute for genomic research |
| UV | Ultraviolet |
| v | volume |

## Symbols

### NUCLEOTIDES

| | |
|---|---|
| A | Adenine |
| C | Cytosine |
| G | Guanine |
| T | Thymine |
| M | A or C |
| R | A or G |
| W | A or T |
| S | C or G |
| Y | C or T |
| K | G or T |
| B | C, G or T |
| D | A, G or T |
| H | A, C or T |
| V | A, C or G |
| N | Any |

## List of common and scientific names of plants

| **Common Name** | **Scientific Name** |
|---|---|
| Arabidopsis (thale cress) | *Arabidopsis thaliana* |
| Barley | *Hordeum vulgare* |
| Carrot | *Daucus carota* |
| Cotton | *Gossypium hirsutum* |
| Maize | *Zea mays* |
| Parsley | *Petroselinum crispum* |
| Pea | *Pisum sativum* |
| Petunia | *Petunia hybrida* |
| Resurrection plant | *Craterostigma plantagineum* |
| Rice | *Oryza sativa* |
| Soybean | *Glycine max* |
| Tobacco | *Nicotiana tabacum* |
| Tomato | *Lycopersicon esculentum* |
| Wheat | *Triticum aestivum* |

# Chapter 1: Introduction

## 1.1 Concepts in gene regulation

Eukaryotic gene regulation is considerably more complex than in bacteria. The increased complexity presumably facilitates the sophisticated regulation needed to direct the activities of many different cell types in a multicellular organisms[24]. Gene expression is regulated at several stages in the pathway from DNA to protein, namely from transcription to protein degradation. Transcriptional regulation includes controls during initiation, elongation and termination of transcription. The produced RNA is also subjected to regulation. Mechanisms controlling the regulation after transcription are regarded as posttranscriptional modifications (posttranscriptional regulation). It includes regulation at the level of RNA stability, translation, modification of amino acid residues by addition of foreign groups such as phosphates or sugars, and regulation at the level of protein degradation[56,111].

In many cases, transcriptional initiation is the most important and tightly regulated level of gene expression[56]. In eukaryotic cells, genes are transcribed by three different RNA polymerases[64]:

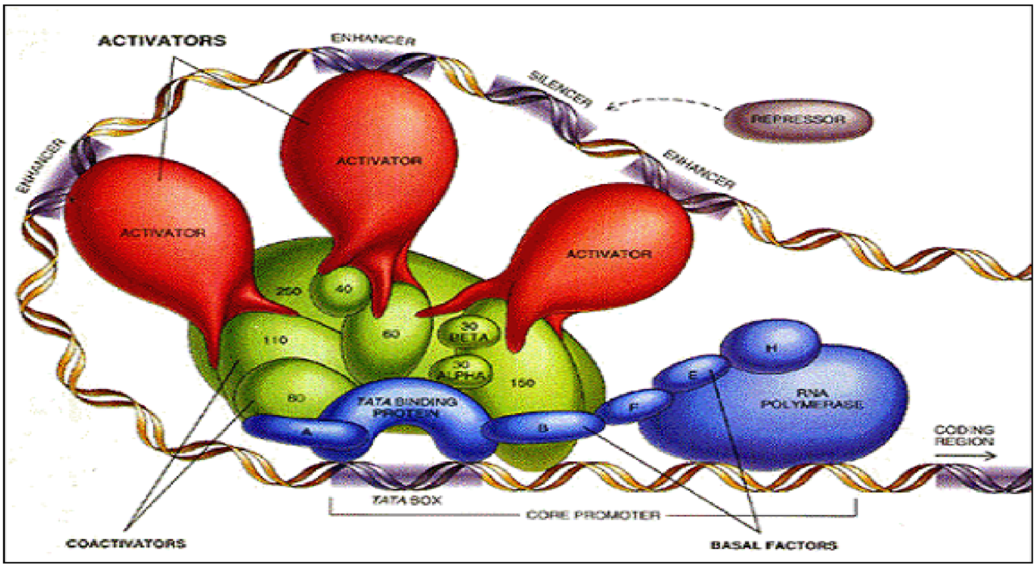
- RNA polymerase I (PolI) it transcribes the ribosomal RNA (rRNA) genes for the precursor 28S, 18S and 5.8S molecules.
- RNA polymerase II (PolII) transcribes messenger RNA (mRNA) genes (protein coding genes), and also some small nuclear RNA genes (snRNAs)
- RNA polymerase III (PolIII) transcribes transfer RNA (tRNA) genes and other small RNA genes.

Any protein that is needed for the initiation of transcription is defined as a transcription factor. Most of the transcription factors are released before RNA polymerase II (RNA PolII) leaves the promoter[43]. The factors rather than the enzymes themselves are responsible for recognizing the sequence components of the promoter[43,64].

Promoters recognized by RNA PolII exhibit more sequence diversity than promoters recognized by the other polymerases, and are modular in design[56]. The factors that assists RNA PolII can be divided into three general groups:

1. General factors are required for the initiation of transcription at all promoters. They join with RNA PolII to form a complex surrounding the startpoint, because RNA PolII alone cannot initiate transcription. These auxiliary factors are called TFIIX (TFII for transcription factor of the RNA PolII, and “X” identifies the individual factor). Together with RNA PolII this complex constitute the **basal transcription apparatus**[64]. A

schematic view of the molecular apparatus controlling transcription in human cells is shown in Figure 1-1[119]. The RNA PolII and the general factors are showed in blue, and are regarded as “basal factors”[43,64]. They are essential for transcription but cannot by themselves increase its rate. The fist step during the formation of the complex at a TATA-box containing promoter is the binding of the factor TFIID to the TATA-box. TFIID itself is composed of multiple subunits, including the TATA-binding protein (TBP), which binds a second factor (TFIIB), and a variety of other subunits called TAFs (for TBP-associated factors). TFIIDs containing different TAFs may recognize different promoters[64]. In Figure 1-1 TAFs are regarded as “coactivators” since TAFs and not TBP itself are the targets for the protein binding of transcription factors (general and inducible). In the figure they are shown in green and named according to their molecular weights (in kilodaltons)[119]. TFIIB bound to TFIID serves in turn as a bridge to RNA polymerase, which binds to the TBP-TFIIB complex in association with the factor TFIIF. Following recruitment of RNA PolII to the promoter, the binding of the factors TFIIE and TFIIH is required for initiation of transcription[24,43,64].



*Picture taken from UC Berkeley website (http://www.berkeley.edu/news/features/1999/12/09_3dimage.html). Original source : Scietific American*

**Figure 1-1: A schematic view of the typical components of a gene transcribed by RNA polymerase II.** Basal factors (blue shapes at bottom) together with the RNA Pol form the basal transcription apparatus. General and inducible transcription factors (activators in red or repressors in gray) interact with the basal factors through coactivators (green), that are proteins linked to the TBP. Coactivators are named according to their molecular weights (in kilodaltons)

2. <u>Upstream factors</u> are DNA binding proteins that recognize specific short DNA sequences located upstream of the startpoint (e.g. Sp1, which binds the GC box). The activity of these factors is not regulated. They are ubiquitous, and act upon any

promoter that contains an appropriate DNA binding site. They increase the efficiency of transcriptional initiation, and are required for a promoter to function at an adequate level[43,64].

3. Inducible factors like the upstream factors they recognize short DNA sequences but they have a regulatory role. They are synthesized or activated at specific developmental stages or in specific tissues. Therefore, inducible factors are responsible for the spatial and temporal control of transcription. In Figure 1-1 inducible factors that increase the level of transcription are shown in red (activators), and inducible factors that decrease or abolish transcription are shown in grey (repressors) [43,56,64].

A promoter that contains only elements recognized by general and upstream factors should be transcribed in any cell type. Such promoters may be responsible for the expression of cellular genes that are constitutively expressed (called sometimes housekeeping genes). The upstream and inducible factors function by interacting with the basal transcription apparatus, typically with certain general factors (shown in green as “coactivators” in Figure 1-1)[64].

### 1.1.1 Transcription factors

Transcription factors are proteins that are needed for the initiation of transcription. They might have a regulatory role mostly by recognizing short DNA sequences (*cis*-acting sites). However, binding to DNA is not the only means of action of a transcription factor. A factor might recognize another factor, or may be incorporated into an initiation complex only in the presence of several proteins[56,64].

The transcription factors that interact with the DNA to control the transcription recognize DNA sequences more or less specifically. Once bound to the DNA, these factors may influence transcription through several mechanisms[56]:

1. In most cases they enhance the formation of the preinitiation complex at the TATA-box/initiator element (a general upstream factor). The interaction with the preinitiation complex is mediated by a trans-activation domain, able to interact with components of the basal transcription apparatus.
2. Some transcription factors cause alterations in the chromosomal architecture, rendering the chromatin more accessible to the RNA polymerase.
3. Some auxiliary factors adjust an optimal DNA conformation for the activity of other transcription factors.
4. Some factors exert repressing influences, either directly by an active inhibiting domain, or indirectly by disturbing the required ensemble of factors within a regulatory sequence.

5. There is a group of transcription factors that do not directly bind to DNA, but assemble into higher-order complexes through protein-protein interactions.

Most transcription factors are modular in structure. Usually the following protein domains are found:

- A DNA binding domain.
- An oligomerization domain that allows to form dimers or higher order complexes with other transcription factors or proteins of the transcriptional machinery.
- A trans-activation domain, which is often characterized by a significant over-representation of certain types of amino acid residues (e.g. glutamine-rich, proline-rich, serine/threonine rich or acidic activation domains).
- A modulating region which is frequently a target of modifying enzymes such as kinases and phosphatases.
- Sometimes they might also have a ligand-binding domain.

The proposed classification scheme of transcription factors made by Wingender[56] is mainly based on the properties of the DNA-binding domain. According to this, four large superclasses of DNA-binding domains are recognized:

1. Basic domains: factors that have a stretch of mainly basic amino acid residues. For example the leucine zipper proteins[5,56]. The leucine zipper is a stretch of amino acids rich in leucine residues that provides a dimerization motif. Zippers may form homo or heterodimers[43,56]. The region adjacent to the leucine repeats is highly basic in each of the zipper proteins. In Figure 1-2 the crystal structure of a protein-DNA complex that represents the leucine zipper GCN4 is shown.
2. Zinc-coordinating DNA-binding domains. The DNA binding domain is brought to a defined conformation by coordinated zinc ion(s). This DNA-binding domain was originally found in the factor TFIIIA, which is required for RNA PolIII to transcribe 5S rRNA genes[43,56]. In Figure 1-2 the crystal structure of the protein-DNA complexes of TFIIIA and Zif268 are shown. Both transcription factors contain zinc finger motifs.
3. Helix-turn-helix: The proteins that have this motif have both the ability to bind DNA and to dimerize, forming both homo and heterodimers. They share a common type of sequence: a stretch of 40-50 amino acids contains 2 amphipathic $\alpha$-helices separated by a linker region (the loop) of varying length[43,56]. Figure 1-2 shows the crystal structure of the protein-DNA complex of ETS-1, a helix-turn-helix transcription factor.
4. $\beta$-Scaffold factors with minor groove contacts: factors where the DNA-contacting interface is exposed by a scaffold of suitably arranged $\beta$-strands and which perform minor groove contacts. In Figure 1-2 the protein-DNA complex of the TBP, a $\beta$-scaffold factor is shown.

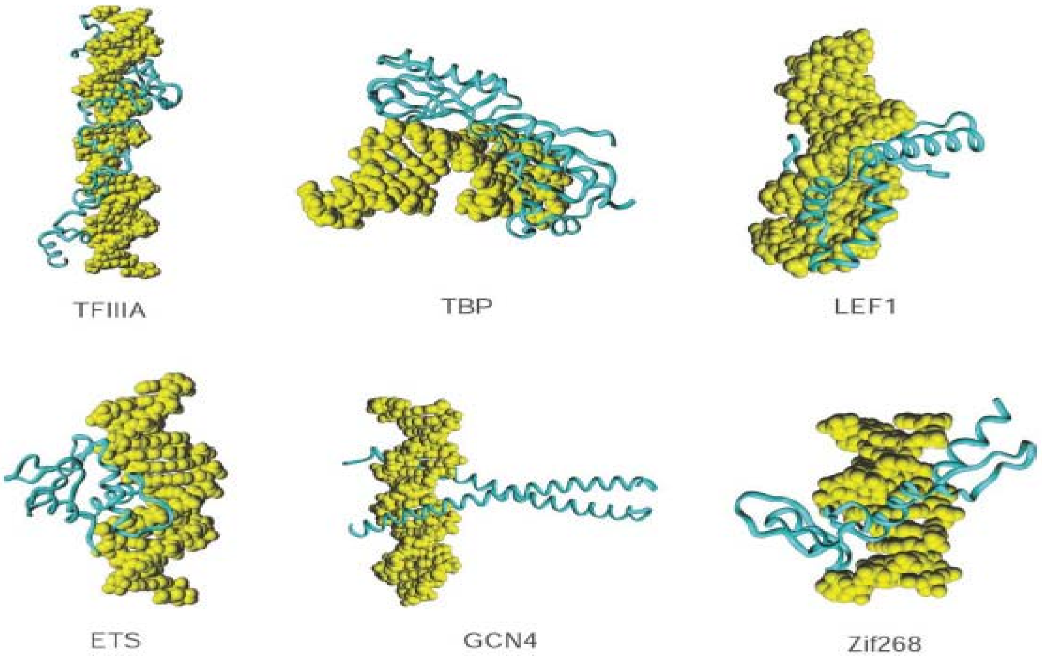


**Figure 1-2: Examples of DNA-binding domains. In TFIIIA the DNA-binding domain contains zinc-finger motifs.** TBP contains a minor groove DNA-binding motif. LEF-1 is a HMG box. ETS-1 is a helix-turn-helix transcription factor. GCN4 is a bZIP transcription factor. In Zif268 the DNA-binding domain contains zinc-finger motifs. Picture taken from Dervan, 2001[28]

### 1.1.2 Representation of transcription factor binding sites

To model a transcription factor binding site (TFBS) the underlying biochemical process has to be taken into account. A transcription factor recognizes a DNA sequence, which shares some common features that almost always appears at the same position in the recognized sequences. Although there is a conserved core within the recognition site, generally certain positions do not influence binding affinity and hence show more variability[56,64,108,111].

The specificity of a transcription factor for its target DNA sequence is different from the specificity of other DNA-binding proteins such as restriction enzymes. The recognition sequence of a restriction enzyme is a defined DNA sequence, in some cases ambiguities are allowed. All sites that match the recognition sequence will be cut (unless modified by methylation) and only matching sequences will be cut. In contrast, TFBSs often show variations in their recognized sequences, and only few of the positions of the binding site are conserved[108]. It makes biological sense that transcription factor binding sites are variable, whereas restriction sites are not. Restriction sites are used as defence mechanisms, and they need to have an all or none activity. But TFBSs can take advantage of the variability in the sites to better control gene transcription. Not all promoters should have the same activity, because some proteins are required by the cell at much higher level than others. The variability in expression can be partially attained by having promoters with different intrinsic affinities for the RNA polymerase, which implies different sequences in the binding sites. Likewise, transcription factors often control the expression of several genes, if these genes are needed to be expressed at different levels, that can be accomplished by having binding sites with different sequences and different affinities for the protein[56,108]. As an example, in Figure 1-3 some binding sites recognized by G-box binding factors (bZIP transcription factors) are shown. It can be observed that all recognition sequences have an invariable core of ACGT shown in red, that starts at the fourth position.

| | |
|---|---|
| | T C C **A C G T** C T C T |
| | C G T **A C G T** G T C G |
| | C C T **A C G T** G G C G |
| | G G G **A C G T** G G C G |
| | C A C **A C G T** C C C G |
| | C G T **A C G T** G T A C |
| | T G T **A C G T** G C T G |
| | G A T **A C G T** G T T T |
| consensus | C G T **A C G T** G T C G |
| alternative consensus | B V B **A C G T** B B V B |

**Figure 1-3: ABRE binding sites.** Sequences recognized by G-box binding factors (bZIP transcription factors)

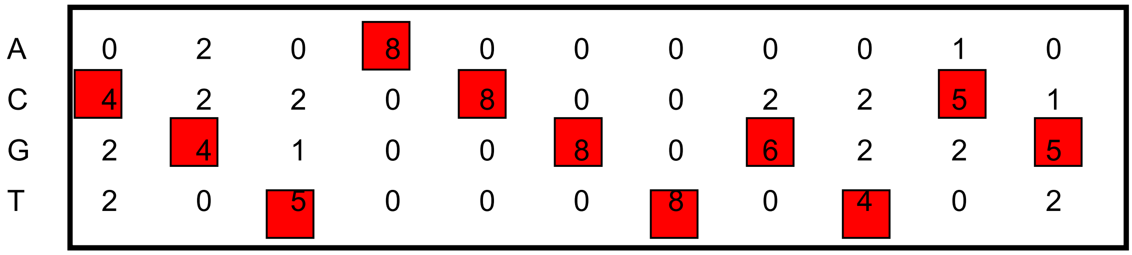
To represent TFBSs the concept of the consensus sequence has been widely used. However, the exact definition is somewhat arbitrary. In general, the consensus sequence refers to a sequence that matches all of the examples of a known binding site closely, but not necessarily exactly. There is a trade-off between the number of mismatches allowed, the ambiguity in the consensus sequence, and the sensitivity and precision of the representation[108]. In Figure 1-3 two possible consensus sequences can be deduced, and are presented. One consensus sequence refers to the most prominent base at each position and does not allow ambiguities. The alternative consensus sequence refers to all observed bases at each position. If the first consensus sequence is used to identify ABRE binding sites only one from eight sequences of the list will be identify. Considering the frequency of A,C,G and T in intergenic sequences of the model plant *Arabidopsis thaliana* (0.34 for A and T, and 0.16 for C and G)[1] there would be about one match each 28 Mb. With the alternative consensus all sequences of the list will be identify. However, a genome wide screening in *A. thaliana* will identify a *cis*-element each 6 kb.

An alternative to consensus sequences is a matrix representation of the site. Figure 1-4 shows a frequency matrix based on the binding sites shown in Figure 1-3. In a frequency matrix the number of occurrences of each nucleotide at a certain position is count. To predict new binding sites the counts of the matrix are taken into account, and each putative binding

[1] See section 5.1

site is predicted with a score[108]. Generally, putative binding sites composed of nucleotides rarely observed in the known binding sites represented in the matrix will have low scores. The decreases of the scores depend on the differences with respect to the most frequent nucleotides.

Several hundred matrices for specific transcription factors are available in databases such as TRANSFAC (http:/www.biobase.de). However, the available matrices for many transcription factors are not specific enough to enable a reliable prediction of sites in long sequences. For large genomes, thousands of potential binding sites are expected to be found just by chance.



| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| A | 0 | 2 | 0 | 8 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| C | 4 | 2 | 2 | 0 | 8 | 0 | 0 | 2 | 2 | 5 | 1 |
| G | 2 | 4 | 1 | 0 | 0 | 8 | 0 | 6 | 2 | 2 | 5 |
| T | 2 | 0 | 5 | 0 | 0 | 0 | 8 | 0 | 4 | 0 | 2 |

**Figure 1-4: Matrix representation of the binding sites presented in Figure 1-3.** The red boxes correspond to the consensus sequence CGTACGTGTCG

### 1.1.3 Identification of known *cis*-elements in DNA sequences

The computational detection of regulatory sites (*cis*-elements) in DNA sequences is a difficult task, especially in eukaryotes where the TFBSs are generally DNA sequences shorter than those found in prokaryotes[108]. Additionally, the recognized sequences are variable, and the binding sites can be dispersed over very large distances. As a consequence, the rate of false positive predictions is very high. Errors in the recognition of putative binding sites occur because of limited knowledge about the structure of a binding site, and the lack of well define models of transcriptional regulation[56].

The analysis of a query DNA sequence with a consensus sequence or a matrix that represents a TFBS results in a list of potential binding sites, and their positions in the query sequence. The output depends critically on the number of binding sites considered to construct the *cis*-element model (principally if the query sequence has been screened with consensus sequences) and the quality of the *cis*-element model (consensus sequence or matrix). A binding site represented either by a matrix or a consensus sequence with low specificity yields frequent matches within a sequence, most of them being false positives[33].

The representation of a binding site as a consensus sequence and posterior screening of a DNA sequence yield a simple yes/no decision. In a screening with consensus sequences mismatches that could be tolerated by the binding protein will not be discriminated from a mismatch that abolishes binding. Matrices are less sensitive to sequence selection, and provide a qualitative rating (score), suggesting the likelihood of the binding site. A single mismatch in a critical position normally greatly reduces the score of the match[56,108,111].

The challenge is to reduce the number of falsely predicted sites. Recently developed methods to detect TFBSs often employ criteria such as conservation of sites across different species, clustering of binding sites in regulatory regions, and enrichment of specific sites in co-regulated genes. These criteria are based on observations from relatively few experimentally dissected regulatory regions. Nevertheless, it is already evident that predictive success increases dramatically as experimental data accumulates and defined models of distribution and interactions between *cis*-elements in regulatory regions become more refined[18,67,124].

## 1.2 ABA signalling in plant development and growth

The phytohormone abscisic acid (ABA) is implicated in a number of developmental events during the life cycle of higher plants. ABA is not only involved in mediating the response to a number of environmental stresses, including drought, cold and high salinity, it also plays a significant role in embryo development and seed maturation [30,117].

ABA was discovered independently through the study of abscission of cotton fruits, and through analysis of dormancy in sycamores (Davies *et al.* 1988 and Addicot *et al.* 1983 cited by Fedoroff, 2002[30]). ABA is synthesized in almost all cells, but it is also transport from roots to shoots, and the circulation of ABA in both xylem and phloem is an important aspect of its physiological role[30]. ABA is sent from the roots to target cells in shoots, as soil is drying, to regulate stomatal closure and leaf development. The most extensively investigated developmental and physiological effects of ABA are those involved in seed maturation and dormancy, and the regulation of stomatal movements[30].

In *A. thaliana* alone, more than 50 loci have been demonstrated to function in various aspects of ABA response. Their products include transcription factors, protein phosphatases and kinases, RNA binding/processing proteins, GTP proteins, enzymes of phospholipid or phosphoinositide metabolism, and proteins regulating vesicle trafficking or membrane localization of specific proteins. Some of these loci have also been identified via diverse genetic screens including defects in response to other phytohormones (ethylene, auxin or brassinosteroids), abiotic stresses (osmotic, salt, cold or UV light), and sugars (glucose or sucrose). The repeated identification of a few loci affecting response to multiple signals had led to the suggestion that these genes are points of "cross-talk" among signalling pathways[12].

To identify the molecular mechanisms of ABA action several approaches have been adopted. One of them is the isolation and characterization of components involved in the transduction of the signal. Taking advantage of ABA response mutants some of the genes involved in the signal transduction pathway have been cloned [12,16,30,101,103]. ABA-deficient mutants (*aba*) as well as ABA-insensitive (*abi*) mutants have been identified. The *aba* mutants have low levels of ABA due to attenuated levels of accumulation in both seeds and leaves. The *abi* mutants are impaired or deficient in various responses regulated by ABA such establishment of dormancy or stomatal closure [30,117].

Because it has been well documented that ABA regulates the expression of a variety of genes, another approach to study the mechanism of ABA action involves the identification of *cis*-acting elements necessary and sufficient for ABA response, and the isolation of trans-acting factors interacting with these DNA sequences[12,16,30,101,103].

The molecule or molecules that perceive the ABA signal (ABA receptor) remain unknown[30]. Different results suggest the presence of both intra and extracellular reception sites for ABA[101].

The signalling mechanisms involved in the transduction of the signal include a variety of phospholipid-based signalling pathways, including phospholipase C (PLC), D (PLD) and $A_2$ ($PLA_2$), and pathways involving the formation of diacylglycerol pyrophosphate (DGPP) and phosphatidylinositol 3,5-bisphosphate ($PI_{(3,5)}P_2$). Protein kinases, and especially mitogen-activated protein kinases (MAPK) are also involved[72]. The signalling components connecting the ABA-activated MAPK cascade to ABA reception have not yet been identified[30,72].

In guard cells there is evidence that PLC- and PLD-generated phosphatidic acid (PA) and inositol 1,4,5-triphosphate ($IP_3$) serve as second messengers. PA promotes inactivation of an inward-rectifying $K^+$ channel, and $IP_3$ stimulates $Ca^{2+}$ release from the vacuole. Cyclic ADP-ribose (cADPR) is required for both, ABA-induced stomatal closure and ABA-activated gene expression. Furthermore, cADPR stimulates also release of $Ca^{2+}$ from the vacuole[30]. Evidence is accumulating that membrane vesicle trafficking and fusion are central to ABA signalling[30].

The signalling mechanisms involved in transcriptional and posttranscriptional regulation by ABA have not been explored as extensively as have been the mechanisms of signalling to membrane channels in guard cells. Nonetheless, several transcription factors have been identified that confer an *abi* phenotype (*abi3*, *abi4* and *abi5*) when mutated[12,30]. In addition, two of the five *abi* mutants isolated so far, carry semidominant mutations in highly similar proteins that belong to the 2C class of serine-threonine protein phosphatases (PP2C) (*abi1* and *abi2*). Both mutants have been used to study the role of protein phosphorylation and dephosphorylation in ABA signalling[30].

In addition to the MAPK cascade, $Ca^{2+}$-dependent protein kinases (CDPKs) are also implicated in transmitting the ABA signal to the transcriptional machinery[30]. Plant CDPKs are similar to mammalian calcium-calmodulin (CaM)-dependent protein kinase II, and also contain an integral CaM domain. Nevertheless, little is known about the role of CaM and CaM-binding proteins in signalling in plants[30]. It has been hypothesized that protein kinases and phosphatases that participate in ABA signalling could regulate some transcription factors constitutively bound to the promoter of ABA-responsive genes, by phosphorylation or dephosphorylation. In mammals for example, the bZIP transcription factor CREB which recognizes an ACGT-core *cis*-element is activated by phosphorylation[16].

## *1.3 ABA-related cis-elements*

The comparison of different promoters of ABA-responsive genes revealed the presence of a conserved sequence. This sequence was first identified as a *cis*-acting element named ABA-responsive element (ABRE) in wheat (*Triticum aestivum*), in the gene *EM* which functions mainly in seeds during late embryogenesis, and in rice (*Oryza sativa*) in the gene *RAB16*, which is expressed in both dehydrated tissue and maturating seeds[117]. ABRE contains an ACGT core similar to the one found in the so-called G-boxes, that are involved in responses to other environmental and physiological cues, such as light, cumaric acid, auxin, jasmonic acid and salicylic acid [32,69]. Studies with several promoters have led to the isolation of G-box binding proteins, all of which are leucine zippers (bZIP)[103]. Considering that G-boxes are found in promoters that respond to diverse environmental and physiological signals, it is proposed that the bases flanking the ACGT-core and the interaction with another *cis*-element close to the G-box determine the specificity of the promoter. Specific point mutations within either box result in a dramatic reduction on the induction of a reporter gene upon specific stimuli (ABA, light)[103,121].

In the case of ABA-induced genes, three coupling elements have been described: coupling element 1 (CE1), coupling element 3 (CE3) and a *cis*-element called Dehydration Responsive Element (DRE)[14,16,101-103,132]. The sequences of CE1 and CE3 are different. However, both elements have a high content of cytosines and/or guanines[16]. Mutational analyses have showed that the essential sequence of CE1 in barley (*Hordeum vulgare*) is CCACC. The most critical base appeared to be the adenine in the middle of the element[103]. The essential sequence of CE3 in barley is GCGTGTC, and because it is nearly identical to the ABRE sequence ACGTGGC, Hobo *et al.* 1999 suggested to classify CE3 as a non-ACGT ABRE[50].

In barley it was found that the orientation of the ABRE and the CE1 is important for a high level of ABA induction. The expression of a reporter gene was higher when the *cis*-elements ABRE and CE1 were separated by 10, 20 and 30 bp, and lower when the separation was 5, 15 or 25 bp. It appears that the protein that recognizes the G-Box (a bZIP transcription factor) and the protein that recognizes the CE1 (possibly an AP2 class transcription factor) have to be located at the same side of the DNA helix in order to interact with each other[103]. In the case of CE3, the orientation of ABRE and CE3 did not play an important role, but the distance between them was important. Closely located ABRE and CE3 *cis*-elements showed more induction of the reporter gene than elements that had a separation of 25 bp. The data suggest that the interaction of ABRE and CE3 is mediated by other proteins than in the case of ABRE-CE1[103].

DRE has been reported to be a coupling element of ABRE only in *A. thaliana*, and has been observed to play a role in the ABA-mediated induction of the gene *RD29A*[75]. The sequence of the *cis*-element is TACCGACAT, with a conserved core (CCGAC)[134]. DRE has been reported in other genes responsive to drought and cold stress like *KIN1*, *COR6.6/KIN2*, *COR15A*, *RD17/COR47* from *A. thaliana*, all genes contain both DRE and ABRE in their promoter regions[75]. DRE is considered also as the most important ABA-independent stress responsive *cis*-element in genes regulated by osmotic stress. A single copy of DRE is sufficient to induce gene expression, which indicates that it does not require other elements for its function in stress-inducible gene expression, unlike ABRE[135]. The similarity between the core sequences of CE1 and DRE suggest that AP2 domain transcription factors that bind CE1 may also be able to bind a different consensus sequence such as DRE although with less affinity[103].

The cloning of stress-related MYB transcription factors from *A. thaliana* and *Craterostigma plantagineum* suggest that this class of transcription factors and their recognition sequences are involved in ABA-induced transcription[16]. A MYB binding sequence is present in the promoter of the ABA-inducible gene *RD22* and is involved in induction by ABA [2,3,54,133]. MYB binding sites have been reported also in promoters of the genes *AtADH1*, *COR6.6/KIN2* and *RD20* from *A. thaliana*. It has been suggested that one of the transcription factors involved in the recognition of these MYB binding sites in response to ABA is the gene *AtMYB2*, and that this system is different from the ABRE-bZIP regulatory system observed in vegetative tissues and seeds[2].

Finally, the *cis*-element As1 (activation sequence 1) has been observed in the promoter region of the gene *RD29A* of *A. thaliana*[75].The *cis*-element was first found in viral and T-DNA promoters. As1-like elements have also been found as functional elements of plant promoters activated in the course of a defence response upon pathogen attack. They are recognized by plant nuclear As1 like binding factors ASF-1, the major component of which is a basic/leucine zipper (bZIP) protein in *Nicotiana tabacum*[61] . In the ABA-responsive gene *RD29A* As1 has been considered as a *cis*-element that confers root-specific. Base substitution analyses in the As1 binding site led to a lower induction of a reporter gene, compared to the wild-type element. However, responsiveness of the promoter to ABA is not completely abolished [75].

## Chapter 2: Aim of this work

Osmotic stress has a major influence on crop production, affecting dramatically yield harvest. The involvement of ABA in osmotic stresses such as drought, cold or high salinity has been extensively documented[2,3,16,30,35,53,58,68,103,105,117,133,135]. Considering the important biological and economical role of osmotic stress and ABA, and the limited knowledge on the interaction between *cis*-elements in ABA-responsive genes, it was proposed to use bioinformatic approaches to detect genes regulated by ABA in *A. thaliana*. The study was made at a genome-wide scale, taking advantage of the genome sequence of *A. thaliana* available since 2000 (Arabidopsis Genome Initiative - AGI[1]).

The whole genome of *A. thaliana* was analysed to identify combinations of *cis*-elements known to be involved in the regulation of ABA-responsive genes. During the formulation of this PhD project two key aspects were considered:

1. The project should provide insights into possible interactions between *cis*-elements involved in the regulation of ABA-responsive genes in *A. thaliana*. The extensive analysis of the *in silico* predictions should help to identify some interaction principles between *cis*-elements, and to propose a model about for the mechanisms of regulation of ABA-responsive genes.

2. The project should provide a comparison of *in silico* predictions with experimental data in order to evaluate the biological reliability of computational predictions. The experiences from this comparison should facilitate the design and refinement of future laboratory experiments.

# Chapter 3: Materials

## 3.1 Bioinformatics

Computational methods were performed on a workstation with a 1.4 GHz Athlon™ processor running on a LINUX operating system (SuSE Linux, version 2.4.20).

## 3.2 Molecular biology

### 3.2.1 Chemical reagents and enzymes

All chemicals used were analytical grade and were from Boehringer (Mannheim), Duchefa (Haarlem, Netherlands), Fluka (Buchs SG, Switzerland), Invitrogen (Invitrogen GmbH, Karlsruhe), Merck (Darmstadt) and Sigma-Aldrich (Taufkirchen-Munich).

DNase I was from Roche (Roche Applied Sciences, Mannheim), and the cDNA synthesis kit from Amershan (Amershan Biosciences, Freiburg).

SYBR® Green Master Mix reagent was obtained from Applied Biosystems (Applied Biosystems, Warrington, UK).

### 3.2.2 Buffers and solutions

Unless otherwise specified, solutions were prepared in Milli-Q-grade deionized Water. Buffers that were used for RNA analyses were made with Milli-Q-Water containing the strong ribonuclease inhibitor diethyl pyrocarbonate (DEPC).

### 3.2.3 Plant material and growth conditions

Wild-type plants of *Arabidopsis thaliana* (L.) Heynth., ecotype C24 were grown hydroponically in 50% Murashige and Skoog medium[73] (MS-medium), supplemented with 2% sucrose (w/v), under controlled conditions in a growth chamber, under long-day conditions (16 h light / 8 h dark), at 21°C and 65% relative humidity

### 3.2.4 Primers

All primers were synthesized by TIP-MOLBIOL (Berlin) and have been described by Czechowski *et al.* 2004[25]

| **Gene** | **Primer** | **Sequence (5’→ 3’)** |
|---|---|---|
| Actin 200/600bp | forward<br>reverse | ACTTTCATCAGCCGTTTTGA<br>ACGATTGGTTGAATATCATCAG |
| *At1g42990*<br>bZIP TF | forward<br>reverse | TGGCTAAAAAACGAAGAAGGAGAG<br>TCAAGCATACGTCCTAGTCTCAAG |
| *At2g46590*<br>DAG2/Dof | forward<br>reverse | TGAAACAGGAGACGACGAGGAACC<br>TCATCAGCAGCAGCCTTCATCATC |
| *At5g10030*<br>OBF4/bZIP TF | forward<br>reverse | TGCGGTAACAGAACCTTGAGAAGC<br>TGTGGAAAACTTCAGCAGAGCGG |
| *At5g39610*<br>NAM/NAC TF | forward<br>reverse | GGCTGGTTCCATTCGGTTAATGTG<br>TCCCCAGCGAATGTCGTAGTGGAT |

# Chapter 4: Methods

## 4.1 Bioinformatics

### 4.1.1 Construction of datasets

Genome sequences and gene coordinates stored at The Institute for Genomic Research (TIGR) were used for the analyses[2]. The gene coordinates define the start and the end of each putative coding region in the *A. thaliana* genome. Sequences upstream of the gene coordinates that indicate the start of a putative coding region were automatically extracted, using a Perl script designed in-house (Riaño-Pachón, unpublished data). Upstream sequences were extracted up to the next stop codon (according to the gene coordinates). In Figure 4-1 intergenic regions are represented in magenta, and coding regions as green boxes.

Additionally to the intergenic sequences, 1 kb upstream sequences were extracted, using a modified version of the Perl script mentioned above (In Figure 4-1, 1 kb up-stream sequences are represented as blue dots).

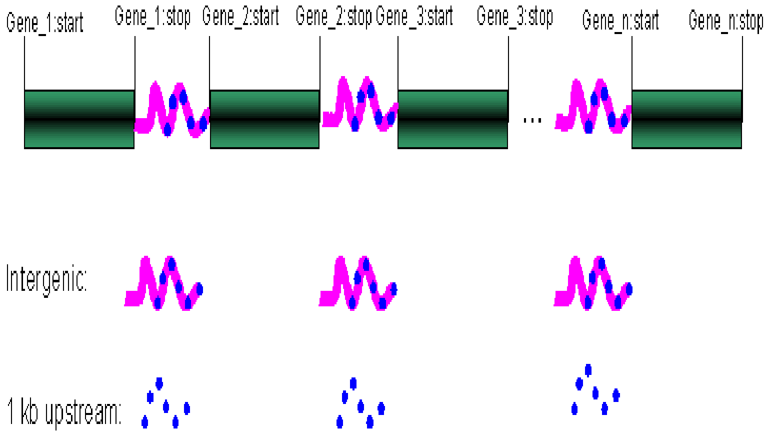


**Figure 4-1: Schematic illustration of gene arrangement in protein coding genes.** A gene comprises a coding region and an upstream sequence. Coding regions are represented as green boxes and upstream sequences as loops in magenta. All sequences between two coding regions were considered to be intergenic sequences (magenta loops). 1 kb upstream sequences are DNA sequences extending 1 kb upstream of the translation start of a coding region (blue dots). Translation start and stop positions were located using the gene coordinates provided by TIGR.

[2] ftp://ftp.tigr.org/pub/data/a_thaliana/

### 4.1.2 Nucleotide and oligonucleotide composition

The single nucleotide composition of every extracted sequence (intergenic and 1 kb upstream) was counted using a Perl script (Riaño-Pachón, unpublished data). All nucleotides in a given sequence were counted, including ambiguities as defined by the Nomenclature Committee of the International Union of Biochemistry – IUPAC.

For the 1 kb upstream sequences, oligonucleotide composition was assessed using the program Compseq from EMBOSS[89] run locally. Default parameters for the program were the following:

```
compseq –sequence [database_file] -reverse –word [size] –outfile [output_file]
```

All possible oligonucleotides of a given size, the number of occurrences and observed frequency in the set of analysed sequences were reported in the output file. Expected frequencies were calculated according to:

$$E = \prod_{i=1}^{w} f_{Xi} \tag{4-1}$$

where $f_{Xi}$ denotes the frequency of base $Xi$.

A ratio of representation was calculated with Equation(4-2):

$$repr = \frac{O - E}{E} \tag{4-2}$$

where $O$ refers to the observed frequency of a given oligonucleotide, and $E$ refers to the expected frequency of the oligonucleotide. Finally, oligonucleotides were grouped together with their reverse complement using a Perl script (Riaño-Pachón, unpublished data).

### 4.1.3 List of known *cis*-elements

*Cis*-elements published in PLACE[3][49] and AGRIS[4][26] were downloaded. In AGRIS, 95 plant-specific *cis*-elements are published, and in PLACE 427 *cis*-elements are published. Considering that the same *cis*-element could have different names in each database, each downloaded *cis*-element was aligned with the other downloaded entries for the detection of duplicated elements (pairwise alignments). In alignments with a similarity of 100% (defined by the identity of the aligned sequences), one of the entries was deleted. The preliminary list of *cis*-elements, from which duplicated entries were deleted contained 429 *cis*-elements (406 from PLACE and 23 from AGRIS). Subsequently, the list of *cis*-elements was further inspected to exclude:

[3] http://www.dna.affrc.go.jp/PLACE/

[4] http://arabidopsis.med.ohio-state.edu/AtcisDB/index.jsp

(i) *Cis*-elements that were longer than 10 nucleotides, and

(ii) *Cis*-elements that were described too vaguely by the nucleotide ambiguity code. A *cis*-element was discarded when it corresponded to more than 16 different oligonucleotides.

The list of known *cis*-elements was reduced from 429 to 198 *cis*-elements, with sizes from 4 to 10 nucleotides. Fifty *cis*-elements in the final list were described with ambiguities.

The list of *cis*-elements with 198 entries did not represented 198 different regulatory elements from plants. *Cis*-elements with similar names were regarded as functionally related, and grouped into subcategories. Each subcategory was composed of 2 to 12 *cis*-elements. In total 28 subcategories of regulatory elements were established (Table 4-1).

**Table 4-1: Groups of similar *cis*-elements.** *Cis*-elements downloaded from PLACE and AGRIS were grouped according to the name of the *cis*-element

| Group | Entries |
|---|---|
| ABRE | 12 |
| MYB | 12 |
| DRE/LTRE | 6 |
| G-box | 6 |
| TATABOX | 6 |
| AMMORES | 3 |
| CEREGLUBOX | 3 |
| E2F | 3 |
| I-box | 3 |
| POLASIG | 3 |
| W-box | 3 |
| -300ELEMENT | 2 |
| AUXRET | 2 |
| CAAT-box | 2 |
| CCA | 2 |
| GARE | 2 |
| GATA | 2 |
| GT1 | 2 |
| HSE | 2 |
| OCTAMER | 2 |
| PYRIMIDINEBOX | 2 |
| RY | 2 |
| S1F | 2 |
| SURE | 2 |
| TATCCA | 2 |
| TGA | 2 |

When the ambiguities in the *cis*-elements were replaced by the corresponding nucleotides, the list with 198 entries corresponded to 406 unique oligonucleotides.

The complete list of known *cis*-elements used is available online under:
www.bio-uni-potsdam.de/jgomez/dbcis.html

### 4.1.4 Frequency matrices and consensus sequences

The *cis*-elements ABRE, DRE, CE1, CE3, MYB and As1 have been experimentally confirmed to be in ABA-responsive genes in different plant species, including *A. thaliana*, rice (*Oryza sativa*), maize (*Zea mays*), barley (*Hordeum vulgare*) and tomato (*Lycopersicon esculentum*). To obtain accurate Position-Specific Frequency Matrices (PSFM) for each *cis*-element, the binding site sequences described in the literature were collected and aligned. To define the consensus sequences, the most prominent base or combination of bases at each position was chosen. The consensus sequences are shown in Table 4-2.
PSFMs for each *cis*-element were deduced from the number of occurrences of each nucleotide at each position. The different binding sites used to construct the matrices for the *cis*-elements ABRE, As1, CE3, DRE and MYB, and the matrices themselves are presented in Appendix 1. The binding sites and the PSFM of the *cis*-element CE1 is presented in section **6.1**.

**Table 4-2: Consensus sequences deduced for ABA-related *cis*-elements.** The corresponding binding sites are presented in Appendix 1 or in section 6.1

| Element | Consensus sequence |
|---|---|
| ABRE | NRYACGTGTM |
| AS1 | TDACGTAA |
| CE1 | SSBCACCSV |
| CE3 | SMCGCSTCGCY |
| DRE | KACCGACMT |
| MYB | MYWAACCA |

### 4.1.5 Pattern-based search

The generated consensus sequences were used to screen *A. thaliana* 1 kb upstream sequences, to identify pairs of ABA-related *cis*-elements using a Perl script designed in-house (Riaño-Pachón, unpublished data) that uses the program fuzznuc from EMBOSS[89]. Pair-wise combinations of *cis*-elements within a maximal distance of 1 kb were localized in the query sequence(s) with fuzznuc, and the output was compiled to deliver three kinds of lists:

1. Pair-wise combinations of *cis*-elements and number of hits found per pair in the screened sequences.

2. Sequence name (or identifier) in which each pair of *cis*-elements was found.
3. Distance between the *cis*-elements of a pair.

## 4.1.6 Matrix-based screening

PSFM were used to screen *A. thaliana* 1 kb upstream sequences. The programs MotifScanner[4,112] and CISTER[34] were used for the screening.

### 4.1.6.1 MotifScanner

MotifScanner was implemented by Gert Thijs at the Catholic University of Leuven[111,115], to localize known transcription factor binding sites in a query sequence. Each short sequence of length x in a query sequence is scored based on a motif and a background model. The motif model was represented by the PSFM (Θ), the background model (Bm) was a 2nd order Markov model, which was estimated from *A. thaliana* intergenic sequences (the Araset published by Pavy *et al.* 1999[80,113]). The probability that the sub-sequence of length x was generated from the background model was calculated, and also the probability that the sub-sequence was generated by the motif model.

Based on these two probabilities, a log-ratio score was calculated with:

$$W(x) = \log\left(\frac{P(x \mid \Theta)}{P(x \mid S, Bm)}\right) \tag{4-3}$$

### 4.1.6.2 CISTER

The algorithm CISTER was implemented by Martin Frith at the Department of Biomedical Engineering in Boston (USA)[34]. A Hidden Markov Model (HMM) is used to detect *cis*-element clusters in a query sequence. The transition probabilities represent the prior expectations concerning the distribution of the *cis*-elements in a query sequence. Emission probabilities describe the nucleotide preferences at each position in the *cis*-element versus the nucleotide preferences in the query sequence (counted around a window size defined by the user).

For the transition probabilities the model assumes that the distance between clusters is geometrically distributed with mean g. The model expects to see any *cis*-element on either strand with equal probability. The distance between motifs in a cluster is modeled as a geometric distribution with mean a, and the number of *cis*-elements in a cluster is supposed to be geometrically distributed with mean b.

For the emission probabilities the nucleotide preferences at each position in the *cis*-element (PSFM), and the background emission probabilities counted around a window around the position of the segment being scanned in the query sequence are compared. The probability of an instance is calculated by multiplying the transition and emission probabilities.

Thus, the parameters defined by the user are:

g for the mean distance between clusters,

a for the mean distance between elements in a cluster,

b for the mean number of elements in a cluster,

w is the window size, nucleotide frequencies are counted around it.

To fix the parameters, the program was run with a small training set, containing upstream sequences from different plant species, including *A. thaliana*. The exact location of the TATA box and other elements under study was known. The screening included the identification of the position of the TATA-box and ABA-related *cis*-elements in the training set (the TATA-box PSFM was downloaded from the web application of CISTER). After different trials, the following parameters recognized the correct position of about 70% of the *cis*-elements present in the training set:

g=1000, a=20, b=10 and w=150

The screenings were made using these parameters.

## 4.2 Molecular biology

### 4.2.1 Sterilization of seeds

Wild-type *A. thaliana* seeds were sterilized with 70% ethanol. After 2 min of incubation in ethanol seeds were centrifuged at 380 ***g*** for 1 min. On a clean bench the ethanol was discarded, and 1.5 mL sodium hypochloride (1:5 v/v dilution) and ca. 20 μL of 0.02% Triton X-100 were added. Seeds were mixed gently and kept at room temperature for 8 min. Afterwards, seeds were centrifuged at 380 ***g*** for 2 min, the supernatant was discarded, and 1.5 mL sterile water and ca. 20 μL 0.02% Triton X-100 were added. Seeds were centrifuged at 380 ***g*** for 2 min. Rinsing with sterile water was repeated three times. Seeds were dried for 3-4 hours, on a sterile filter paper prior to transfer to MS agar medium.

### 4.2.2 Hormone treatment

*A. thaliana* plants were grown hydroponically in 50% MS-medium supplemented with 2% sucrose. Four-week old plants were stimulated with 100 μM cis(+) ABA added to the fresh medium. Control plants were treated with an equivalent amount of 1 N NaOH used to dissolve ABA. Leaf samples were harvested 30, 60, 90, 120 and 300 min after addition of ABA or NaOH into the medium, and immediately frozen in liquid nitrogen for further analysis.

### 4.2.3 Isolation of RNA

Total RNA from control and ABA-treated plants was isolated using TRIZOL reagent. Leaf tissue was ground to a fine powder in liquid nitrogen, homogenized in 1.5 mL TRIZOL, and incubated at room temperature for 5 min. The samples were centrifuged at 16000 ***g*** for 10 min at 4°C to remove cell debris. The supernatant was carefully collected in a new tube and 400 µL chloroform were added. Samples were mixed by vortexing for 15 s, incubated at room temperature 5 min, and then centrifuged at 9500 ***g*** for 15 min at 4°C to separate phases.

The upper aqueous layer was carefully removed to a clean tube and RNA was precipitated with 0.6 volumes of isopropanol and 0.1 volumes of 3 M sodium acetate. Samples were left overnight at –20°C. To pellet the RNA the samples were centrifuged at 13600 ***g*** for 15 min at 4°C. The pellet was washed with 500 µL 70% ethanol, and pelleted again by centrifugation. The pellet was allowed to dry for 5 to 10 min at room temperature, and dissolved in 50 µL $H_20$-DEPC. Samples were stored at –20°C.

### 4.2.4 Spectrophotometric determination of RNA concentration

The RNA concentration was measured at 260 nm wavelength. For the measurement 1µL of RNA was dissolved in 99 µL $H_20$-DEPC. An $OD_{260nm}$ of 1 corresponds to a concentration of 40 µg/mL of single-strand RNA. The ratio between $OD_{260nm}$ and $OD_{280nm}$ provides and estimation of the purity of RNA. Pure preparations of RNA have an $OD_{260nm}/OD_{280nm}$ of 2.0[94].

### 4.2.5 DNase I digestion

Prior to the transcription of RNA into cDNA, RNA was digested to destroy traces of contaminating genomic DNA. In a final volume of 50 µL approximately 10µg of total RNA were incubated with RNase-free DNase I at 37°C for one hour. The enzyme was inactivated by heating at 75°C for 10 min. Absence of genomic DNA was confirmed by PCR using the pair of primers Actin 200/600, designed on an intronic sequence of actin.

After digestion RNA was purified with EGTA, LiCl and glycogen. For the purification of RNA 6 µL 20 mM EGTA, 5,2 µL 8 M LiCl and 7.4 µL glycogen (5 mg/mL) were added. The mixture was incubated at –20°C for 20 min, 300 µL 100% ethanol were added, and samples were incubated for 1 hour at –20°C. Samples were centrifuged at 13600 ***g*** for 30 min at 4°C. Precipitated RNA was rinsed with 70% ethanol-DEPC, and dissolved in 30 µL $H_2O$-DEPC.

### 4.2.6 cDNA synthesis

RNA was transcribed *in-vitro* into cDNA using the first-strand cDNA synthesis kit from Amershan, following the instructions of the manufacter.

Each reaction was performed in 33 µL, containing about 5 µg total RNA, 200 ng oligo dT (18mer), 11 µL bulk reaction mix (contains FPLCpure$^{TM}$ murine reverse transcriptase, RNAguard$^{TM}$, RNase/DNase free BSA and dNTPs in aqueous buffer), and 1µL 0.2 M DTT. After 1 hour of incubation at 37°C enzyme was inactivated at 70°C for 15 min.

### 4.2.7 Polymerase chain reaction (PCR)

To check for genomic DNA contaminations, samples were checked by PCR using the pair of primers actin 200/600 that amplified two fragments, a 200 bp long fragment when cDNA is used as template, and a 600bp long fragment when genomic DNA is used as template.

Each reaction was performed in a volume of 20 µL, containing 1µL template cDNA, 2 µL 10x buffer-$MgCl_2$ free, 2 µL 25 mM $MgCl_2$, 15 pmol forward and reverse primer, 0.4 µL dNTPs (10mM each), 1µL self-made *Taq*-polymerase and water.

DNA amplification was performed in a Robocycler (Stratagene, Heidelberg). A typical protocol for PCR is shown in Table 4-3.

**Table 4-3: Typical PCR protocol**

| Denaturation | Annealing | Polymerisation | Cycles |
|---|---|---|---|
| 94°C – 5 min | | | Initial cycle |
| 94°C – 30sec | 65°C – 30 sec (-1°C/cycle) | 72°C – 90 s | 9 cycles |
| 94°C – 30sec | 55°C – 30 sec | 72°C – 90 s | 20 cycles |
| | | 72°C – 5 min | Final cycle |

### 4.2.8 Real-time PCR (RT-PCR)

cDNA samples that amplified only the 200 bp long fragment with the pair of primers actin 200/600 were used for RT-PCR. Amplification was performed in a 7300 RT-PCR System (Applied Biosystems, Darmstadt), using SYBR® Green to monitor double strand DNA synthesis, in a 96-optical reaction plates. Reactions contained 5 µL 2x SYBR® Green PCR-Master Mix reagent, 1µL 1:10 dilution of template cDNA, and 500 nM of each gene-specific primer in a final volume of 10 µL. To minimize pipetting errors, and ensure that each reaction contained an equal amount of cDNA and primers, an electronic Eppendorf pipette (Eppendorf, Hamburg) was used to pipette the cDNA and the primer mix, while 2x SYBR® Green reagent was aliquoted with a 8-well multichannel pipette (ABIMED GmbH, Langenfeld).

The following standard thermal profile was used for all PCRs: 50°C for 2 min, 95°C for 10 min, 40 cycles at 95°C for 15 s, and 60°C for 1 min. Data acquisition was made using the SDS 1.2.2 software (Applied Biosystems, Darmstadt).

The level of cDNA was assessed by comparison with values obtained for a control gene (*Ubiquitin 10 – At4g05320*). The $\Delta C_t$ method was used for relative quantification, and values were expressed as $2^{-\Delta\Delta Ct}$[81].

In Figure 4-2 a typical amplification plot is shown. Note the increase in SYBR® Green fluorescence with increasing PCR cycle number. For most of the samples the slope of the curves was quite similar, reflecting similar amplification efficiencies.

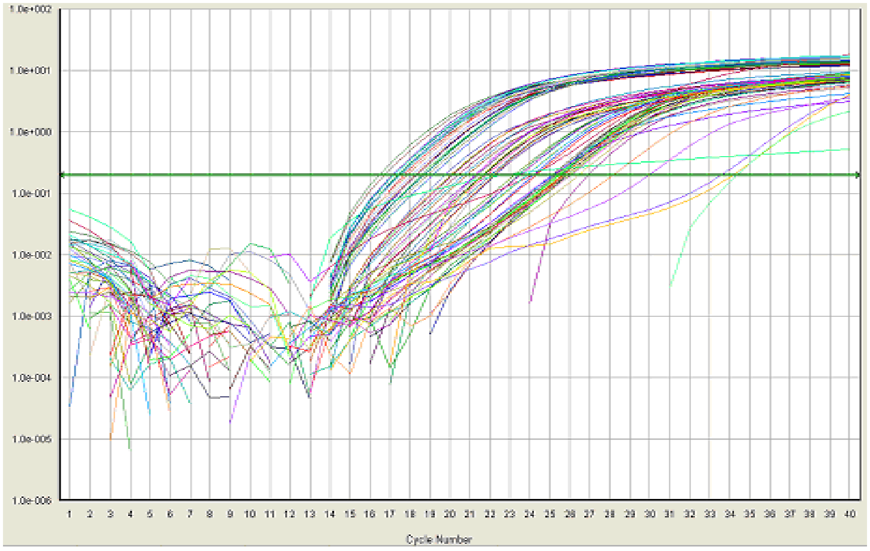


**Figure 4-2: Amplification plot of a RT-PCR.** The cycle number is indicated at the x-axis, the y-axis denotes SYBR® Green fluorescence (log scale)

For quantification an arbitrary threshold cycle was defined ($C_t$), that indicates the fractional cycle number at which the targets and the reference gene reached a fluorescence level of 0.2 (marked green line in Figure 4-2). $\Delta C_t$ is the difference in threshold cycles for target and reference of samples subjected to the same treatment ($\Delta Ct = C_{target} – C_{reference}$). To compare treated and untreated samples the $\Delta\Delta C_t$ value was calculated, which describes the difference between $\Delta C_t$ values in treated (T) and control (C) samples ($\Delta\Delta C_t = \Delta C_{t\text{-}C} - \Delta C_{t\text{-}T}$). Relative changes in expression with an amplification efficiency of 1 are given by $2^{-\Delta\Delta Ct}$. If $2^{-\Delta\Delta Ct}$ is equal to 1, there is no difference in the expression level upon treatment. A $2^{-\Delta\Delta Ct}$ larger than one indicates up-regulation of gene expression upon treatment, whereas a $2^{-\Delta\Delta Ct}$ lower than 1 indicates down-regulation of gene expression upon treatment. *Actin* was used to test the reliability of the amplification experiment. It is supposed that the expression of *actin* and other housekeeping genes is not largely influenced by the treatment, and the $2^{-\Delta\Delta Ct}$ must be close to one.

### 4.2.9 DNA array data analysis

Nylon membrane data kindly provided by Dr. Magdalena Ornatowska[78] were used to determine genes differentially expressed after ABA treatment. Gene expression in *A. thaliana* leaves was monitored at 30, 60, 90, 120 and 300 min after ABA or control treatment.

After data acquisition measured radioactivity was normalized by the software Haruspex developed at the Max-Planck Institute for Molecular Plant Physiology in Golm - Germany (S. Kloska, B. Essigmann and T. Altmann, unpublished data). Normalization included the subtraction of local background and indication of values below threshold level (set-up to twice local background). Haruspex calculated a gene activity value, corresponding to the ratio between the signal measured in a complex hybridisation and the signal measured in a reference hybridisation.

The analysis of the data started with the selection of clones that had been measured successfully in all experiments (control and treatment, all data points). Subsequently, clone names were substituted by the gene identifiers established by the Arabidopsis Genome Initiative (regarded as AGI codes)[1].

The software GEPAS (http://gepas.bioinfo.cnio.es/cgi-bin/preprocess)[48] was used to detect and delete inconsistent replicates. A replicate measurement was considered inconsistent if the gene activity was two-times above/below the standard deviation of the other measurements for the same gene. In the case of genes with at least three valid measurements, missing values were calculated using KNN, run locally[120]. Replicates were merged by the mean.

A ratio of expression was calculated as gene activity of gene *i* measured in treatment divided by gene activity of gene *i* measured in control plants. Ratios were scaled by logarithmic transformation ($Log_2$), where 0 $Log_2U$ stands for no changes. Values greater than 0 denote positive changes (measured gene activity was higher in treated plants), and values smaller than zero denote negative changes (measured gene activity was lower in treated plants).

To select differentially expressed genes the recommendation of Thimm *et al.* 2001[116] was followed. A gene was considered to be regulated if it showed at least a three-fold change in gene expression ($\pm1{,}58$ $Log_2U$).

# Chapter 5: Analysis of intergenic sequences

*During the past years many genomes have been fully sequenced. For the genomic sequence annotations several prediction programs have been used to deduce gene models, and similarity searches have helped to assign protein functions[80]. Nevertheless, the proportion of predicted genes for which no functional annotation has been found still large. For example, five years after the publication of the genome sequence of the model plant A. thaliana about half of the genes currently do not have any definitive functional annotation[137]. One fundamental problem that remains is the identification of genes that respond to a given external or endogenous stimulus, especially if no functional annotation is available for these genes.*

*In this study, the genome sequence of the model plant A. thaliana was used to predict genes putatively regulated by the phytohormone abscisic acid (ABA). The major role of the combinatorial action of transcription factors was explored. Putative target genes were identified by the identification of regulatory cis-elements in their intergenic regions.*

*Given the importance of the selected intergenic sequences in the quality of the predictions, general features of the intergenic regions of A. thaliana were studied. In this chapter the results concerning nucleotide composition, intergenic region lengths and oligonucleotide composition of intergenic regions are presented.*

## *5.1* *A. thaliana* intergenic regions

The analysis of DNA sequences from bacteria (e.g. *Rhizobium meliloti*, *Agrobacterium tumefaciens*, *Escherichia coli*), human viruses (e.g. EBV, CMV, Vaccinia), and eukaryotes (e.g. *Saccharomyces cerevisiae*, *Neurospora crassa*, *Zea mays*, *Homo sapiens*) has shown that the nucleotide composition changes along different regions of the genome[13]. For instance, the significant differences between non-coding regions and coding regions are exploited by all gene prediction methods[80].

The selection of the DNA sequences for the identification of **T**ranscription **F**actor **B**inding **S**ites (TFBS) is driven by the typical location of the binding sites. In *A. thaliana* experimental and computational approaches have shown that TFBSs are mostly located in intergenic regions[2,35,75,107,111]. The intergenic region in such cases was defined as the non-coding region between two genes. Nevertheless, some exceptions are known were regulatory sites are located in introns, conferring tissue specific expression[51,91].

The sizes of the intergenic regions depend on the compactness of the genome[111]. There is not a detailed report about lengths and nucleotide compositions of intergenic regions from *A. thaliana* performed on the whole genome, which has been available since 2000 [107].

The work presented here utilizes the whole genome sequence of *A. thaliana* to establish default parameters to be used with TFBS-predicting algorithms, by carefully analysing the lengths and single nucleotide compositions of intergenic regions.

In this study, the DNA sequence that extends upstream from a coding region to the end of the preceding coding region was considered as the intergenic region of the mentioned gene (Figure 4-1). An in-house designed Perl script (Riaño-Pachón, unpublished data) was used to automatically extract such sequences, using the gene coordinates published in the last genome release in January 2004[5]. About 30.000 (29885) intergenic regions of *A. thaliana* ecotype Columbia-0 were retrieved, and their lengths and single strand nucleotide compositions were determined.

As already found for other genomes[13,123], intergenic regions in *A. thaliana* exhibit a strong A+T over C+G bias. According to the single nucleotide composition counted over a single strand (extracted sequences) the overall frequency of A was equal to the overall frequency of T, equal to 0.34 ($f_A$ = $f_T$ = 0.34). The overall frequency of C was equal to the overall frequency of G, equal to 0.16 ($f_G$ = $f_C$ =0.16). Undetermined bases (N) account for only 0.33% of the total. The average %GC in intergenic regions was 32.4.

Few sequences (12) showed a low A +T frequency (A+T <0.4), and corresponded to genome regions not fully sequenced yet, that are close to centromeres and/or telomeres, or to extremely short intergenic regions (shorter than 30 bp).

Intergenic sequence lengths ranged from 1 base to 72 kb. Regarding the lengths it must be kept in mind that intergenic regions, mainly the longer ones, may include DNA that codes for non-protein-coding RNAs. This is usually overlooked during the process of genome annotation, where genome sequences are screened for the presence of relatively long open reading frames (minimum 50 amino acid residues)[80,88].

The mean length found for intergenic regions was 1.8 kb with a standard deviation of 2.5 kb. The median length was 1.1 kb. These results were substantially different from the results presented by Steffens *et al.* 2004, where the average length of *A. thaliana* intergenic sequences is regarded to range from 2 to 2.5kb[107]. Although no additional information is provided by Steffens and co-workers, it is likely that the deduced lengths might be based in calculations made with the data presented in the analysis of the genome[1], and not the result of a careful analysis like the one present here.

The length distribution of intergenic regions is shown in Figure 5-1. Approximately 54% of the intergenic regions retrieved have a length of at least 1 kb.

[5] ftp://ftp.tigr.org/pub/data/a_thaliana/

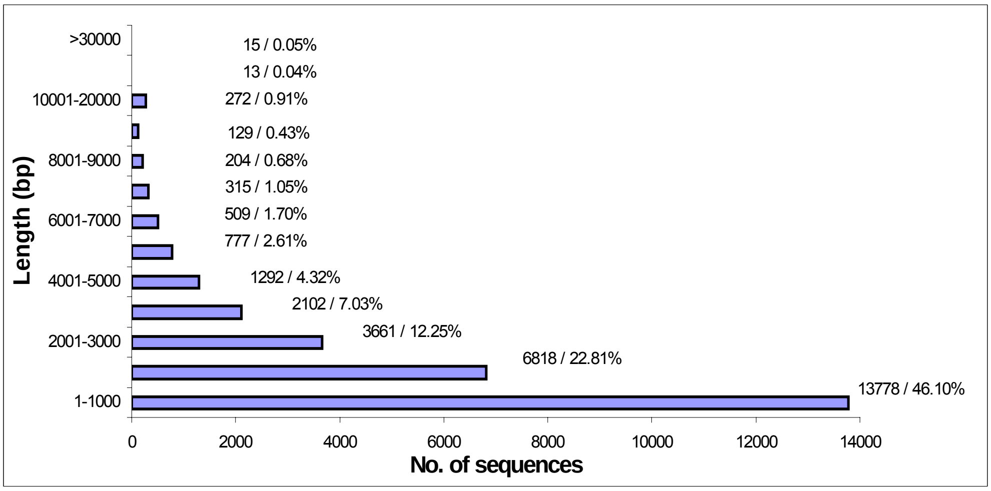


**Figure 5-1 Distribution of the lengths of intergenic regions in *A. thaliana*.** An intergenic region was regarded as the sequence upstream of a coding region (Figure 4-1). The start and stop of a coding region were defined by the gene coordinates provided by TIGR[6]

The wide range in lengths confirms the length polymorphism of intergenic regions of *A. thaliana*, originally outlined by Pavy *et al.* 1999[80] The work of Pavy and co-workers was prior to the publication of the genome sequence of *A. thaliana*[1], and focused on the evaluation of gene prediction software using annotated BAC sequences.

Nearly 10% of the extracted intergenic regions (2933 sequences) were shorter than 200 bp. Open Reading Frame (ORF) annotations showed that most of the corresponding genes were expressed proteins (618 genes), hypothetical proteins (133 genes), pentatricopeptide (PPR) repeat-containing proteins (85 genes), hypothetical proteins similar to pseudogenes (43 genes), copia-like retroposon family members (30 genes) and gypsy-like retroposon (Athila) family members (21 genes), among others.

For 300 intergenic regions, i.e. 1% of all extracted intergenic sequences, the length was more than 10 kb. As for the short intergenic regions, ORF annotations showed that most of the corresponding genes were hypothetical proteins (28 genes), expressed proteins (26 genes) gypsy-like retroposon family members (21 genes), gypsi-like retroposon (Athila) family members (19 genes), copia-like retroposon family members (12 genes) and pseudogenes, among others.

## 5.2 *Characterisation of 1 kb upstream sequences*

Most regulatory sequences in *A. thaliana* have been located experimentally upstream of the transcription start site[2,35,75,107,111]. However, for most *A. thaliana* genes no information of the exact transcription start site was available. Therefore, considering that the analysis of

[6] ftp://ftp.tigr.org/pub/data/a_thaliana/

intergenic regions revealed a median length of 1.1. kb, the computational prediction of TFBSs was made using 1 kb long sequences upstream of the start of a coding region. Similar as for intergenic regions, the exact location of the start of a coding region was assessed using the gene coordinates of the last release of the genome sequence (January 2004).

Normally, putative regulatory regions have been located in *A. thaliana* within 500 bp and 1 kb upstream of the transcription or translation start point (e.g., some putative regulatory sequences responsible for auxin and brassinosteroids responsiveness[39], putative MYB binding sites involved in the up-regulation of genes in plants overexpressing AtMYB2[2], and regulatory sequences involved in the regulation by cold stress[46]). The default length of 1 kb was applied in all cases, including intergenic regions smaller than 1 kb, because it was not discarded that some coding sequences exert regulatory actions on a neighbouring gene, moreover if the intergenic sequence of the neighbouring gene is too short.

After retrieving the set of upstream sequences, the single nucleotide compositions of the sense DNA strands were analysed. The prevalence of A+T and C+G was slightly changed compared with the results for whole intergenic regions ($f_A$ = $\mathrm{f}_T$ = 0.32 and $f_G$ = $f_C$ =0.18), but the bias towards an increased A+T content was maintained.

Forty sequences were excluded from further analysis, because at least 50% of the sequence was composed by undetermined nucleotides (N).

### 5.2.1 Oligonucleotide composition

The specificity of a transcription factor for a binding site arises from the specific interaction between the DNA binding domain of the transcription factor and the DNA sequence at the binding site[114]. Especially in eukaryotes the sequences recognized by the transcription factors are generally short and variable[123]. In yeast for instance, the number of well-conserved bases in a collection of binding sites of a single transcription factor is typically six to ten[129].

The set of *A. thaliana* 1 kb upstream sequences was analysed with respect to their oligonucleotide composition, under the assumption that sequences with a regulatory role cannot be abundantly distributed throughout the genome, if precise control of gene-expression is to be achieved. The analysis was conducted with a particular interest to find under-represented oligonucleotides.

The number of occurrences of all possible oligonucleotides of a given size (w=2 to w=10) were counted on both strands of all 1 kb upstream sequences extracted, using the program compseq from EMBOSS[89].

The output of the program was the number of occurrences of every possible oligonucleotide (occ) and the frequency of observation (O). The frequency of observation was the number of occurrences of each oligonucleotide divided by total number of oligonucleotides possible of the given size in the set of sequences evaluated.

Under-representation of an oligonucleotide was assessed with respect to a statistical background model. The background model was the expected frequency of each oligonucleotide (E), based on the single-nucleotide frequencies ($f_A$, $f_C$, $f_G$, and $f_T$) (Equation (4-1) – section 4.1.2).

Observed and expected frequencies (O and E, respectively) were compared, and a ratio of representation was defined according to Equation (4-2) (section 4.1.2)

$$repr = \frac{O - E}{E}$$

According to the ratio of representation defined, under-represented oligonucleotides were those with negative values, the maximum under-representation score was –1, defining oligonucleotides not observed in the set of evaluated sequences. Oligonucleotides with *repr* values close to 0 do not occur more or less frequently than expected according to the background model, whereas positive values account for over-represented oligonucleotides, these values can tend to infinity.

Considering that the number of expected oligonucleotides of increasing size grows exponentially, and that a transcription factor does not discriminate between binding sequences on the sense or reverse strand of a DNA fragment, the set of oligonucleotides was reduced by grouping each oligonucleotide with its reverse complement.

In the case of oligonucleotides of an even-numbered size (w), the number of palindromes of size **w** was calculated as $n_{pal}=4^{w/2}$[123]. Therefore, the number of unique oligonucleotides can be calculated according to Equation (5-1):

$$U = \frac{4^w}{2} + \frac{n_{pal}}{2} \tag{5-1}$$

Table 5-1 shows the number of unique oligonucleotides, number of palindromes and number and percentage of oligonucleotides found under-represented in *A. thaliana* 1 kb upstream sequences.

It was observed that at least 50% of the oligonucleotides of any size were slightly to highly under-represented in *A. thaliana* 1 kb upstream sequences (Table 5-1). Additionally, the linear regression coefficient ($R^2$=0.82) between oligonucleotide size and percentage of under-represented oligonucleotides indicated that there was a strong linear correlation between both.

When the corresponding equation of the regression curve (y=1.3167x+49.989, where x is oligonucleotide size) was used to calculate the percentage of under-represented oligonucleotides with size w=11, the result of 64% was very close to the observed result (65%, data no shown).

**Table 5-1: Under-represented oligonucleotides in *A. thaliana* 1 kb upstream sequences.** w=oligonucleotide size, u=number of expected unique oligonucleotides, $n_{pal}$=number of palindromes, under-represented=number of oligonucleotides found under-represented, percentage in brackets

| w | u | $n_{pal}$ | under-represented |
|---|---|---|---|
| 2 | 10 | 4 | 5 (50) |
| 3 | 32 | 0 | 17 (53) |
| 4 | 136 | 16 | 79 (58) |
| 5 | 512 | 0 | 298 (58) |
| 6 | 2080 | 64 | 1236 (59) |
| 7 | 8192 | 0 | 4941 (60) |
| 8 | 32896 | 256 | 19638 (59) |
| 9 | 131072 | 0 | 80969 (62) |
| 10 | 524800 | 1024 | 325888 (62) |

From the data it can be concluded that at least for the analysed oligonucleotide sizes, the percentage of under-represented oligonucleotides increases linearly with oligonucleotide size. However, because no data are available for longer oligonucleotide sizes, and the calculation of the number of occurrences of all possible oligonucleotides of longer sizes is computationally exhaustive, it cannot be established up to what extent the linear behaviour can be extrapolated.

The analysis of under-represented oligonucleotides revealed that up to oligonucleotide size w=8 none of the possible oligonucleotides obtained a ratio of representation equal to –1 (meaning that the given oligonucleotide was not found in the set of *A. thaliana* 1 kb upstream sequences). For oligonucleotide size w=9, 0.02% of the possible oligonucleotides were not observed (20), and in the case of oligonucleotide size w=10 the percentage increases up to 0.92 (4806 oligonucleotides). It is expected that the number of observed oligonucleotides of a given size decreases when oligonucleotide size increases. It is also expected that the correlation between both variables do not show a linear correlation.

For dinucleotides (w=2), five from ten possible oligonucleotides were under-represented (see Table 5-1). Among them, the dinucleotides TA and CG showed the lowest scores (-0.25 and –0.28, respectively), the other 3 under-represented dinucleotides showed negative scores which were close to zero. Therefore, their occurrence did not differ markedly from the background model (single-nucleotide frequencies).

Both dinucleotides that were highly under-represented (TA and CG) were previously found under-represented in other eukaryotic genomes[13]. It has been speculated that the under-representation of the dinucleotide TA in intergenic regions that are A+T rich might be directly related to the fact that AT or TA containing oligonucleotides possibly adversely affect supercoiling and/or chromatin structure[13]. From another perspective, taking into account the prominent role of the basal regulatory *cis*-element TATA-box, the suppression of TA might minimize the inappropriate binding of general transcription factors.

It has been found that the dinucleotide CG has a relatively normal occurrence in organisms were no methylase activity has been reported (e.g. *Drosophila melanogaster* or *Caenorhabditis elegans*)[13]. In eukaryotes were methylation of nucleotides has been reported, the methylation of the cytosine in the dinucleotide CG is considered a regulatory mechanism allowing the suppression of gene activity[64]. For that reason, the under-representation of the dinucleotide CG in *A. thaliana* might be directly related to the methylation machinery.

Among the trinucleotides (w=3), GTA(TAC), ACG(CGT) and CGC(GCG) showed the lowest scores. These sequences represent extensions of the already mentioned under-represented dinucleotides. The trinucleotides CNG, that play an important role in plants because the cytosine is often subjected to methylation[19], were also under-represented. The oligonucleotide CAG(CTG) was found just slightly under-represented (-0,002), compared with a score of –0,16 for the oligonucleotide CCG(CGG).

The strong under-representation of the trinucleotide CCG(CGG) compared with the trinucleotide CAG(CTG) cannot be explained simply by its putative role in DNA methylation, unless it can be proven experimentally that it is more often subjected to methylation than the other.

In eukaryotes, sequences with regulatory roles are between 4 and 10 nucleotides. To test whether such sequences were under-represented on a genome-wide scale, the ratio of representation of oligonucleotides with size w >= 4 were compared with already described *cis*-elements.

The evaluated *cis*-elements represented a subset of 198 *cis*-elements retrieved from the databases PLACE[7] and AGRIS[8]. The 198 *cis*-elements corresponded to 406 unique oligonucleotides. The differences between the number of *cis*-elements and unique oligonucleotides is due to the fact that 50 *cis*-elements are described using the nucleotide ambiguity code. For instance, one of the *cis*-elements describing a MYB binding site was MWCCWAMC. This transcription factor binding site corresponds to sixteen different oligonucleotides, since M represents the nucleotides A and C, and W represents the nucleotides A and T.

The occurrence of the 406 oligonucleotides was then analysed in *A. thaliana* 1 kb upstream sequences. In Table 5-2the total number of oligonucleotides representing *cis*-elements is

[7] http://www.dna.affrc.go.jp/PLACE/

[8] http://arabidopsis.med.ohio-state.edu/AtcisDB/index.jsp

displayed for each oligonucleotide size (T). Likewise, the number and percentage of these oligonucleotides found under-represented in the set of *A. thaliana* 1 kb upstream sequences are given.

**Table 5-2: Under-represented oligonucleotides in *A. thaliana* 1 kb upstream sequences that corresponded to known *cis*-elements.** w=oligonucleotide size, UR=Under-represented oligonucleotides, T=total number of oligonucleotides of the given size representing *cis*-elements, %=(UR/T)*100

| w | UR/T | % |
|---|---|---|
| 4 | 5/7 | 71,43 |
| 5 | 8/12 | 66,67 |
| 6 | 37/75 | 49,33 |
| 7 | 36/70 | 51,43 |
| 8 | 47/100 | 47,00 |
| 9 | 32/56 | 57,14 |
| 10 | 30/86 | 34,88 |

From 406 unique oligonucleotides representing the 198 regulatory *cis*-elements, 195 were found under-represented. Table 5-2 shows that when the size of the *cis*-element increases, the number of under-represented oligonucleotides corresponding to the *cis*-element decreases, although the total number of under-represented oligonucleotides increases with oligonucleotide size (Table 5-1). Nevertheless, nearly in every dataset at least half of the oligonucleotides corresponding to described *cis*-elements were found under-represented in 1 kb upstream sequences.

The list of *cis*-elements from plants does not refer to 198 different regulatory sequences, some of the *cis*-elements are described by more than one sequence. In those cases, these are *cis*-elements that were observed in different genes and/or plant species, or are different sequences of the same kind of *cis*-element in a given gene.

Based on the name of the *cis*-elements, regulatory sequences that refer to the same *cis*-element were grouped (section 4.1.3). As an example, Table 5-3 shows the three *cis*-elements that refer to the cereal glutenin box in the pea (*Pisum sativum*) *LEGA* gene, grouped as CEREGLUBOX.

**Table 5-3: *Cis*-elements retrieved from PLACE that refer to the cereal glutenin box. w=oligonucleotide size.** All elements were grouped as members of CEREGLUBOX based on the name of the *cis*-element

| w | *Cis*-element name | Sequence | Group |
|---|---|---|---|
| 9 | CEREGLUBOX1PSLEGA | TGTTAAAGT | CEREGLUBOX |
| 8 | CEREGLUBOX2PSLEGA | TGAAAACT | CEREGLUBOX |
| 9 | CEREGLUBOX3PSLEGA | TGTAAAAGT | CEREGLUBOX |

Twenty-six groups were established. The groups that contained the highest number of regulatory sequences for a given *cis*-element were (i) ABRE, and (ii) MYB, i.e. MYB factor binding site. For most of the groups only two regulatory sequences refer to the same *cis*-element. Nevertheless, most of the *cis*-elements in the list were represented by a single regulatory sequence (108 *cis*-elements). The groups established, number of regulatory sequences per group (entries), number of oligonucleotides corresponding to the given number of entries per group, and number of under-represented oligonucleotides in *A. thaliana* 1 kb upstream sequences are shown in Table 5-4. The last row in the table indicates the results for the 108 *cis*-elements represented by only a single regulatory sequence (entry).

**Table 5-4: Groups of known *cis*-elements established according to the name of the *cis*-element.** *Cis*-elements were retrieved from PLACE and AGRIS. UR=number of under-represented oligonucleotides in *A. thaliana* 1 kb upstream sequences

| Group | Entries | Oligonucleotides | UR |
|---|---|---|---|
| ABRE | 12 | 55 | 32 |
| MYB | 12 | 47 | 23 |
| DRE/LTRE | 6 | 7 | 7 |
| G-box | 6 | 6 | 4 |
| TATABOX | 6 | 6 | 4 |
| AMMORES | 3 | 4 | 3 |
| CEREGLUBOX | 3 | 3 | 2 |
| E2F | 3 | 5 | 1 |
| I-box | 3 | 4 | 3 |
| POLASIG | 3 | 3 | 1 |
| W-box | 3 | 3 | 1 |
| -300ELEMENT | 2 | 14 | 6 |
| AUXRET | 2 | 3 | 2 |
| CAAT-box | 2 | 2 | 2 |
| CCA | 2 | 3 | 0 |
| GARE | 2 | 2 | 2 |
| GATA | 2 | 5 | 5 |
| GT1 | 2 | 9 | 5 |
| HSE | 2 | 17 | 3 |
| OCTAMER | 2 | 2 | 2 |
| PYRIMIDINEBOX | 2 | 2 | 0 |
| RY | 2 | 3 | 0 |
| S1F | 2 | 2 | 2 |
| SURE | 2 | 2 | 1 |
| TATCCA | 2 | 3 | 3 |
| TGA | 2 | 2 | 1 |
| Other / one entry | 108 | 192 | 80 |

Table 5-4 shows that some groups with few entries (2 or 3) have a large number of corresponding oligonucleotides, due to the use of the nucleotide ambiguity code to describe the sequence of the TFBS (e.g. –300 element or HSE).

The results reported in Table 5-4 demonstrate that for seven groups of *cis*-elements all corresponding oligonucleotides were under-represented in *A. thaliana* 1 kb upstream sequences (DRE/LTRE, CAAT-box, GARE, GATA, OCTAMER, S1F, and TATCCA). For four groups of *cis*-elements nearly all oligonucleotides were under-represented (AMMORES, CEREGLUBOX, I-box and AUXRET). For three groups of *cis*-elements only a single nucleotide was under-represented (E2F, POLASIG and W-box). Finally, none of the oligonucleotides of the groups CCA, PYRIMIDINEBOX and RY9 were found under-represented in *A. thaliana* 1 kb upstream sequences. In general, in 17 of the 26 groups of *cis*-elements more than 50% of the oligonucleotides corresponding to *cis*-elements were found under-represented.

Exemplarily, a few detailed results are displayed in Table 5-5. All oligonucleotides corresponding to the *cis*-elements G-box and TATA-box are shown. The results for the G-box were chosen because G-boxes are regulatory sequences that are specifically found in plants. They have been described in promoters of genes that are responsive to a wide variety of stimuli such as light, or cumaric acid or abscisic acid[69]. The core sequence of a G-box is the tetranucleotide ACGT, and the specificity of the *cis*-element is achieved through the flanking nucleotides, and through the coupling with other *cis*-elements. An example of a G-box coupling element is the I-box found in the tobacco (*Nicotiana tabacum*) *RBCS8* promoter. The group G-box - I-box represents the shortest promoter capable of conferring light-responsiveness [121]. The results for the TATA-box were chosen because the TATA-box represents a basal regulatory element in eukaryotic promoters.

**Table 5-5: Ratio of representation of the oligonucleotides corresponding to G-box and TATA-box.** Ratios were calculated based on the observed occurrence of each oligonucleotide in *A. thaliana* 1 kb upstream sequences, according to Equation (4-2)

| Name | Oligonucleotide | Group | Ratio |
|---|---|---|---|
| ACGTABOX | TACGTA | GBOX | -0,540 |
| ACGTCBOX | GACGTC | GBOX | -0,270 |
| ACGTTBOX | AACGTT | GBOX | -0,300 |
| ACGTOSGLUB1 | GTACGTG | GBOX | -0,400 |
| GBOXLERBCS | CCACGTGGC | GBOX | 5,220 |
| GBOXPC | ACCACGTGGC | GBOX | 3,290 |
| TATABOX2 | TATAAAT | TATABOX | -0,150 |
| TATABOX3 | TATTAAT | TATABOX | -0,320 |
| TATABOX4 | TATATAA | TATABOX | -0,009 |
| TATABOX5 | TTATTT | TATABOX | 0,160 |
| TATABOXOSPAL | TATTTAA | TATABOX | -0,300 |
| TATABOX1 | CTATAAATAC | TATABOX | 0,060 |

Table 5-5 illustrates that four from six oligonucleotides corresponding to G-box sequences were under-represented in 1 kb upstream sequences. All under-represented oligonucleotides achieved high scores (-1 is the maximum score for under-representation). Over-represented oligonucleotides achieved also high scores. According to the documentation provided by PLACE, only one of the G-boxes was found originally in *A. thaliana*. One of the sequences over-represented corresponded to the G-box GBOXLERBCS, that was observed in the gene *RBCS* from tomato. The other (GBOXPC) was observed in parsley (*Petroselinum crispum*). Both genes are regulated by light.

The G-box originally described in *A. thaliana* was found under-represented (ACGTTBOX), and corresponds to a *cis*-element of the *RBCS-3A* gene[32]. The G-boxes ACGTOSGLUB1, ACGTABOX and ACGTCBOX corresponded to the genes *GLUB1* and *RAB16A* from rice. All the G-boxes from rice are involved in tissue specific gene expression. *GLUB1* in endosperm tissue[9] and *RAB16A* in vegetative and floral organ tissues[77].

Under or over-representation of oligonucleotides underlying G-boxes does not seem to be directly related to the organism where the transcription factor binding sites were originally found. However, regarding the over-represented oligonucleotides corresponding to light responsive elements in tomato and parsley, it was found that the sequences that confer light-responsiveness in *A. thaliana* were: (i) under-represented, and (ii) the nucleotides flanking the ACGT-core involved in the specificity of the *cis*-element were different from those found in the light responsive elements of parsley and tomato, as can be observed in Table 5-6.

**Table 5-6: Ratio of representation of the oligonucleotides corresponding to light responsive elements found in *A. thaliana*, tomato and parsley.** Ratios were calculated based on the observed occurrence of each oligonucleotide in *A. thaliana* 1 kb upstream sequences, according to Equation (4-2)

| Organism | *Cis*-element | Ratio |
|---|---|---|
| *A. thaliana* | GAC ACGT AGA | -0.06 |
| *A. thaliana* | A ACGT AT | -0.47 |
| Tomato | CC ACGT GGC | 5.22 |
| Parsley | ACC ACGT GGC | 3.29 |

Regarding the TATA-box, this *cis*-element might be considered *a priori* as a very frequent *cis*-element in intergenic regions due to its A+T rich content and its general role as a transcriptional regulator. It was found that in *A. thaliana* 1 kb upstream sequences four from six oligonucleotides corresponding to the TATA-box were under-represented, i.e. the sequences TATABOX2, 3, 4 and TATABOXOSPAL (see Table 5-5). The other two oligonucleotides that were over-represented achieved relatively low ratio of representation

[9] PLACE documentation (http://ftp.dna.affrc.go.jp/pub/dna_place/place.fasta)

(repr=0.160 and repr=0.060, respectively), and correspond to binding sites described in pea and in rice.

Also in the case of the TATA-box, none of the *cis*-elements listed corresponded to the TATA-box of *A. thaliana*, and the under-representation or slightly over-representation of the corresponding oligonucleotides might be related to the suppression of unspecific transcription initiation.

### 5.2.2 Oligonucleotide composition of ABA-related *cis*-elements

ABRE (for Abscisic Acid Responsive Element) and DRE/LTRE (for Dehydration Responsive Element / Low Temperature Responsive Element) are the two major *cis*-acting elements involved in the regulation of gene expression in response to osmotic stress in ABA-independent and ABA-dependent signalling pathways respectively[135].

In addition to these *cis*-elements, MYB binding sites and Coupling elements (CE1 and CE3) also play an important role in ABA-mediated gene expression[2,3,50,100,134]. The element As1 (Activation sequence 1) was also found in the ABA-responsive gene RD29A from *A. thaliana*[75].

Although the *cis*-element DRE/LTRE modulates ABA-independent gene expression in response to osmotic stress, cross-talk between different osmotic stresses like drought and high salinity has been documented. Additionally, it has been shown that DRE can also act as a coupling element of ABRE[75,105,135].

The list of 198 *cis*-elements (corresponding to 406 oligonucleotides underlying regulatory sequences) has 32 sequences that refer to the *cis*-elements ABRE, DRE, MYB, CE1 and CE3, no entries of the *cis*-element As-1 were found. The 32 entries correspond to 111 unique oligonucleotides, and their sizes range from w=5 to w=10 nucleotides.

As has been already shown, most of the *cis*-elements retrieved from PLACE and AGRIS corresponded to ABRE and MYB (see Table 5-4). Additionally, 6 entries for the *cis*-element DRE/LTRE were found (7 oligonucleotides corresponded to the 6 entries), and only one entry was found for the *cis*-elements CE1 and one for CE3.

From the 111 oligonucleotides underlying *cis*-elements involved in ABA-mediated gene regulation, 60 oligonucleotides (54%) were found under-represented in *A. thaliana* 1 kb upstream sequences.

All oligonucleotides corresponding to the 6 entries of the *cis*-element DRE/LTRE were under-represented, 32 out of 55 oligonucleotides corresponding to the 12 entries of the *cis*-element ABRE were under-represented, 23 out of 47 oligonucleotides corresponding to the 12 entries of the *cis*-element MYB were under-represented. In contrast, the oligonucleotide corresponding to the *cis*-element CE1 was not under-represented (repr=0.91).

For 13 out of 32 *cis*-elements involved in ABA-mediated gene regulation, the nucleotide ambiguity code has been used in the description of the binding sites. For these *cis*-elements the name of the *cis*-element, the number of corresponding oligonucleotides and the number of under-represented oligonucleotides in *A. thaliana* 1 kb upstream sequences are presented in Table 5-7.

**Table 5-7: Ratio of representation of oligonucleotides corresponding to ABA-related *cis*-elements for which the regulatory sequences have been described using the nucleotide ambiguity code.** The column "Ambiguities" denotes the number of corresponding oligonucleotides for the given *cis*-element. UR=number of oligonucleotides found under-represented in *A. thaliana* 1 kb upstream sequences

| *Cis*-Element | Group | Sequence | Ambiguities | UR |
|---|---|---|---|---|
| ABREMOTIF1 | ABRE | RYACGTGGC | 4 | 2 |
| ABREBZMRAB28 | ABRE | TCCACGTSKY | 8 | 0 |
| ABRE-like | ABRE | BACGTGKM | 12 | 7 |
| ABREOSRAB27 | ABRE | ACGTSSSC | 8 | 7 |
| MYB | MYB | MWCCWAMC | 16 | 7 |
| MYB1AT | MYB | WAACCA | 2 | 0 |
| MYB2 | MYB | TAACTSGTT | 2 | 2 |
| MYB2CONSENSUSAT | MYB | YAACKG | 4 | 2 |
| MYB4 | MYB | AMCWAMC | 8 | 4 |
| MYBCORE | MYB | CNGTTR | 8 | 4 |
| MYBPZM | MYB | CCWACC | 2 | 1 |
| ABREATRD22 | ABRE | RYACGTGGYR | 16 | 10 |
| DRECRTCOREAT | DRE/LTRE | RCCGAC | 2 | 2 |

Table 5-7 shows that for two *cis*-elements (ABREBZMRAB28 and MYB1AT) none of the underlying oligonucleotides were under-represented in 1 kb upstream sequences. The ABRE *cis*-element ABREBZMRAB28 corresponds to an ABRE from maize, found in the regulatory region of the gene *RAB28*. The MYB binding site corresponds to a *cis*-element found in *A. thaliana*. Both binding sites have been experimentally proven to be active in the respective organisms.

All underlying oligonucleotides for the *cis*-elements MYB2 and DRECRTCOREAT were found under-represented. The MYB binding site corresponds to a sequence in *Petunia*, and the DRE/LTRE *cis*-element corresponds to a sequence in *A. thaliana*.

The under-representation of all oligonucleotides underlying the *cis*-element DRE/LTRE in 1 kb upstream sequences of *A. thaliana* might plausibly be explained by the fact that this *cis*-element activates the transcription of the down-stream gene as a single copy. If the corresponding oligonucleotides would be over-represented in regulatory sequences genome-wide, a precise regulation of gene expression would be difficult to be achieved.

Over-representation of some or all oligonucleotides corresponding to MYB binding sites and ABRE *cis*-elements might be explained by the high variability of the sequences that describe these binding sites. In the case of the *cis*-elements described as MYB binding sites it was impossible to establish a minimal conserved core. High variability in the description of a binding site reflects the current knowledge on the underlying regulatory process. Different members of a family of transcription factors recognize their target sequences with more or less affinity, leading to the expression of the target genes at different levels[108]. Unfortunately, there is no information available about the affinity of ABRE and MYB binding transcription factors to their target sequence, that could help to distinguish binding sites that would be bind with high affinity from others than not. It can be speculated that over-represented ABRE and MYB binding sites are recognized by transcription factors with high affinity to these sequences, because the target genes must be activated rapidly within few minutes. These rapid activation enables the plant to react efficiently to transient stimuli such as stresses. In such cases, a precise regulation is achieved through the tight regulation of the transcription factors that are activated or deactivated by mean of posttranscriptional modifications carried-out by members of the corresponding signalling pathways (kinases or phosphatases). In contrast, under-represented binding sites might be recognized by transcription factors with low affinity to these sequences. The corresponding target genes are transcribed at low levels, and are involved in developmental gene expression (tissue-specific expression). In that sense, the gene *DC3* from carrot (*Daucus carota*) which encodes a LEA (Late Embryogenesis Abundant) protein was found expressed in response to exogenous ABA and in seeds. The promoter of the gene is composed by two regulatory regions, the distal region confers ABA responsiveness in vegetative tissues, and the proximal region confers seed-specific expression. The binding sites involved are different in sequence[117].

Examples of ratios of under-represented *cis*-elements observed in *A. thaliana* 1 kb upstream sequences, that were described with only one entry and one oligonucleotide, are displayed in Table 5-8.

The *cis*-elements have been originally described in barley (ABRE2HVA22, ABRE3HVA1 and LTRE1HVBLT49), *A. thaliana* (ABRELATERD1, LTRECOREATCOR15 and MYB3) and rice (ABREMOTIFIIIOSRAB16B and CE3OSOSEM). As described earlier, it appears that the original species for which the *cis*-elements have been described does not play an important role when testing whether the corresponding sequence is over or under-represented in another plant genome.

**Table 5-8: Ratio of representation of oligonucleotides corresponding to ABA-related *cis*-elements.** Ratios were calculated based on the observed occurrence of each oligonucleotide in *A. thaliana* 1 kb upstream sequences, according to Equation (4-2)

| w | Name | Oligonucleotide | Group | Ratio |
|---|---|---|---|---|
| 10 | ABRE2HVA22 | CGCACGTGTC | ABRE | -0,41 |
| 10 | ABRE3HVA1 | GCAACGTGTC | ABRE | -0,47 |
| 5 | ABRELATERD1 | ACGTG | ABRE | -0,27 |
| 10 | ABREMOTIFIIIOSRAB16B | GCCGCGTGGC | ABRE | -0,46 |
| 10 | CE3OSOSEM | AACGCGTGTC |  | -0,66 |
| 6 | LTRE1HVBLT49 | CCGAAA | DRE/LTRE | -0,08 |
| 5 | LTRECOREATCOR15 | CCGAC | DRE/LTRE | -0,17 |
| 8 | MYB3 | TAACTAAC | MYB | -0,29 |

## 5.3 Conclusion

1. The analysis of intergenic regions from *A. thaliana* revealed that although some intergenic regions are extremely long (72 kb), most protein coding genes have intergenic regions not longer than 5 kb, with a median of 1.1 kb. According to Pavy *et al.* 1999[80] it was expected that *A. thaliana* intergenic regions vary over at least two orders of magnitude. Here, the analysis based on the genome sequence of *A. thaliana* showed intergenic lengths polymorphism of four orders of magnitude.

2. The analysis of oligonucleotide frequencies showed that around 50% of all possible nucleotides of different sizes were from slightly to strongly under-represented in *A. thaliana* 1 kb upstream sequences. For the analysed oligonucleotide sizes a linear correlation between the percentage of under-represented oligonucleotides and the oligonucleotide size was established. The linear correlation could be confirmed for an oligonucleotide size w=11, but it could not be established whether the linear relationship applies for longer oligonucleotide sizes, and up to which size.

3. Biologically relevant oligonucleotides like methylation and hemimethylation sequences (CG and CNG) were among the most highly under-represented dinucleotides and trinucleotides. However, not all possible hemimethylation sequences were equally under-represented. In the case of *A. thaliana*, experimental results showed that the highly under-represented trinucleotides (CCG-CGG) were found less often methylated than the trinucleotides CAG(CTG). In addition, it was found that the genome of *A. thaliana* is lightly methylated, with approximately 4% of methylated cytosines[60].

4. Some sequences underlying *cis*-elements were found under-represented, like different versions of the TATA-box and well characterised stress responsive *cis*-elements. Showing that indeed sequences important for the tight regulation of expression are sparingly distributed in *A. thaliana* 1 kb upstream sequences. Nonetheless, the reason why some sequences that play a role in regulation were not under-represented could be manifold. Some plausible explanations are:
    1. Not all *cis*-elements listed in PLACE and AGRIS have been proven to be functional experimentally. Some sequences found in the databases were predicted on the basis of expression profiling experiments, using computational tools. It might be possible that the deduced sequences are in fact not functional.

2. Combinatorial control plays a very important role in gene regulation. It might be possible that (i) over-represented *cis*-elements act together with other *cis*-elements that are in fact under-represented, or (ii) that some *cis*-elements correspond to sequences bound with low affinity by the corresponding transcription factors, conferring a low level of expression. It is also important to note, than the distribution of combinations of *cis*-elements follows a complete different statistical model that the distribution of single *cis*-elements.
3. Importantly, some regulatory *cis*-elements are tightly connected to other mechanisms of regulation of gene expression such as DNA methylation or chromatin modification. The influence of these additional components in the regulation of gene expression cannot be efficiently judged merely from sequence data.
4. The protein level is also important in the determination of the regulatory network at a specific spatial and temporal time point. Only the general transcription factors belonging to the basal transcription apparatus are constitutively expressed. Inducible transcription factors are tightly regulated (transcriptionally and post-transcriptionally). The precise control of gene expression in the case of transcription factors that recognize over-represented binding sites could be achieved through other mechanisms, like regulation of the activity of the transcription factor by phosphorylation or dephosphorylation. Such events may be connected to specific signalling pathway components.

5. Various oligonucleotides similar to regulatory *cis*-elements involved in ABA-dependent and ABA-independent gene regulation due to osmotic stress were under-represented genome-wide. Over-representation of such oligonucleotides in a small subset of upstream sequences (e.g. coregulated genes) might eventually mean that the genes evaluated are regulated by ABA or ABA-related stimuli.

# Chapter 6: Computational prediction of genes putatively regulated by ABA

*A genome-wide screening of A. thaliana towards the identification of genes that are potentially responsive to ABA was carried out by using publicly available software. The cis-elements that confer ABA responsiveness were used to screen 1 kb upstream sequences. Subsequently, various statistical analyses were applied to identify statistically significant instances, since not all predicted cis-elements are expected to be true TFBSs.*

## 6.1 Generation of consensus sequences and matrices

For the genome-wide identification of genes putatively regulated by ABA, 1 kb upstream sequences were screened using PSFMs and consensus sequences representing the regulatory elements ABRE, MYB, CE1, CE3 and DRE. Additionally, the matrix and consensus sequence representing the *cis*-element As1 was included in the screening[75].

The consensus sequences and the frequency matrices were derived from aligned binding sites. There are not exact rules how to deduce a consensus sequence[113]. In this study the following rules were used:

1. A single nucleotide was chosen for a position, if the nucleotide occurred in at least 60% of the binding sites at that position.
2. A nucleotide ambiguity code (as defined by IUPAC) representing two nucleotides was used if a single nucleotide did not occur in at least 60% of the binding sites at that position, and one of the following three cases applied:
   i. Only two different nucleotides were observed at that position.
   ii. From three occurring nucleotides, one was present in less than 25% of the binding sites at that position, and the other two in more than 25% of the binding sites at that position, or
   iii. Two nucleotides were observed in at least 60% of the binding sites at that position, none of them present in less than 25% of the binding sites, and the other two were observed in less than 25% of the binding sites at that position.
3. A nucleotide ambiguity code representing three nucleotides was used if three of the four nucleotides were observed in more than 25% of the binding sites at that position, or if only three different nucleotides were observed at that position and each nucleotide was observed in at least 25% of the binding sites.
4. N (any nucleotide) was used in all other cases.

As an example the binding sites and the consensus sequence derived for CE1 are shown in Table 6-1. The binding site has a core region of four nucleotides (CACC), shown in capital letters. All sequences used have a length of nine nucleotides.

**Table 6-1: Annotated sequences of the Coupling Element 1 (CE1) found in the literature.** CONSENSUS denotes the deduced consensus sequence

| Organism | Gene | Sequence | Reference |
|---|---|---|---|
| Hordeum vulgare | *HVA22* | tgcCACCgg | [102] |
| *Zea mays* | *RAB17* | ggcCACCga | [101] |
| *Craterostigma plantagineum* | *CDET27-45* | ttgCACCgt | [101] |
| *Triticum aestivum* | *EM* | acgCACCgc | [101] |
| *Hordeum vulgare* | *HVA1* | gagCACCgc | [101] |
| *Oryza sativa* | *RAB16D* | gccCACCtg | [101] |
| *Oryza sativa* | *RAB 16B* | gctCACCca | [101] |
| *Oryza sativa* | *RAB16C* | gctCACCcc | [101] |
| *Oryza sativa* | *RAB16C* | acgCACCa | [101] |
| *Oryza sativa* | *RAB16C* | cgtCACCga | [101] |
| *Lycopersicon esculentum* | *LE25* | actCACCac | [101] |
| *Arabidopsis thaliana* | *RAB18* | cagCACCct | [101] |
| Oryza sativa | *RAB16A* | cacCACCcg | [71] |
| *Arabidopsis thaliana* | *ATMYB74* | cggCACCga | [27] |
| *Arabidopsis thaliana* | *ATMYB102* | cggCACCga | [27] |
| CONSENSUS | | ssbCACCsv | |

PSFMs were deduced from counts, i.e. the number of occurrences of each nucleotide at each position. As an example, Figure 6-1 shows the matrix for the element CE1, which is based on the binding sites presented in Table 6-1. Information about the different binding sites used to construct the matrices for the *cis*-elements ABRE, As1, CE3, DRE and MYB, and the matrices themselves are presented in Appendix 1.

$$\begin{array}{cc} \begin{array}{c} \mathbf{A} \\ \mathbf{C} \\ \mathbf{G} \\ \mathbf{T} \end{array} & \begin{pmatrix} 3 & 3 & 0 & 0 & 15 & 0 & 0 & 2 & 5 \\ 5 & 6 & 4 & 15 & 0 & 15 & 15 & 4 & 4 \\ 5 & 5 & 7 & 0 & 0 & 0 & 0 & 8 & 4 \\ 2 & 1 & 4 & 0 & 0 & 0 & 0 & 1 & 2 \end{pmatrix} \end{array}$$

**Figure 6-1: Position-specific frequency matrix constructed from the binding sites presented in Table 6-1.** Columns refer to positions one to nine, rows refers to counts for the nucleotides shown at the left side of the matrix

## 6.2 Ratio of representation of the consensus sequences

The ratio of representation (section 4.1.2) of the oligonucleotides corresponding to each consensus sequence generated for the *cis*-elements ABRE, As1, CE1, CE3, DRE and MYB was computed for *A. thaliana* 1 kb upstream sequences. Table 6-2 shows the average ratio of representation for the oligonucleotides corresponding to the consensus sequence of each *cis*-element. The total number of oligonucleotides corresponding to each consensus sequence, the percentage of under-represented oligonucleotides, and the maximum and minimum ratio of representation observed for the oligonucleotides corresponding to each consensus sequence are included.

It was found that all oligonucleotides corresponding to the consensus sequence of the *cis*-element As1 were under-represented. For this *cis*-element a low number of experimental data is provided in the literature. Therefore, the matrix and the consensus sequence generated were derived from practically identical sequences. More than fifty percent of the oligonucleotides corresponding to the consensus sequences of the *cis*-elements DRE and MYB were under-represented. Exactly fifty percent of the oligonucleotides corresponding to the consensus sequence of the *cis*-element CE3 were under-represented, and so were around forty percent of the oligonucleotides corresponding to the consensus sequences of the *cis*-elements ABRE and CE1.

**Table 6-2: Ratio of representation of the oligonucleotides corresponding to the consensus sequence of each ABA-related *cis*-element.** Ratios were calculated based on the observed occurrence of each oligonucleotide in *A. thaliana* 1 kb upstream sequences, according to Equation (4-2). SD=Standard deviation, TO=Total number of oligonucleotides corresponding to the *cis*-element, UR=Percentage of under-represented oligonucleotides, Max and Min ratio=maximum and minimum ratio of representation observed for the corresponding oligonucleotides

| *Cis*-element | Average ± SD | TO | UR | Max. ratio | Min. ratio |
|---|---|---|---|---|---|
| ABRE | 2.38 ± 4.93 | 32 | 38 | 18.68 | -1 |
| As1 | -0.21 ± 0.12 | 3 | 100 | -0.01 | -0.37 |
| CE1 | 0.44 ± 1.28 | 72 | 42 | 15.83 | -0.95 |
| CE3 | 0.26 ± 0.45 | 8 | 50 | 0.58 | -0.06 |
| DRE | -0.24 ± 0.30 | 4 | 75 | 0.65 | -0.68 |
| MYB | -0.17 ± 0.39 | 8 | 75 | 0.93 | -0.79 |

The oligonucleotides corresponding to the consensus sequence of the *cis*-element CE1 were found on average over-represented. The *cis*-element CE1 has a short core sequence (CACC) (Table 6-1). This tetranucleotide was over-represented in 1 kb upstream sequences, with a ratio of representation of 0.24, and longer oligonucleotides containing the tetranucleotide CACC were found also over-represented.

The oligonucleotides corresponding to the consensus sequence of the *cis*-element ABRE were on average over-represented. ABRE is a subtype of the plant-specific *cis*-element so-called G-box. The core sequence of a G-box is the tetranucleotide ACGT that was under-represented in 1 kb upstream sequences (ratio of representation of –0.41). Analysing the results for longer oligonucleotides harbouring the tetranucleotide ACGT it was found that up to heptamers, oligonucleotides containing the ACGT-core were generally under-represented. However, the octamer CACGTGGC was highly over-represented in 1 kb upstream sequences (ratio of representation of 2,75). Thereafter, oligonucleotides longer than w=8 containing the core of the over-represented octamer were also over-represented.

The fact that some oligonucleotides corresponding to the generated consensus sequences were over-represented in 1 kb upstream sequences complicated the distinction between spurious matches and real instances of a *cis*-element in a genome-wide screening. To overcome this problem, *A. thaliana* 1 kb upstream sequences were screened looking for pairs or clusters of ABA-related *cis*-elements. It has been proven that the identification of combinations of *cis*-elements significantly improves the probability to detect the functionally active *cis*-elements [44].

Because the background distribution of combinations of *cis*-elements is unknown in *A. thaliana* 1 kb upstream sequences, the significance of the predictions was assessed by comparing the results obtained for 1 kb upstream sequences with results obtained for random sequences. In that sense random sequences were used as an approximation to the background model. Random sequences were generated by randomly shuffling the sequences while keeping the single nucleotide composition and the length. Hundred datasets were generated in this way, and all programs used for the determination of combinations of *cis*-elements were used with the real and the shuffled datasets.

To test if random sequences were a good approximation to the background model, the ratio of representation of each oligonucleotide corresponding to the consensus sequence of the *cis*-element CE1 was computed in each random dataset. The average ratio of representation of each oligonucleotide in random datasets was compared with results for the same oligonucleotide in 1 kb upstream sequences. In random sequences values close to zero are expected, meaning that the given oligonucleotide do not occur more or less frequently than expected according to the single nucleotide frequencies.

The results observed in random sequences were in agreement with the expected results. For most of the oligonucleotides corresponding to the consensus sequence of the *cis*-element CE1 the ratio of representation was close to zero. In contrast, about 60% of the oligonucleotides were over-represented in 1 kb upstream sequences, and the ratio of representation ranged from 15,83 to –0.95. In Figure 6-2 every bar along the x-axis represents an oligonucleotide, the y-axis corresponds to the values for the ratio of representation (repr) for the given oligonucleotide.

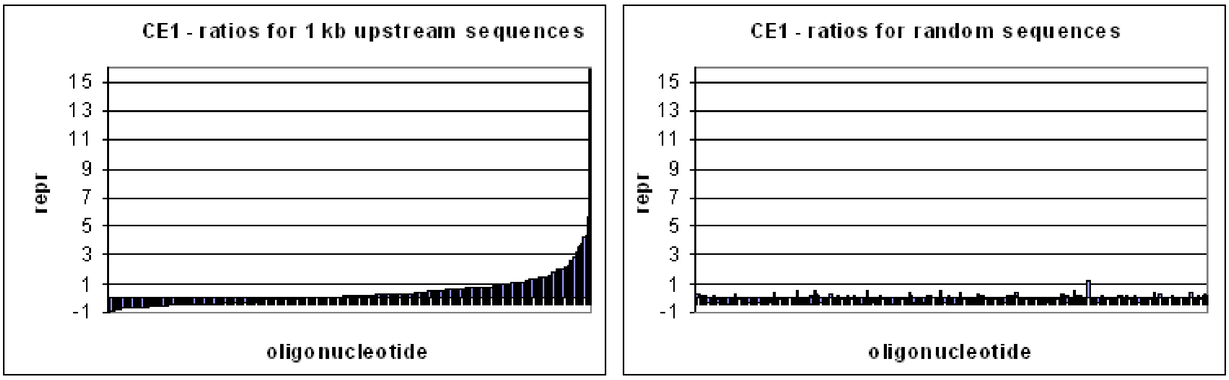


**Figure 6-2: Ratio of representation of oligonucleotides corresponding to the consensus sequence of the *cis*-element CE1.** Ratios were calculated in 1 kb upstream sequences and in random sequences. Every bar along the x-axis corresponds to an oligonucleotide, y-axis shows the ratio of representation calculated according to Equation (4-2)

## 6.3 Pattern-based search

*Cis*-elements can be represented as matrices or consensus sequences. One of the approaches used to screen upstream sequences for ABA-related *cis*-elements uses consensus sequences. Positive instances are those subsequences within the input sequence(s) that exactly match the consensus sequence of the binding site.

To localize pairs of ABA-related *cis*-elements a program was created that uses fuzznuc from EMBOSS[89]. The program looks for all possible pair-wise combinations of *cis*-elements separated by a maximal distance of 1 kb. The order of the *cis*-elements was taken into account, i.e. pair 1-2 was considered to be different from pair 2-1. After the screening the output of the program provides three lists: one list with pair-wise combinations of *cis*-elements and the number of hits found in the query sequence(s), one with the gene number (or sequence identifier) in which a given pair has been found, and one with the distances between the *cis*-elements of a pair.

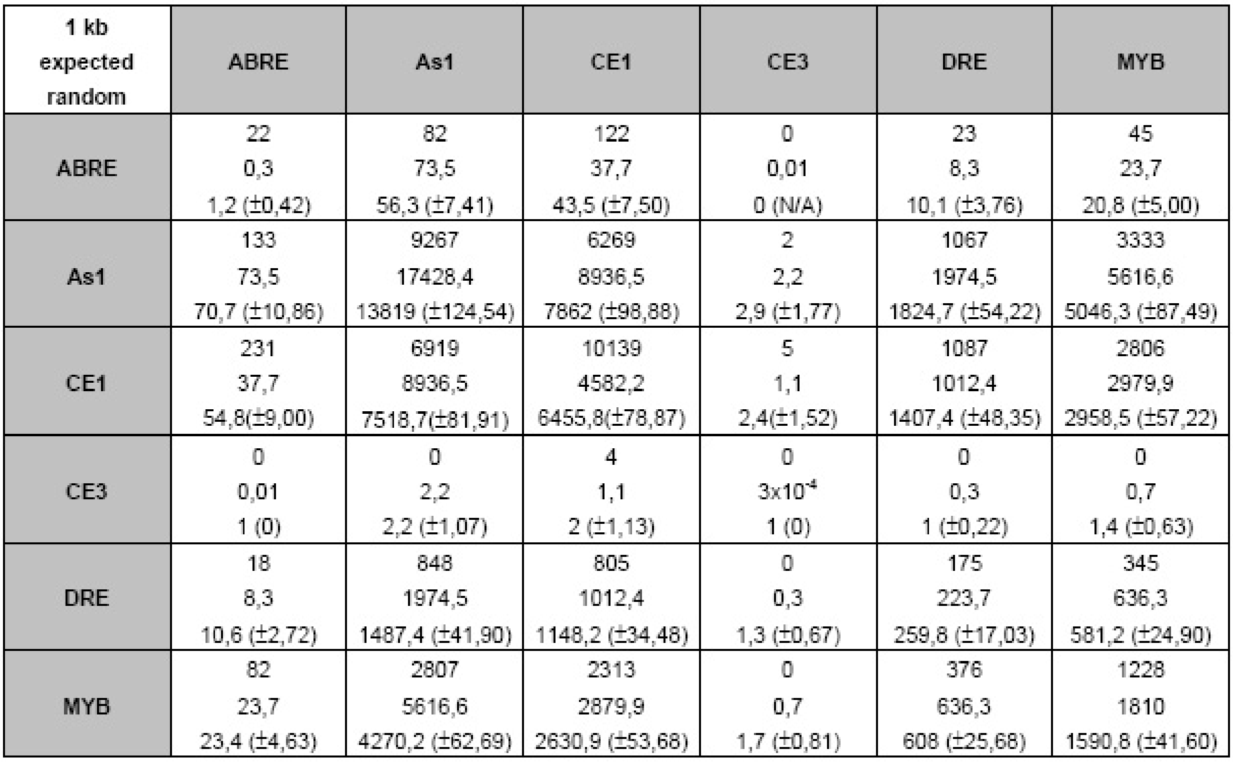
For the pattern-based search a mathematical model for the calculation of the number of expected pairs in a single sequence was deduced [Equation (6-1)]. Simple combinatorial considerations revealed that there are $(N-w_m-w_n+1)*(N-w_m-w_n+2)$ possibilities to place a *cis*-element of length $w_m$ together with a second *cis*-element of length $w_n$, in a sequence of length N. According to this, the number of expected pairs (E) in a sequence of length N depends on the number of possibilities to place both *cis*-elements, and the probability to find each *cis*-element. These probabilities were calculated according to the single nucleotide frequencies in 1 kb upstream sequences, using Equation (4-1).

$$E = \left[(N - w_m - w_n + 1) * (N - w_m - w_n + 2)\right] * P_m P_n \qquad (6\text{-}1)$$

### 6.3.1 Pairs of ABA-related *cis*-elements in 1 kb upstream sequences

The total number of occurrences of a given pair of ABA-related *cis*-elements was counted in 1kb upstream sequences and in each of the hundred random datasets. The results for 1 kb upstream sequences and average results for random datasets are presented in Table 6-3. Additionally, according to Equation (6-1), the expected number of occurrences for each combination of ABA-related *cis*-elements in 29845 sequences was calculated, and is also shown. The columns denote the first *cis*-element of the pair, and the rows denote the second.

**Table 6-3: Total number of occurrences of pairs of ABA-related *cis*-elements in 1 kb upstream sequences (1 kb), expected number of occurrences calculated with Equation (6-1) and average number of occurrences in random datasets.** The first column in the table indicates the first *cis*-element of the pair, the first row indicates the second. For every pair the first row corresponds to the number of occurrences in 1 kb, the second to the expected number of occurrences [Equation (6-1)], and the third to the mean number of occurrences in random datasets, ± SD. N/A=No applicable



| 1 kb<br>expected<br>random | ABRE | As1 | CE1 | CE3 | DRE | MYB |
|---|---|---|---|---|---|---|
| ABRE | 22<br>0,3<br>1,2 (±0,42) | 82<br>73,5<br>56,3 (±7,41) | 122<br>37,7<br>43,5 (±7,50) | 0<br>0,01<br>0 (N/A) | 23<br>8,3<br>10,1 (±3,76) | 45<br>23,7<br>20,8 (±5,00) |
| As1 | 133<br>73,5<br>70,7 (±10,86) | 9267<br>17428,4<br>13819 (±124,54) | 6269<br>8936,5<br>7862 (±98,88) | 2<br>2,2<br>2,9 (±1,77) | 1067<br>1974,5<br>1824,7 (±54,22) | 3333<br>5616,6<br>5046,3 (±87,49) |
| CE1 | 231<br>37,7<br>54,8(±9,00) | 6919<br>8936,5<br>7518,7(±81,91) | 10139<br>4582,2<br>6455,8(±78,87) | 5<br>1,1<br>2,4(±1,52) | 1087<br>1012,4<br>1407,4 (±48,35) | 2806<br>2979,9<br>2958,5 (±57,22) |
| CE3 | 0<br>0,01<br>1 (0) | 0<br>2,2<br>2,2 (±1,07) | 4<br>1,1<br>2 (±1,13) | 0<br>$3x10^{-4}$<br>1 (0) | 0<br>0,3<br>1 (±0,22) | 0<br>0,7<br>1,4 (±0,63) |
| DRE | 18<br>8,3<br>10,6 (±2,72) | 848<br>1974,5<br>1487,4 (±41,90) | 805<br>1012,4<br>1148,2 (±34,48) | 0<br>0,3<br>1,3 (±0,67) | 175<br>223,7<br>259,8 (±17,03) | 345<br>636,3<br>581,2 (±24,90) |
| MYB | 82<br>23,7<br>23,4 (±4,63) | 2807<br>5616,6<br>4270,2 (±62,69) | 2313<br>2879,9<br>2630,9 (±53,68) | 0<br>0,7<br>1,7 (±0,81) | 376<br>636,3<br>608 (±25,68) | 1228<br>1810<br>1590,8 (±41,60) |

After the screening with the consensus sequences no combinations of the *cis*-element CE3 with ABRE, CE3, DRE or MYB were counted in 1 kb upstream sequences. In random datasets few occurrences of CE3 with ABRE, CE3, DRE or MYB, and DRE-CE3, DRE-MYB arouse from the shuffling process.

Regarding the order of the pairs, approximately the same number of occurrences of the reciprocal pairs[10] ABRE and DRE, As1 and CE1, As1 and DRE, As1 and MYB, CE1 and CE3, CE1 and DRE, CE1 and MYB, and DRE and MYB were counted in 1 kb upstream sequences. In contrast, the reciprocal pairs of the *cis*-elements ABRE and As1, ABRE and CE1, and ABRE and MYB showed nearly a two-fold difference. The pairs *cis-element2* – ABRE were counted nearly two times more frequently than ABRE – *cis-element2* (Table 6-3). To select pairs that showed statistically significant differences with respect to random sequences, the number of instances counted in 1 kb upstream sequences and random datasets was compared. It was assumed that the number of instances for any combination of *cis*-elements was normally distributed. For each combination of *cis*-elements the probability (P) that the number of instances in 1 kb upstream sequences belonged to the normal distribution of values in random datasets was computed, taking into account the mean and the standard deviation of the number of instances in random datasets. The significance is given by the confidence interval, here defined as $0.01<P<0.99$. Consequently, if the computed probability is $\geq 0.99$ or $\leq 0.01$, the counted number of pairs in 1 kb upstream sequences is significantly different from the counted number of pairs in random datasets. If the mean or the standard deviation of the counted number of pairs in random datasets is equal to zero, the probability is not defined (e.g. for the pairs ABRE-CE3, CE3-ABRE and CE3-CE3, Table 6-3).

In some cases, the number of expected instances calculated according to Equation (6-1) was larger than the number of counts in random datasets and/or larger than the number of counts in 1 kb upstream sequences. The significance of these differences was computed as described above. The probability that the calculated number of instances belonged to the normal distribution of the counted number of pairs in random datasets was computed. If the calculated probability falls above or below the confidence interval ($0.01<P<0.99$), the calculated number of instances is significantly different than the counted number of pairs in random datasets. Both, the number of expected pairs calculated according to Equation (6-1), and the average number of pairs counted in random datasets are related to background expectations. Thus, if the differences between the number of pairs counted in random datasets and the number of pairs calculated is statistically significant, it can be stated that for this combination of *cis*-elements a reliable expectation model cannot be established. For that reason, the respective pair was not considered for further analysis, even if the number of counts in 1 kb upstream sequences was significantly different from the number of counts in random datasets.

[10] As an example the reciprocal pairs of the cis-elements DRE and MYB are DRE-MYB and MYB–DRE.

In Table 6-4 the computed probabilities for the comparison between 1 kb upstream sequences and random datasets are shown. Probabilities that indicate significant differences between 1 kb upstream sequences and random datasets, and on which no significant differences between background expectation models[11] were found are highlighted in blue. For the pairs where the comparison between background expectation models showed significant differences, the corresponding cells are shaded in grey.

**Table 6-4: Probability to observe the counted number of pairs in 1 kb upstream sequences with regard to counted number of pairs in random datasets.** The fist column of the table indicates the first element of the pair; the first row indicates the second. N/A.=No applicable. Cells shaded in grey indicate significant differences between background expectation models (P≥0.99 or P≤0.01). Pairs that showed significant differences (P≥0.99 or P≤0.01) are highlighted in blue

| | ABRE | As1 | CE1 | CE3 | DRE | MYB |
|---|---|---|---|---|---|---|
| **ABRE** | **1** | 1 | **1** | N/A | **1** | **1** |
| **As1** | **1** | 0 | 0 | 0.313 | 0 | 0 |
| **CE1** | **1** | 0 | **1** | 0.955 | **0** | **0.004** |
| **CE3** | N/A | 0.018 | 0.961 | N/A | **0** | 0.014 |
| **DRE** | **0.997** | 0 | 0 | 0.024 | **0** | **0** |
| **MYB** | **1** | 0 | 0 | 0.019 | **0** | 0 |

Overall the results showed:

1. For fifteen pairs the differences observed between 1 kb upstream sequences and random datasets were significant.
2. The only CE3-containing pair that showed significant differences between 1 kb upstream sequences and random datasets was CE3-DRE.
3. With the exception of the pairs ABRE-CE3 and CE3-ABRE that were not counted in 1 kb upstream sequences, and the pair ABRE-As1, all other pairs involving the *cis*-element ABRE showed significant differences between 1 kb upstream sequences and random datasets.
4. Almost all pairs involving the *cis*-element As1 showed significant differences between background expectation models (observed instances in random datasets and calculated number of instances according to the deduced model). The only exceptions were the pairs As1-ABRE, and the reciprocal pairs of the *cis*-elements As1 and CE3.

[11] The background expectation models represent the number of instances counted in random datasets, or the expected number of instances calculated according to Equation (6-1).

In Table 6-4 if the computed probability was close to zero, the number of instances in 1 kb upstream sequences was significantly smaller than in random datasets. If the computed probability was close to one the number of instances in 1 kb upstream sequences was significantly larger than in random datasets.

From 15 pairs that showed significant differences, six exhibited significant under-representation in the number of instances compared with random datasets, and nine exhibited significant over-representation in the number of instances compared with random datasets. Regarding over-represented pairs, eight are combinations of ABRE with other *cis*-elements and with itself. The other over-represented pair was CE1-CE1.

### 6.3.2 Number of genes that showed pairs of ABA-related *cis*-elements in 1 kb upstream sequences

In *A. thaliana* 1 kb upstream sequences it was observed that 17935 genes were predicted to have at least one pair of ABA-related *cis*-elements. Therefore, according to this result 60% of the annotated *A. thaliana* genes harbour at least one pair of ABA-related *cis*-elements.

In Table 6-5 the number of genes that showed any pair of ABA-related *cis*-elements in the set of 1 kb upstream sequences and the average number of sequences that showed any pair of ABA-related *cis*-elements in the random datasets are shown. The columns denote the first *cis*-element of the pair, and the rows denote the second *cis*-element.

The identified pairs of ABA-related *cis*-elements may be equally distributed among a large number of genes, or alternatively, may reside in a small number of genes. To test whether the number of genes showing a specific pair of ABA-related *cis*-elements was different from the number of genes showing the elements separately, the significance score introduced by Manke *et al.* 2003 was used[65].

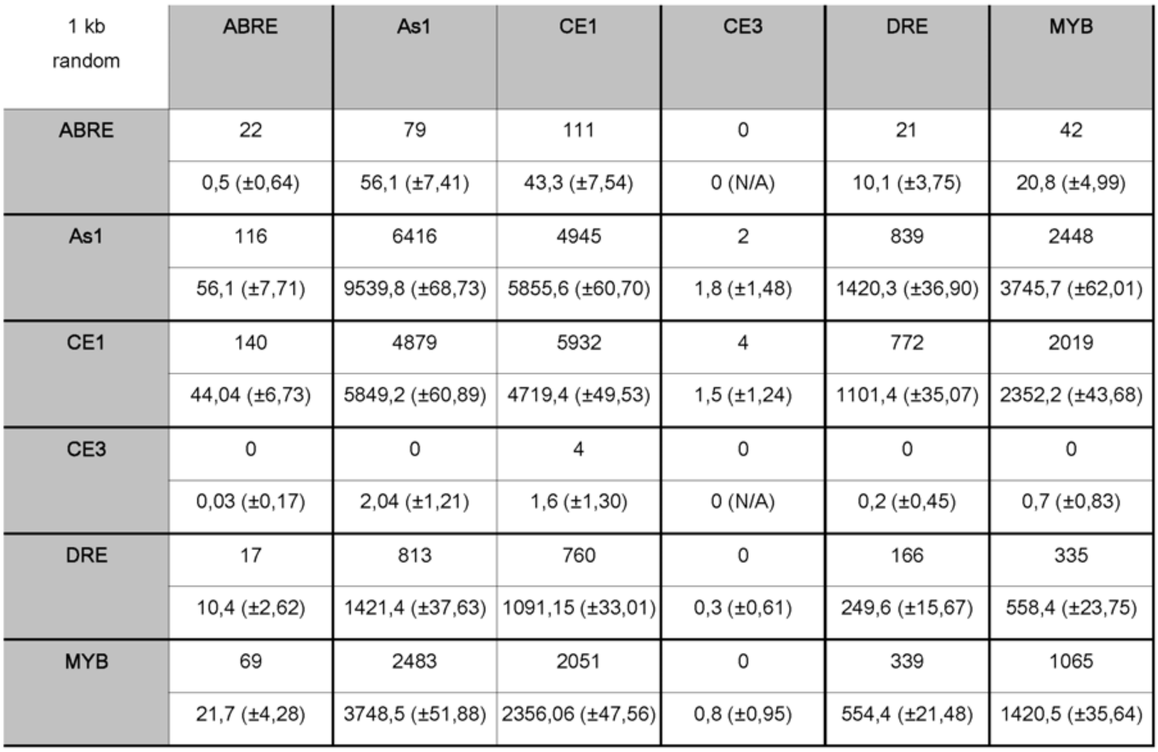
The significance score compares as probabilities the frequency of a pair *ij* with the frequency of the independent *cis*-elements *i* and *j*. The logarithm of the probability ratio defines the significance of finding a pair compared with the expectation for the single *cis*-elements[65,86].

$$S_{ij} = \log\left(\frac{p_{ij}}{p_i p_j}\right) \tag{6-2}$$

Here $p_{ij}$ is $n_{ij}/N$ (frequency of pair *ij* in N upstream regions), and $p_i$ and $p_j$ are $n_i/N$ or $n_j/N$, respectively (frequency of the single *cis*-elements). Every pair *ij* and *cis*-element *i* and *j* was counted only once per upstream sequence, even if it occurred multiple times.

Positive $S_{ij}$ scores denote pairs of *cis*-elements that preferentially occur together[65,86]. If the significance score is equal to zero, then there is no difference between the frequency of the pair *ij* compared with the frequencies of the independent *cis*-elements. Negative scores indicate that the pair *ij* is less frequent than the separate *cis*-elements.

**Table 6-5: Number of genes that showed any kind of pair of ABA-related *cis*-elements.** The first column of the table indicates the first element of the pair, the first row indicates the second. For every pair the first row correspond to the number of genes that showed the given pair of *cis*-elements in 1 kb upstream sequences (1 kb) and the second row correspond to the mean number of sequences that showed the given pair of *cis*-elements in random datasets ± SD. N/A: No applicable



| 1 kb<br>random | ABRE | As1 | CE1 | CE3 | DRE | MYB |
|---|---|---|---|---|---|---|
| ABRE | 22 | 79 | 111 | 0 | 21 | 42 |
| | 0,5 (±0,64) | 56,1 (±7,41) | 43,3 (±7,54) | 0 (N/A) | 10,1 (±3,75) | 20,8 (±4,99) |
| As1 | 116 | 6416 | 4945 | 2 | 839 | 2448 |
| | 56,1 (±7,71) | 9539,8 (±68,73) | 5855,6 (±60,70) | 1,8 (±1,48) | 1420,3 (±36,90) | 3745,7 (±62,01) |
| CE1 | 140 | 4879 | 5932 | 4 | 772 | 2019 |
| | 44,04 (±6,73) | 5849,2 (±60,89) | 4719,4 (±49,53) | 1,5 (±1,24) | 1101,4 (±35,07) | 2352,2 (±43,68) |
| CE3 | 0 | 0 | 4 | 0 | 0 | 0 |
| | 0,03 (±0,17) | 2,04 (±1,21) | 1,6 (±1,30) | 0 (N/A) | 0,2 (±0,45) | 0,7 (±0,83) |
| DRE | 17 | 813 | 760 | 0 | 166 | 335 |
| | 10,4 (±2,62) | 1421,4 (±37,63) | 1091,15 (±33,01) | 0,3 (±0,61) | 249,6 (±15,67) | 558,4 (±23,75) |
| MYB | 69 | 2483 | 2051 | 0 | 339 | 1065 |
| | 21,7 (±4,28) | 3748,5 (±51,88) | 2356,06 (±47,56) | 0,8 (±0,95) | 554,4 (±21,48) | 1420,5 (±35,64) |

The $S_{ij}$ score is undefined if no instances of the pair or of the independent *cis*-elements have been found (i.e. pairs ABRE-CE3, CE3-ABRE, CE3-As1, CE3-CE3, CE3-DRE, CE3-MYB, DRE-CE3 and MYB-CE3). Table 6-6 shows the calculated *Sij* scores. For the eight pairs that were not observed the $S_{ij}$ score is regarded as N/A. (no applicable). In the cases were the number of instances of a pair in 1 kb upstream sequences was significantly over-represented compared with random datasets (Table 6-4 section 6.3.1), and where the background expectation models did not show statistically significant differences (Table 6-4, section 6.3.1) the corresponding cells are shaded grey. The columns denote the first *cis*-element of the pair, and the rows the second *cis*-element.

The results shown in the Table 6-6 revealed that most of the observed pairs of ABA-related *cis*-elements obtained a positive $S_{ij}$ score, indicating that the pairs were found more frequently than the independent *cis*-elements. The negative $S_{ij}$ scores observed for the pairs ABRE-As1 and As1-CE3 indicate that the independent *cis*-elements were found more frequently than the pairs. The pair ABRE – MYB was found as frequently as the independent *cis*-elements ($S_{ij}$=0). Finally, the pair ABRE-ABRE achieved the largest $S_{ij}$ score from all observed pairs.

**Table 6-6. Significance score ($S_{ij}$) calculated according to the number of genes that showed a give pair of *cis*-elements using Equation (6-2).** The frequency of a pair is compared with the frequency of the independent *cis*-elements. The first column of the table indicates the first *cis*-element of the pair, the first row indicates the second *cis*-element. N/A=No applicable. Grey shaded cells indicate cases where the number of pairs counted in 1 kb was significantly over-represented compared with random datasets, and where no statistically significant differences between background expectation models were observed (see section 6.3.1)

| | ABRE | As1 | CE1 | CE3 | DRE | MYB |
|---|---|---|---|---|---|---|
| **ABRE** | 1,028 | -0,112 | 0,071 | N/A | 0,159 | 0 |
| **As1** | 0,055 | 0,102 | 0,024 | -0,013 | 0,065 | 0,070 |
| **CE1** | 0,172 | 0,019 | 0,139 | 0,324 | 0,065 | 0,022 |
| **CE3** | N/A | N/A | 0,324 | N/A | N/A | N/A |
| **DRE** | 0,068 | 0,052 | 0,058 | N/A | 0,208 | 0,053 |
| **MYB** | 0,216 | 0,076 | 0,029 | N/A | 0,058 | 0,096 |

Interestingly, reciprocal pairs showed different $S_{ij}$ scores. It was observed that the reciprocal pair that was observed in more genes obtained the higher $S_{ij}$ score. If the number of genes that showed a reciprocal pair was the same, the $S_{ij}$ score was equal for both pairs.

The reciprocal pair between CE1 and CE3 was observed in the same number of genes (4). Thus, both pairs obtained the same $S_{ij}$ score. Interestingly, the comparison between the number of pairs observed in 1 kb upstream sequences and in random sequences revealed that there were no statistical differences between both datasets (see Table 6-4, section 6.3.1). However, these pairs achieved the second highest $S_{ij}$ score. This result shows clearly that the affirmation made by Manke *et al.* 2003[65] that $S_{ij}$ scores define the significance of finding a pair in comparison with random expectation is not correct. For that reason, only pairs that showed significantly over-representation in 1kb upstream sequences compared with random datasets, and where no statistically significant differences between background expectation models were found (grey-shaded cells in Table 6-6) were considered further, and analysed in more detail. The $S_{ij}$ score was used then to assess the significance of finding a pair in comparison with the independent occurrence of the *cis*-elements.

### 6.3.3 Distance between ABA-related *cis*-elements

$S_{ij}$ scores larger than 0 indicate pairs of *cis*-elements that occur preferentially together. It has been found that *cis*-elements conferring responsiveness to the same stimulus or group of stimuli tend to form clusters showing defined distances between *cis*-elements[67,137]. To test whether pairs that showed high $S_{ij}$ scores might also show a defined distance between *cis*-elements, and whether reciprocal pairs that achieved higher $S_{ij}$ score shown more defined distance between *cis*-elements compared with results in random datasets, the distance distribution of ABA-related *cis*-elements statistically over-represented in 1 kb upstream sequences was plotted in a histogram. The window length of the histogram was 1000 bp, with mutually exclusive distance intervals of 50 bp. Overlapping *cis*-elements were excluded from the analysis. The minimal distance between *cis*-elements was found to be never smaller than 15 bp. The distance distribution was compared with the results obtained for the same pair in random datasets. To allow a direct comparison the number of pairs observed at each distance interval in each dataset was displayed as percentage of the total number of observed pairs. Results are shown in Figure 6-3.

The results for the homologous pair ABRE-ABRE (Figure 6-3A) showed that in 1 kb upstream sequences about 60% of the pairs were separated by a maximum of 50 bp. Additionally, the distance distribution was clearly different from the distribution observed in random datasets. The pair ABRE-ABRE showed also the largest $S_{ij}$ score (1.028). Considering that ABRE needs a coupling element to activate transcription, these results indicate that in *A. thaliana* ABRE might act as a coupling element of itself.

Other *cis*-elements that have been reported as coupling elements of ABRE are CE1, CE3 and DRE[14-16,47,50,54,75,77,83,103,132,134,135]. The combinations ABRE-CE3 or CE3-ABRE were not found in any of the *A. thaliana* 1 kb upstream sequences. The combinations of ABRE and CE1 (Figure 6-3C, D) were over-represented in 1 kb, with positive $S_{ij}$ scores. The pair ABRE-CE1, had a distance distribution of its *cis*-elements that was similar to the results obtained for random datasets.

In addition, a fixed distance between *cis*-elements was not observed (Figure 6-3D). The pair ABRE-CE1, that had a lower $S_{ij}$ score (0.071) showed a small peak of pairs having a distance between *cis*-elements of 201 to 250 bp. This peak was not observed in random datasets. The other distance intervals showed approximately the same percentage of pairs as in random datasets (Figure 6-3C). The positive $S_{ij}$ scores obtained for both pairs (ABRE-CE1 and CE1-ABRE) clearly indicate that these *cis*-elements preferentially occur together in 1 kb upstream sequences. However, the distance distribution of the *cis*-elements in 1kb upstream sequences is closely similar to the distance distribution observed in random datasets. The observed distance between *cis*-elements of a pair do not support the observations that were made in monocots, where CE1 was found in close vicinity to ABRE, in ABA-responsive genes [101].

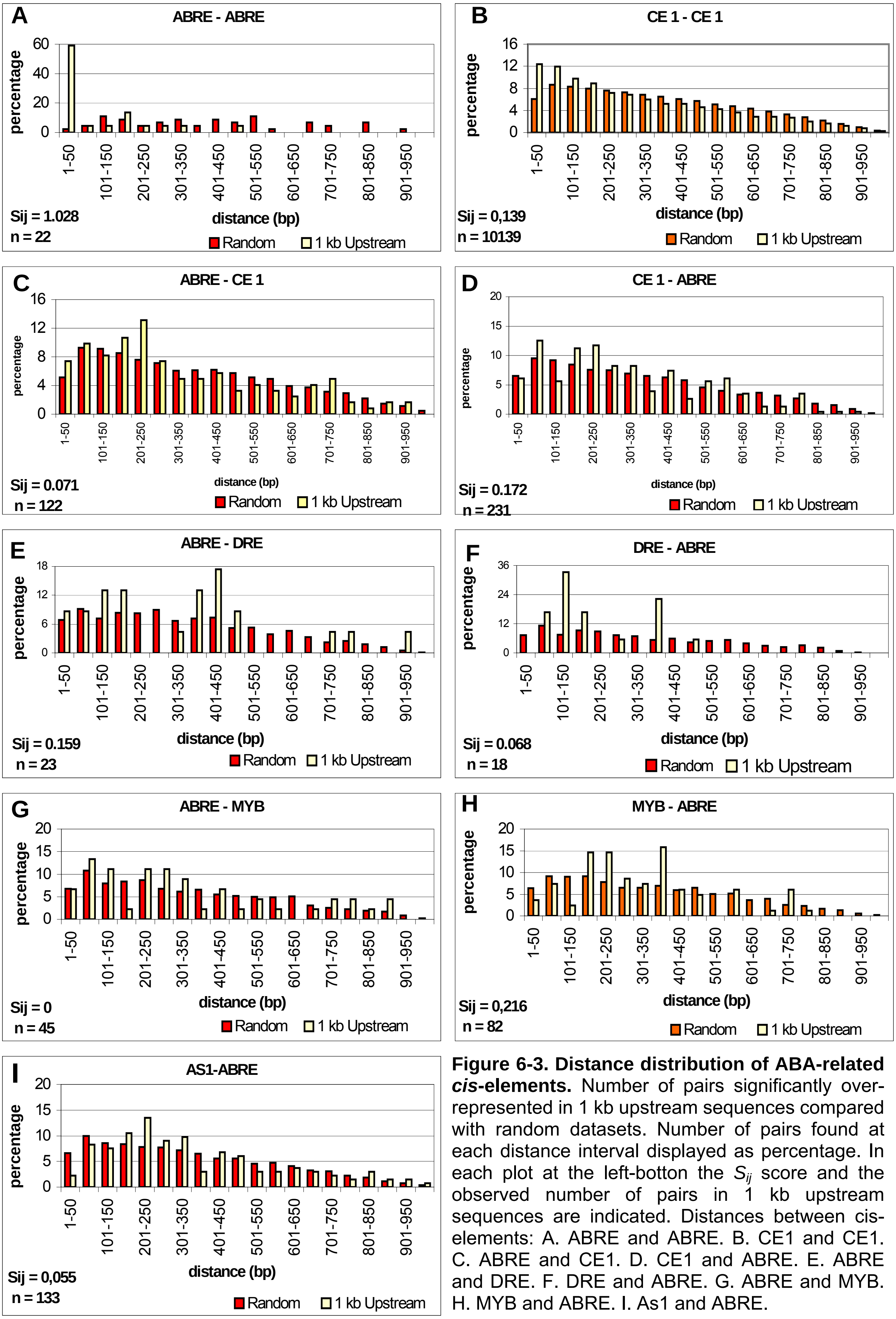


**Figure 6-3. Distance distribution of ABA-related *cis*-elements.** Number of pairs significantly over-represented in 1 kb upstream sequences compared with random datasets. Number of pairs found at each distance interval displayed as percentage. In each plot at the left-botton the $S_{ij}$ score and the observed number of pairs in 1 kb upstream sequences are indicated. Distances between cis-elements: A. ABRE and ABRE. B. CE1 and CE1. C. ABRE and CE1. D. CE1 and ABRE. E. ABRE and DRE. F. DRE and ABRE. G. ABRE and MYB. H. MYB and ABRE. I. As1 and ABRE.

Considering that pairs between the *cis*-elements ABRE and CE3 were not observed in 1 kb upstream sequences, and that neither CE1 nor CE3 *cis*-elements have been observed in ABA-responsive genes of *A. thaliana*, it might be possible that none of these coupling elements of monocots act as coupling elements of ABRE in *A. thaliana*. It might also be possible that the sequence(s) of the coupling element(s) of ABRE in *A. thaliana* are different from those present in monocots. In addition, the results observed by Zhang *et al.* 2005[137] supported only a weak role of coupling elements in the regulation of ABA-responsive genes. The authors screened *A. thaliana* promoter sequences with frequency matrices of the *cis*-elements ABRE and CE. Computationally predicted ABA-responsive genes that harbour ABRE-CE *cis*-elements could not be confirmed experimentally.

Another *cis*-element described as coupling element of ABRE in *A. thaliana* is DRE[75]. In this study it was observed that the pair that achieved the lowest $S_{ij}$ score (DRE-ABRE, Figure 6-3F) showed a distance distribution of *cis*-elements that was clearly different from the distribution observed in random datasets. Two tendencies are highlighted by the distance histogram: (i) large distances between *cis*-elements (more than 500 bp) were avoided, and also (ii) short distances (less than 50 bp) were not observed. Two clear peaks of distances were observed for 1 kb upstream sequences at 101 to 150 bp, and 351 to 400 bp. The pair ABRE-DRE, despite a larger $S_{ij}$ score showed less pronounced peaks in the distance distribution. The peaks were observed at the following distance intervals: 101 to 200 bp, and 351 to 450 bp, and most of the pairs were observed in the interval 401 to 450 bp (18%)(Figure 6-3E).

The positive $S_{ij}$ scores for both combinations of ABRE and DRE indicate that these *cis*-elements preferentially occur together. However, the distance of the *cis*-elements makes unlikely that DRE acts as a coupling element of ABRE. Instead, it is more likely that combinations of ABRE and DRE are found in genes that respond to osmotic stresses in an ABA-dependent (mediated by ABRE) and ABA-independent (mediated by DRE) signalling pathway, allowing cross-talk during osmotic stress. Accordingly, Naruzaka *et al.* 2003[75] proposed that DRE-binding proteins (DREB) may co-operate with ABRE-binding proteins (AREB), coordinating the ABA-dependent gene expression of *RD29A*. The presence of both kinds of elements in the promoter of the gene allows that *RD29A* is induced rapidly or slowly under dehydration and high salinity stresses. The rapid induction appears to be ABA-independent, mediated by DRE, whereas the slow induction after the accumulation of ABA is mediated by the cooperation of ABRE and DRE binding proteins.

The pair ABRE-MYB (Figure 6-3G) was the only pair that showed a $S_{ij}$ score of zero, clearly indicating that the pair of *cis*-elements does not occur more frequently than the *cis*-elements independently. The distance distribution histogram additionally shows that the distance between *cis*-elements is not different from the results obtained for random datasets. In contrast, the pair MYB-ABRE that had one of the highest $S_{ij}$ scores (0,216) showed three distinct peaks in the distance distribution of *cis*-elements at 151 to 200 bp, 201 to 250 bp, and 351 to 400 bp (Figure 6-3H). Approximately 45% (45.1) of the pairs counted in 1 kb upstream sequences showed one of the already mentioned separation distance between *cis*-elements. In random datasets approximately 24% (23.87) of the pairs showed the same separation distances. The fact that the pair MYB-ABRE achieved an $S_{ij}$ score higher than other *cis*-elements documented as coupling elements of ABRE (for example CE1 or DRE) is a very interesting result, and might be related to the following observations:

1. MYB binding sites have been found in ABA-inducible genes[2,135,138]. For the induction of the gene *RD22* by ABA additional to the MYB binding site, a G-box which is similar in sequence to ABRE is needed [2]. The induction of genes harbouring a MYB-ABRE pair requires protein synthesis[3]. Thus, as in the case of combinations of DRE and ABRE, the pair MYB-ABRE might be involved in the slow induction of ABA-regulated genes.
2. The association found between MYB-ABRE was stronger than the association between combinations of DRE and ABRE (according to the $S_{ij}$ scores). As mentioned earlier, DRE is a TFBS that can drive the transcription in ABA-dependent and ABA-independent signalling pathways after osmotic stress, whereas MYB-binding sites have been associated to the ABA-dependent signalling pathway. Thus, considering this fact it is not surprising that *cis*-elements linked to the transcriptional regulation of genes activated in the same signalling pathway showed a stronger association than *cis*-elements involved in gene regulation in separated signalling pathways.
3. The association between MYB-ABRE according to the $S_{ij}$ score is stronger than the association between the *cis*-elements ABRE and CE1. According to the results presented here, CE1 might not be a functional binding site in *A. thaliana*. This conclusion is supported by the results presented by Zhang *et al.* 2005[137], indicating that in contrast to the ABA-transcriptional regulation observed in mococotyledoneous species, in *A. thaliana* is more likely that ABRE couples with itself to activate the ABA-mediated transcription.

MYB might act as a coupling element of ABRE, this hypothesis cannot be discarded completely. However, the mechanism of action implies the formation of dimmers between the ABRE binding protein (bZIP transcription factors) and the protein binding to the coupling element (AP transcription factors in the case of CE1). The large distance found between MYB and ABRE (usually larger than 150 bp), provides only weak support for a direct interaction between the ABRE binding protein and MYB transcription factors. Instead, it is more likely that MYB-binding proteins interact with ABRE-binding proteins by forming higher-order protein complexes. The separation between MYB-ABRE *cis*-elements (151 to 250 bp and 351 to 400 bp) provides some hints about the size of the protein complexes necessary to allow the contact between these *cis*-elements.

For the homologous pair CE1-CE1 (Figure 6-3B), a defined distance between *cis*-elements could not be clearly defined. Approximately 24% (24.3) of the pairs observed in 1 kb upstream sequences showed a separation between *cis*-elements of 1 to 100 bp, compared with 14.7% in random datasets. For the other distance intervals the percentage of pairs was almost equal in both cases. Until now, unlike homologous pairs of ABRE elements, homologous pairs of CEs have not been experimentally described to activate the transcription of ABA-regulated genes. For that reason, genes predicted to have only CEs in their upstream sequences are good candidates to be tested experimentally, to establish whether they are in fact regulated by ABA. Such experiments could also demonstrate whether in *A. thaliana*, in contrast to observations made in monocots, homologous pairs of CE binding proteins confer ABA-responsiveness.

Finally, for the pair As1-ABRE that also showed a positive $S_{ij}$ score, the distance distribution between *cis*-elements is almost identical in 1 kb upstream sequences and in random datasets. The only difference between both distributions consisted in a small peak of pairs of elements with a distance between them of 201 to 250 bp, observed in 1 kb upstream sequences (Figure 6-3I).

The analysis of the number of pairs and number of genes that harbour ABA-related *cis*-elements in 1 kb upstream sequences revealed that:

1. Some combinations of *cis*-elements were significantly under-represented in 1 kb upstream sequences compared with random datasets, and some combinations were significantly over-represented.
2. The $S_{ij}$ score defined by Manke *et al.* 2003[65], could not be used to evaluate whether the observed number of pairs was significantly different from the number of pairs expected by chance. However, was useful to define pairs of *cis*-elements that occur preferentially together (positive $S_{ij}$ score) compared with the frequency of the independent *cis*-elements

The selection of pairs of *cis*-elements that are significantly over-represented in 1 kb upstream sequences compared with background expectations notably reduces the number of genes to be considered putatively regulated by ABA in *A. thaliana*. From 17935 genes initially predicted, only 6132 remained as significant predictions, implicating that about 20% of the annotated genes in *A. thaliana* might be regulated by ABA.

## 6.4 Matrix-based search

To identify putative ABA-responsive genes, two matrix based search algorithms were used. In contrast to pattern-based searches where a hit is defined as a subsequence identical to the consensus sequence of the binding site, matrix-based methods are probabilistic methods, and each putative *cis*-element obtains a score, which is useful to judge the quality of the match. Matrix-based searches can be considered as complementary to pattern-based searches, since such approaches allow the evaluation of higher-order combinations of *cis*-elements (and not only pairs). The scores can be used to select high-scoring *cis*-elements as the putative functional *cis*-elements.

The programs MotifScanner[4] and CISTER[34] were used to screen *A. thaliana* 1 kb upstream sequences, to search for combinations of ABA-related *cis*-elements. In both programs each *cis*-element is represented as a PSFM. However, the probability of a match is defined according to different parameters in each program.

### 6.4.1 MotifScanner

The program was implemented by Gert Thijs at the Catholic University of Leuven[4,112]. For the detection of matches, every query sequence is scanned with each matrix separately and sequentially, and compared with a background model. The background model used is a second-order HMM constructed from a subset of *A. thaliana* intergenic sequences[80,111]. The score that every match obtains was defined as the ratio between the probability that the matching subsequence has been generated by the background model, and the probability that the matching subsequence has been generated by the motif model (determined by the frequency matrix). High scores indicate large differences between the match and the background model, i.e. the match is closer to the motif model.

A large number of hits were found in 1 kb upstream sequences. The obtained scores were scatter values. To select relevant instances and to compare results obtained for different matrices, the scores were rescaled to values between 0 and 1. For that, the minimum and maximum scores for each *cis*-element were computed ($W\min$ and $W\max$, respectively), and values were rescaled according to Equation (6-3).

$$\bar{W}x = \frac{Wx - W\min}{W\max - W\min} \tag{6-3}$$

The minimum and maximum scores largely depend on the motif model used (frequency matrix).

After the screening with MotifScanner, nearly all 1 kb upstream sequences were predicted to have at least one of the investigated *cis*-elements (28935 sequences from 29845). Only few sequences were predicted to have exactly one *cis*-element (3356). In this study, the main focus was on the prediction of genes putatively regulated by ABA, based on the presence of more than one *cis*-element that confers ABA-responsiveness. Genes predicted to have exactly one *cis*-element were not considered for further analyses.

It was observed that independent from the considered *cis*-element, rescaled scores were always very low, indicating a poor difference between the background model and the motif model. Additionally, the *cis*-element that showed the lowest scores was ABRE (about 98% of the predicted *cis*-element achieved a score below 0.1), and the *cis*-element that showed better scores was As1, with about 50% of the predicted *cis*-elements that achieved a score larger than 0.1. In Figure 6-4 the distribution of scores for each considered *cis*-element is shown.

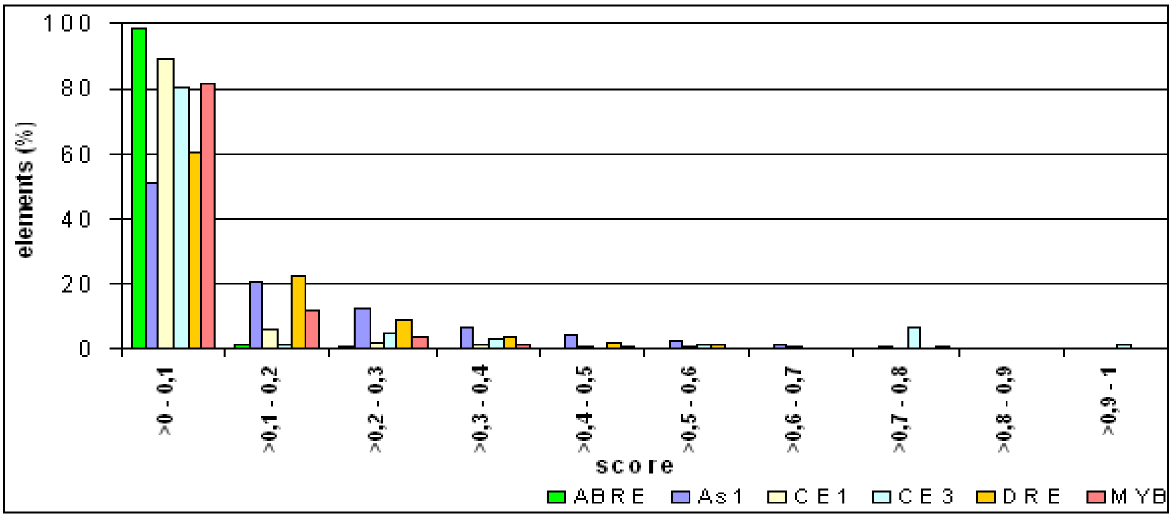


**Figure 6-4: Distribution of rescaled scores for MotifScanner.** Results observed for *A. thaliana* 1 kb upstream sequences. Scores were rescaled according to Equation (6-3)

MotifScanner was also used with the hundred random datasets, to evaluate the scores obtained for random sequences. It was observed that the shuffling process generated some random *cis*-elements. In comparison with 1 kb upstream sequences, more sequences were predicted on average to have two or more ABA-related *cis*-elements in random datasets (27863 ± 46 sequences per random dataset, compared to 25578 in 1 kb upstream sequences). Furthermore, also the mean number of predicted *cis*-elements in random datasets was larger than in 1 kb upstream sequences, as can be observed in Table 6-7. Another observation made is that not all *cis*-elements were predicted with the same frequency. Notably the *cis*-element CE3 was very rarely predicted in either dataset.

Scores observed in the random datasets were compared with the scores observed in 1 kb upstream sequences, in order to establish whether the two kind of datasets can be differentiated or not. First, the scores were rescaled according to Equation (6-3), using the minimum and maximum values found in 1 kb upstream sequences. During the rescaling process some hits achieved a score higher than 1. However, rescaling the scores observed in random datasets with the corresponding minimum and maximum values did not lead to any change in the observed distribution (data no shown). To allow for direct comparison, the number of predicted *cis*-elements for each score interval was displayed as percentage of the total number of predicted *cis*-elements (Figure 6-5).

**Table 6-7: Number of predicted *cis*-elements in *A. thaliana* 1 kb upstream sequences and average number of *cis*-elements predicted in hundred random datasets using MotifScanner.** Results for random datasets are given ± SD

| *Cis*-element | 1 kb upstream | Random datasets |
|---|---:|---:|
| ABRE | 7753 | 11534,9 ± 122,6 |
| As1 | 34533 | 41803,8 ± 111,0 |
| CE1 | 25917 | 26928,9 ± 98,3 |
| CE3 | 61 | 81,4 ± 10,4 |
| DRE | 6173 | 9022,1 ± 99,7 |
| MYB | 13914 | 17770,9 ± 117,1 |

Independent of the *cis*-element considered, it was observed that the score distribution in either dataset follows the same tendency. In both datasets the predicted ABRE *cis*-elements generally obtained very low scores (≤0.1) and more than 40% of the predicted As1 *cis*-elements obtained scores ≥0.1. The only small difference observed between both datasets is related to the predicted CE3 *cis*-elements. While nearly no *cis*-elements were observed in random datasets for the score interval >0-7 to 0.80, a very small peak was observed in 1 kb upstream sequences (Figure 6-5D).

Considering that no clear differences were observed between both datasets, and that most of the predicted *cis*-elements achieved scores ≤0.1, a threshold score of 0.1 was set. From the 25578 sequences predicted in the set of 1 kb upstream sequences, only 1882 sequences achieved scores equal to or above the threshold for most of the predicted *cis*-elements. Unfortunately, the filter threshold cannot be directly introduced as a parameter into MotifScanner, and for that reason some of the predicted *cis*-elements in the subset of 1882 sequences obtained a rescaled score lower than the chosen threshold.

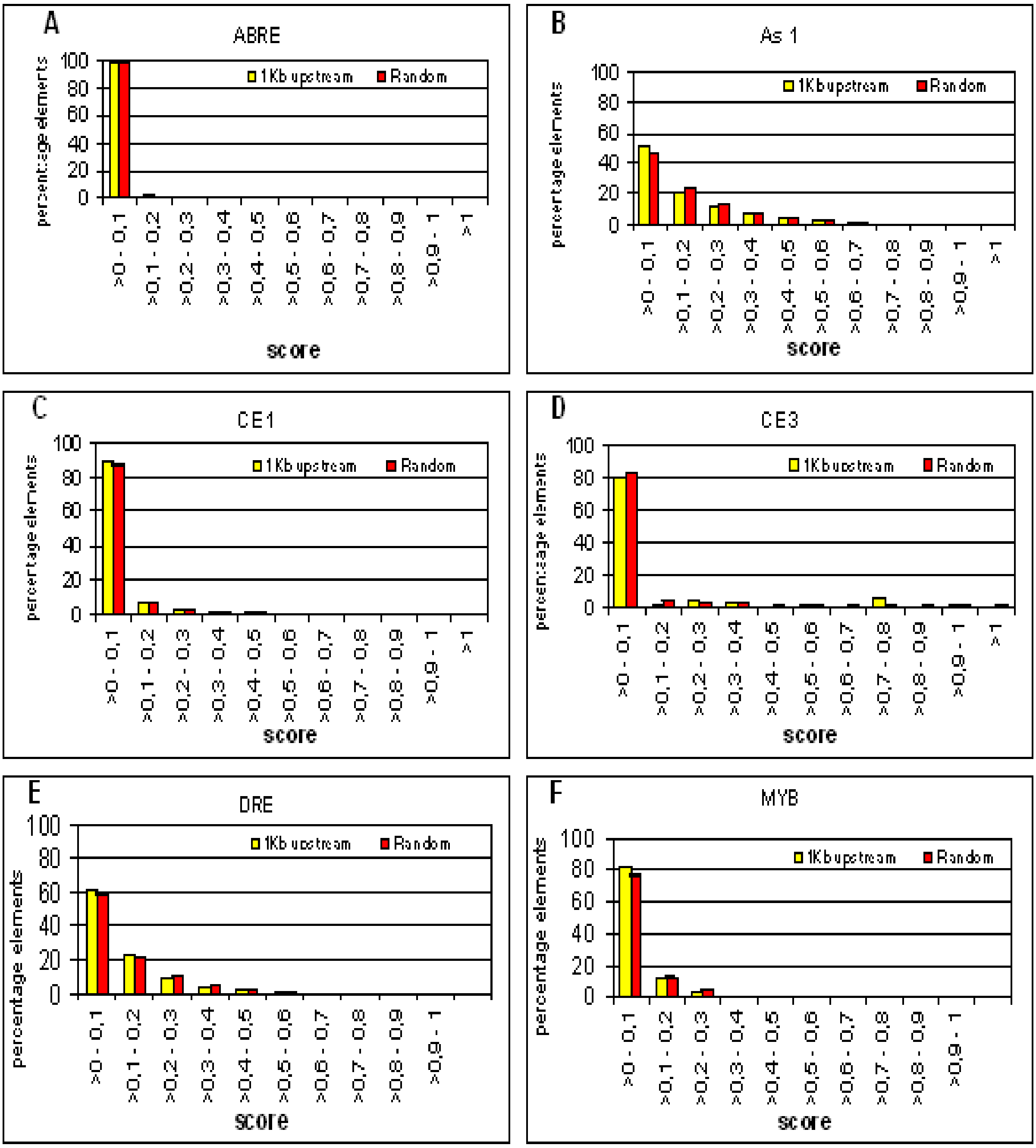


**Figure 6-5 Distribution of rescaled scores for MotifScanner in *A. thaliana* 1 kb upstream sequences and in random datasets.** Scores for A. ABRE. B: As1. C. CE1. D. CE3. E. DRE. F. MYB

The number of predicted *cis*-elements in the subset of 1882 sequences was computed. Furthermore, for each of the hundred shuffled datasets the corresponding 1882 shuffled sequences were extracted. The average number of predicted *cis*-elements in this subset of shuffled sequences was also computed. Results are shown in Table 6-8.

When the results presented in Table 6-7 are compared with the results presented in Table 6-8, it can be observed that for every *cis*-element under study, only approximately 5% of the predicted *cis*-elements were left in either dataset. The number of predicted *cis*-elements in the subset of random datasets was still larger than in 1 kb upstream sequences.

**Table 6-8: Number of predicted *cis*-elements found with MotifScanner in a subset of 1882 1 kb upstream sequences and average number of predicted *cis*-elements in the corresponding subset of shuffled sequences.** Results for random datasets are average values given ± SD

| ***Cis*-element** | **1 kb upstream** | **Random datasets** |
|---|---:|---:|
| ABRE | 31 | 628,5 ± 25 |
| As1 | 3165 | 2713,4 ± 31 |
| CE1 | 782 | 1422,2 ± 23 |
| CE3 | 3 | 4 ± 2 |
| DRE | 266 | 488,2 ± 24 |
| MYB | 333 | 1121 ± 30 |

The score distribution of the subset of 1 kb upstream sequences and the corresponding distribution observed in random datasets were plotted in a histogram, to evaluate whether the results of both datasets were different. The number of predicted *cis*-elements for each score interval was displayed as percentages of the total number of predicted *cis*-elements (Figure 6-6).

The score distribution of the predicted ABRE, CE1 and MYB *cis*-elements (Figure 6-6A, C and F) showed clear differences between the subset of 1 kb upstream sequences and the corresponding random sequences in random datasets. In the case of ABRE (Figure 6-6A) the percentage of predicted *cis*-elements that obtained a score ≤0.1 in 1 kb upstream sequences was reduced from about 98% to about 60%. In contrast, all predicted ABRE *cis*-elements in the set of corresponding shuffled sequences obtained scores below the threshold. In the case of CE1 (Figure 6-6C) the percentage of predicted *cis*-elements that obtained a score equal to or smaller than the threshold of 0.1 in 1 kb upstream sequences is reduced from about 85% to 38%. Most of the CE1 *cis*-elements predicted in the corresponding shuffled sequences achieved a score below the threshold. In the case of MYB (Figure 6-6F) the percentage of predicted *cis*-elements that obtained a score equal to or smaller than the threshold of 0.1 in 1k upstream sequences is reduced from 80% to about 20%, whereas for the corresponding subset of shuffled sequences most of the scores (80%) were equal to or below the threshold. For the *cis*-element As1 (Figure 6-6B) it was observed that even if only a subset of sequences were compared, the distribution of scores was still very similar for 1 kb upstream sequences and random sequences. The *cis*-element CE3 (Figure 6-6D) was very rarely predicted in both datasets (Table 6-7 and Table 6-8). For the 3 hits predicted in the subset of 1 kb upstream sequences the score was equal to 0.1. The predicted CE3 *cis*-elements in the corresponding subset of shuffled sequences obtained higher scores. In the case of DRE (Figure 6-6E) the distribution of scores for the subset of 1 kb upstream sequences and corresponding shuffled sequences showed some differences. About 60% of the predicted *cis*-elements in the subset of random sequences obtained a score equal to or below the threshold of 0.1, compared with 20% in the subset of 1kb upstream sequences. Most of the *cis*-elements predicted in the subset of 1 kb upstream sequences were in the score interval >0.1 to 0.2.

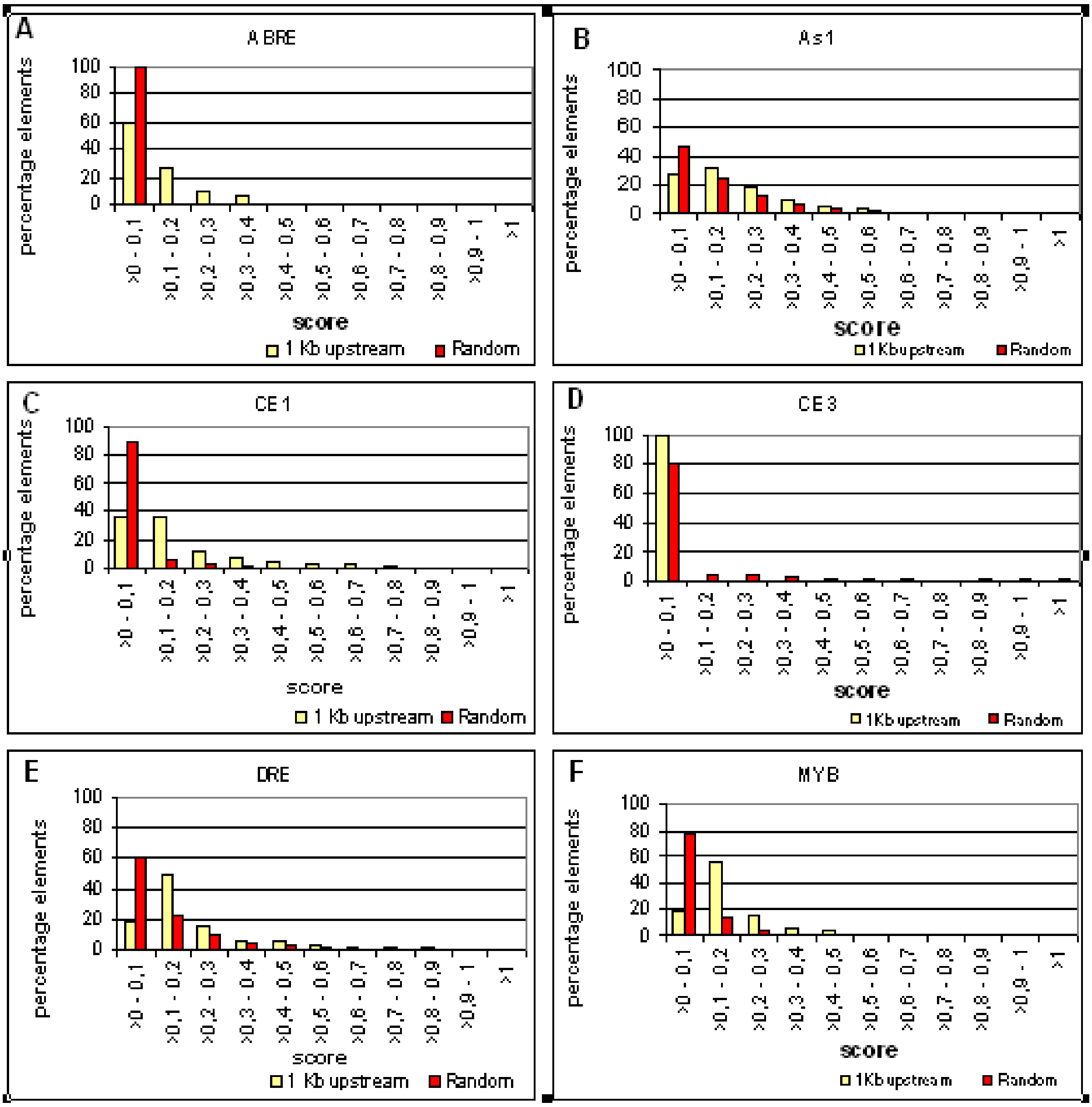


**Figure 6-6: Distribution of rescaled scores for a subset of 1882 1 kb upstream sequences and corresponding shuffled sequences in the random datasets.** Rescaled scores for A. ABRE. B. As1. C. CE1. D: CE3. E. DRE. F. MYB *cis*-elements

In general, even though only a subset of sequences was analysed, some predicted *cis*-elements in the subset of 1 kb upstream sequences showed a score distribution similar to or worse than corresponding random sequences (e.g. *cis*-elements As1 and CE3). Other *cis*-elements such as ABRE, CE1 and MYB showed a score distribution clearly different from that observed for the corresponding random sequences.

To investigate whether the results obtained for both subsets of sequences can be taken as different, a Mantel test[66] was performed.

The Mantel test is computed over two distance matrices. To compute the distance matrix of the subset of 1 kb upstream sequences, each sequence was associated to a 6-dimensional vector. The vector indicates the number of predicted *cis*-elements of each kind. The first dimension refers to the number of predicted ABRE *cis*-elements, the second to the number of predicted As1 *cis*-elements, the third to the number of predicted CE1 *cis*-elements, the fourth to the number of predicted CE3 *cis*-elements, the fifth to the number of predicted DRE *cis*-elements, and the sixth to the number of predicted MYB *cis*-elements. The upper panel in Figure 6-7 shows an example of the scores for each of the predicted *cis*-elements of five sequences (Figure 6-7A), and in the lower panel the resulting 6-dimensional vector of each sequence (Figure 6-7B). The vectors were used to compute the distance matrix between sequences in the dataset. Distance matrices were calculated using three measurements to access similarity: Euclidean distance, Hamming distance and Pearson correlation coefficient.

**A**

| AGI | ABRE | AS1 | CE1 | CE3 | DRE | MYB |
|---|---:|---:|---:|---:|---:|---:|
| At1g01100 | 0 | 0,22 | 0,11 | 0 | 0 | 0 |
| | 0 | 0,49 | 0 | 0 | 0 | 0 |
| At1g01480 | 0 | 0,25 | 0 | 0 | 0 | 0,10 |
| At1g02620 | 0 | 0,11 | 0 | 0 | 0,15 | 0 |
| | 0 | 0,19 | 0 | 0 | 0 | 0 |
| At1g02920 | 0 | 0,54 | 0 | 0 | 0 | 0 |
| | 0 | 0,17 | 0 | 0 | 0 | 0 |
| At1g03010 | 0 | 0,19 | 0 | 0 | 0 | 0 |
| | 0 | 0,10 | 0 | 0 | 0 | 0 |

**B**

| AGI | PATTERN OF COUNTS |
|---|---:|
| At1g01100 | 0 2 1 0 0 0 |
| At1g01480 | 0 1 0 0 0 1 |
| At1g02620 | 0 2 0 0 1 0 |
| At1g02920 | 0 2 0 0 0 0 |
| At1g03010 | 0 2 0 0 0 0 |

**Figure 6-7: MotifScanner results for a subset of 5 sequences (*A. thaliana* 1 kb upstream sequences).** A. Rescaled scores. B. 6-dimensional vector that represents the number of *cis*-element of each kind predicted per sequence, each of the six digits refers to the counts of a given *cis*-element

To compute the 6-dimensional vector for random datasets a slightly different process was followed. First, the number of predicted *cis*-elements of each kind per sequence in each random datasets was counted. Then, the mean number of *cis*-elements of each kind per sequence in all random datasets was computed. Afterwards, each sequence was defined with a 6-dimensional vector. Each dimension corresponded to the mean number of predicted *cis*-elements per sequence, i.e. the first dimension to ABRE, the second to As1, the third to CE1, the fourth to CE3, the fifth to DRE and the sixth to MYB.

The 6-dimensional vectors were used to calculate the distance matrix between sequences in this dataset, referred as mean_random in Table 6-9. The same similarity measurements mentioned above were used to calculate the distance between pairs of sequences.

An additional set of distance matrices was computed for a subset of random sequences (dataset R1). The 6-dimensional vectors corresponded to the number of *cis*-elements per sequence counted in one random dataset. The order of the dimensions is the same as before, as are the similarity measurements used to compute the distance matrices.

The Mantel test assumes that the distances in a matrix A are independent of the distances for the same objects in another matrix B[11,66]. The reason for using different similarity measurements was to test whether the results of the Mantel test are independent of the similarity measurements used. The reason for using two different distance matrices based on results for random datasets was to gain some insights into the expected scores for the test, when the assumption of no correlation between matrices is violated (comparison random vs. random).

Once the distance matrices have been calculated, the computation of the test starts with the random permutation of the rows and corresponding columns of one of the two matrices (arbitrarily chosen). The number of permutations determine the overall precision of the test (Manly, 1997, cited in Bonnet *et al.* 2002[11]). After each permutation the Pearson correlation coefficient between matrix A and matrix B was calculated. In general, the Pearson correlation coefficient between non-correlated matrices is low, and with each permutation is degraded or lost.

The function mantel.randtest of the ade4 software package[21] of the statistical language R[84] was used to compute the correlation coefficient and an associated P-value for the Mantel test. As recommended by Bonnet *et al.* 2002[11] 5000 permutations were carried out. Generally, this number of permutations is considered to produce very robust results for an $\alpha = 0.01$. If the null hypothesis (no correlation between distance matrices) holds, the correlation coefficient between matrices is zero or close to zero. The associated P-value estimates whether the correlation found after the permutations is significant or not. A P-value equal to or below the significance $\alpha$ (0.01 in this case) indicates a significant correlation between matrices. In this case the null hypothesis is rejected. The results of the Mantel test are shown in Table 6-9.

In Table 6-9 the correlation coefficient between the distance matrix for 1 kb upstream sequences and random sequences was around 0,3 independent of the similarity measurement used. These correlation coefficients were significantly different from zero, with a P-value smaller than the $\alpha$=0.01. Therefore, the null hypothesis was rejected. The comparison between random matrices showed a correlation coefficient close to one, with a P-value smaller than the significant $\alpha$ (0.01). The distance matrices in the case of the random datasets were not independent, and clearly the null hypothesis must be rejected.

Taking all results together (distribution of scores and correlation between distance matrices) it had to be concluded that the predictions for 1 kb upstream sequences were statistically similar to predictions for random datasets.

**Table 6-9: Mantel Test for results of MotifScanner.** Matrix A and B correspond to distance matrices. The number of hits per element and sequence were counted for 1882 sequences that achieved a rescaled score ≥0.1 in *A. thaliana* 1 kb upstream sequence. The number of hits per sequence and element were counted for the corresponding shuffled sequences in each random dataset. R1 is the distance matrix calculated from the results observed in random dataset one. Mean random is the distance matrix calculated from the average results observed in the hundred random data sets. Correlation=Pearson correlation coefficient. α=0,01

| Matrix A | Matrix B | Similarity measurement | Correlation | P-value |
|---|---|---|---|---|
| 1 kb upstream | Mean random | Euclidean distance | 0.2749 | <0,01 |
| 1 kb upstream | Mean random | Pearson correlation coefficient | 0.3185 | <0,01 |
| 1 kb upstream | Mean random | Hamming distance | 0.2301 | <0,01 |
| 1 kb upstream | R1 | Euclidean distance | 0.2731 | <0,01 |
| 1 kb upstream | R1 | Pearson correlation coefficient | 0.3164 | <0,01 |
| 1 kb upstream | R1 | Hamming distance | 0.2289 | <0,01 |
| Mean random | R1 | Euclidean distance | 0.9705 | <0,01 |
| Mean random | R1 | Pearson correlation coefficient | 0.9802 | <0,01 |
| Mean random | R1 | Hamming distance | 0.9779 | <0,01 |

Another method was used to evaluate the presence of ABA-related *cis*-elements in 1 kb upstream sequences from *A. thaliana*. The results will be described in the following section. As an important aspect, the method used (based on frequency matrices as MotifScanner) identifies the combination of *cis*-elements simultaneously, and not sequentially.

### 6.4.2 CISTER

CISTER stands for *Cis*-elements Clusters. Martin Frith at the Department of Biomedical Engineering in Boston (USA) implemented the algorithm, and particular features have been described in section 4.1.6.2[34]. Briefly, the query sequence is analysed using a HMM. Parameters such as “mean number of elements expected per sequence“, “mean distance between elements“ and “mean number of clusters of elements“ are considered in the calculations. The scores obtained with CISTER for a predicted binding site give some hints about the similarity of the predicted binding site with the matrix, but also the likelihood of the distribution of elements.

As described in section 4.1.6.2 the parameters “mean number of clusters”, “mean number of elements per cluster”, “mean distance between elements” and “window length on which nucleotide frequencies are counted”, were optimised using a subset of promoters, in which the exact location of the *cis*-elements was known.

In the best results achieved, seven out of ten *cis*-elements were correctly identified. In this case the parameters were following:

1. mean distance between clusters: g=1000;
2. mean distance between elements in a cluster: a=20;
3. mean number of elements in a cluster: b=10;
4. window length were nucleotide frequencies are counted: w=150.

From the analysed 29845 1 kb upstream sequences, 6317 were predicted to have at least one of the *cis*-elements under study. Since this study concentrates on the prediction of genes putatively regulated by ABA, based on the presence of more than one *cis*-element conferring ABA-responsiveness, genes predicted to have exactly one *cis*-element were not considered for further analyses.

About half of the positively predicted sequences have exactly one *cis*-element (3024). Sequences predicted to have more than one *cis*-element (3293) corresponded to 11% of the total number of analysed sequences (29845). This number of positively predicted sequences is evidently lower than the number of positively predicted sequences found using MotifScanner (see section 6.4.1). Using MotifScanner 86% of the analysed sequences were predicted to have more than one ABA-related *cis*-elements.

As mentioned before, with CISTER the probability of each match reflects the similarity of the matching subsequence and the motif model (frequency matrix), but it also reflects the similarity of the query sequence with the model distribution of *cis*-elements. The probability threshold for a match is 0.1. In Figure 6-8 the frequency for each observed probability was plotted in a histogram, and displayed as percentages of the predicted number of *cis*-elements.

Figure 6-8 indicates that independent of the considered *cis*-element only few predictions achieved a probability equal to the threshold (0.1) or reached a high probability (higher than 0.8). The percentage of predicted *cis*-elements that achieved a probability between 0.21 and 0.3 was very similar for all *cis*-elements studied in the course of this work. Interestingly, some pairs of *cis*-elements showed practically the same probability distribution, e.g. ABRE and DRE, CE1 and CE3 and, As1 and MYB.

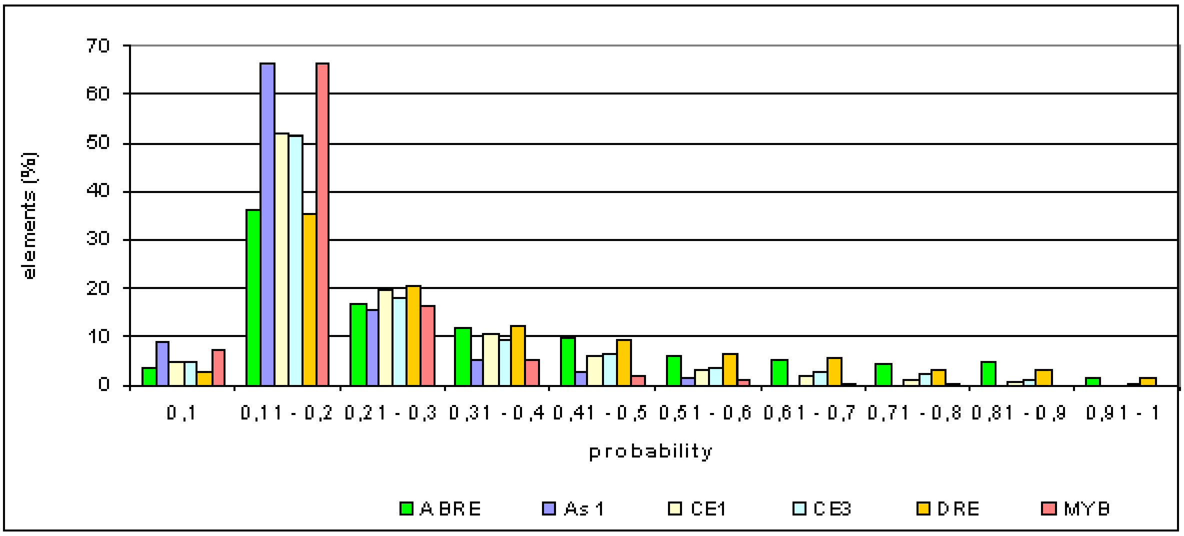


**Figure 6-8: Distribution of probabilities for CISTER results.** Probabilities observed for *A. thaliana* 1 kb upstream sequences

The program CISTER was also used to predict *cis*-elements in the hundred random datasets generated by shuffling. For CISTER as well as for MotifScanner the shuffling process generated more putative binding sites in the set of random sequences (7448.9 ±70 positive sequences per random dataset, compared to 6317 in *A. thaliana* 1 kb upstream sequences). Differences in the number of predicted sequences having more than one *cis*-element were not very pronounced between both datasets (upstream sequences or shuffled sequences). In the case of random datasets 3469 ±56 sequences were predicted to have only one *cis*-element, while in 1 kb upstream 3293 sequences were predicted.

The screening of both datasets with MotifScanner showed that on average not only more sequences were predicted in random datasets, but also more *cis*-elements (Table 6-7). Results obtained with CISTER were different on that respect. Although the number of positively predicted sequences in 1kb upstream was smaller (3293), more putative *cis*-elements were predicted compared with random datasets. The only exceptions were the *cis*-elements As1 and DRE (Table 6-10). Furthermore, the number of predicted *cis*-elements did not show the extreme differences that were evident using MotifScanner. The *cis*-elements As1 and CE1 for example, were predicted by MotifScanner almost five thousand times more often than the most rarely predicted *cis*-element CE3.

To test whether the inclusion of a spatial model for the distribution of the *cis*-elements results in a better discrimination between probabilities observed in 1 kb upstream sequences compared with random datasets, the probability distributions for each predicted *cis*-element were compared. Results are shown in Figure 6-9.

**Table 6-10: Number of *cis*-elements predicted with CISTER in *A. thaliana* 1 kb upstream sequences and in random datasets.** Average number of *cis*-elements in random datasets are given ± SD

| *Cis*-element | 1 kb upstream | Random datasets |
|---|---:|---:|
| ABRE | 3208 | 2340,8 ± 69,6 |
| As1 | 1369 | 1175,5 ± 45,8 |
| CE1 | 2815 | 2104,6 ± 57,7 |
| CE3 | 1974 | 2069,1 ± 54,6 |
| DRE | 2277 | 2752,4 ± 63,9 |
| MYB | 1647 | 1569,9 ± 48,4 |

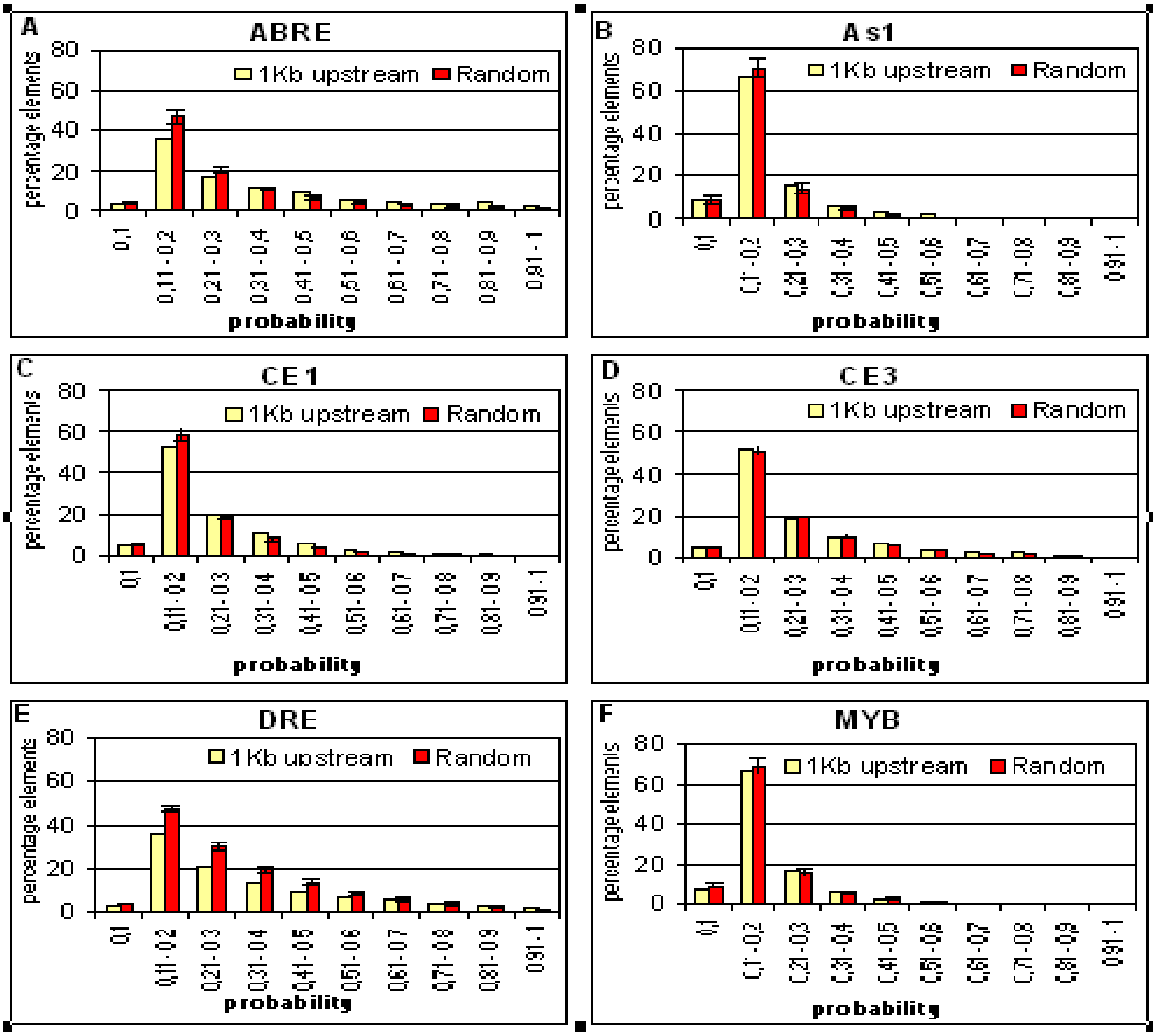


**Figure 6-9: Distribution of probabilities for CISTER results.** Probabilities for A. ABRE. B: As1. C. CE1. D. CE3. E. DRE. F. MYB in 1 kb upstream sequences and in random datasets

Despite the introduction of a spatial model, the distribution of probabilities in both datasets followed the same tendencies. Some slight differences were observed for the *cis*-elements ABRE and CE1 (Figure 6-9A, C), where the percentage of predicted *cis*-elements that achieved high probability values in the set of 1 kb upstream sequences was larger than in random datasets. In contrast, predictions of DRE achieved always better scores in the set of random datasets (Figure 6-9E).

To test whether different results are observed if only a subset of sequences is compared, the positive sequences predicted with CISTER were compared with the results for the corresponding shuffled sequences in each random dataset. The results in Table 6-11 shown that the average number of putative *cis*-elements predicted in the subset of shuffled sequences corresponding to the positive 1 kb upstream sequences was dramatically smaller than in 1 kb upstream sequences. The largest difference (13-fold) was observed in the number of putative ABRE *cis*-elements predicted with CISTER compared with results observed for the the corresponding shuffled sequences. Nevertheless, for all *cis*-element here under study, the differences in the number of predicted *cis*-elements in both subsets of sequences was generally close to an order of magnitude (see Table 6-11).

To assess if the differences in the number of predicted sequences in 1 kb upstream sequences and in corresponding suffled sequences were statistically significant, a Mantel test[66] was performed. The test is calculated using distance matrices. To calculate the distance matrix in the case of 1 kb upstream sequences, each positive sequence was converted into a 6-dimensional vector that contains the information about the number of ABRE, As1, CE1, CE3, DRE and MYB *cis*-elements observed in each sequence (section 6.4.1).

**Table 6-11: CISTER, number of predicted *cis*-elements in 1 kb upstream sequences and in the corresponding shuffled sequences of hundred random datasets.** Average number of predicted *cis*-elements in the subset of random datasets are given ± SD

| ***Cis*-element** | **1 kb upstream** | **Random datasets** |
|---|---:|---:|
| ABRE | 3208 | 238 ± 22 |
| As1 | 1369 | 120 ± 14 |
| CE1 | 2815 | 216 ± 20 |
| CE3 | 1974 | 210 ± 19 |
| DRE | 2277 | 280 ± 21 |
| MYB | 1647 | 163 ±18 |

Distance matrices were calculated using Euclidean and the Hamming distances. The Pearson correlation coefficient could not be used, because the distance matrix in the case of random datasets could not be computed

Similar procedures were used than for caculations of the Mantel test for results with MotifScanner, including the calculation of two distance matrices for random datasets. One distance matrix was computed using the average counts observed per sequence and per element in random datasets, and is referred as mean_random in Table 6-12. The second distance matrix refers to the results observed in one of the hundred datasets.

As described at the end of the previous section (see section 6.4.1) the Mantel test assumes that the distances in a matrix A are independent of the distances for the same object in matrix B. Objects in one of the matrices are randomly permutated (5000 in this case), and the Pearson correlation coefficient and a P-value for the correlation are calculated after each permutation. As in the previous case, the P-value is assessed by comparison with $\alpha$ ($\alpha$=0.01 in this case). If the P-value is equal to or lower than the $\alpha$, the null hypothesis of independence of the distance matrices is rejected. Results are presented in Table 6-12.

**Table 6-12: Mantel Test for results of CISTER. Matrices A and B correspond to distance matrices.** The number of hits per sequence and element were counted in 1 kb upstream sequences, and in the corresponding shuffled sequences in each random dataset. R1 is the distance matrix calculated from the results observed in the random dataset one. Mean random is the distance matrix calculated from the average results observed in the hundred random data sets. Correlation=Pearson correlation coefficient. α=0,01

| Matrix A | Matrix B | Similarity measurement | Correlation | P-value |
|---|---|---|---|---|
| 1 kb upstream | Mean random | Euclidean | $7.2 \times 10^{-4}$ | 0.454 |
| 1 kb upstream | Mean random | Hamming distance | 0.0087 | 0.04 |
| 1 kb upstream | R1 | Euclidean | $-6.4 \times 10^{-4}$ | 0.515 |
| 1 kb upstream | R1 | Hamming distance | 0.0083 | 0.05 |
| Mean random | R1 | Euclidean | 0.9666 | <0.01 |
| Mean random | R1 | Hamming distance | 0.9682 | <0.01 |

Table 6-12 shows that the correlation coefficient for all comparisons between the distance matrices calculated for random sequences and 1 kb upstream sequences achieved values close to zero independent of the similarity measurement used, and a P-value larger than the $\alpha$=0.01 (rows 1 – 4). Therefore, the null hypothesis of no correlation between distance matrices was accepted. The correlation coefficient in the case of the comparison between random matrices was close to one, and the P-value below the significant $\alpha$, indicating that the null hypothesis of independence between both matrices is not valid (rows 5-6).

With the results obtained for the Mantel test it can be concluded that the distances in the matrix of 1 kb upstream sequences were independent of the distances for the same objects in either random matrix.

After verifying that the observed differences between random datasets and 1 kb upstream sequences were statistically significant, the distance matrices calculated for 1 kb upstream sequences were used to cluster the results, with the aim to find sequences with similar combinations of *cis*-elements.

Distance matrices were clustered using the k-means clustering algorithm of the statistical program R[84]. To evaluate the best number of clusters that fits the data, the silhouette coefficient was calculated when the data were grouped in 2 to 100 clusters. The silhouette coefficient is a measurement of the compactness of the clusters. It is defined as the ratio between the inter-cluster distances (that should be maximized) and the intra-cluster distances (that should be minimized)[93], and the values obtained are between zero and one. A silhouette coefficient close to one indicate compact clusters, with small intra-cluster distances, and large inter-cluster distances. Some of the computed results of the silhouette coefficient are shown in Table 6-13. According to the silhouette coefficient the best numbers of clusters that fit the data is 60. A number of cluster above or below 60 results in a small silhouette coefficient (silhouette coefficients for k < 60 are no shown in Table 6-13)

**Table 6-13: Average silhouette calculated for different cluster number.** The distance matrix calculated for the Mantel test were used to calculate the silhouette coefficient. Similarity measurement: Euclidean distance. k= cluster number

| k | Silhouette |
|---|---|
| 60 | 0.91 |
| 62 | 0.85 |
| 64 | 0.79 |
| 66 | 0.77 |
| 68 | 0.75 |

As expected, sequences belonging to the same cluster were predicted to have the same assembly of *cis*-elements (see Appendix 2). In order to asses the degree of separation between clusters, Principal Component Analysis (PCA) was used. PCA revealed that most of the clusters were very close (Figure 6-10). Only the clusters 59, 25, 19, 5 and 46 were relatively far from the other clusters. The sequences included in each clusters showed the following combinations of ABA-related *cis*-elements:

Cluster 59: As1, CE1, CE3, DRE and MYB (no ABRE *cis*-element)
Cluster 25: CE1 and MYB binding sites.
Cluster 19: CE1.
Cluster 46: As1 and MYB binding sites.
Cluster 5: All elements with the exception of CE3.

The number of sequences per cluster was normally less than hundred, only the clusters 1, 9, 12, 15, 19, 21, 24, 27 and 30 grouped more than 100 sequences (maximum 217 in cluster 15), indicating that ABA-related *cis*-elements can combine in a relatively large number of arrangements. However, the differences between combinations are very small.

The functional annotations found for genes belonging to the same cluster were explored. Functional categories were retrieved from the Gene Ontology annotation (GO)[7]. The categories employed referred to the third level in the GO hierarchy. It was observed that all clusters showed very similar GO-categories. The most abundantly observed were “metabolism”, “intracellular”, “membrane”, “cell growth and/or maintenance” and “nucleic acid binding”.

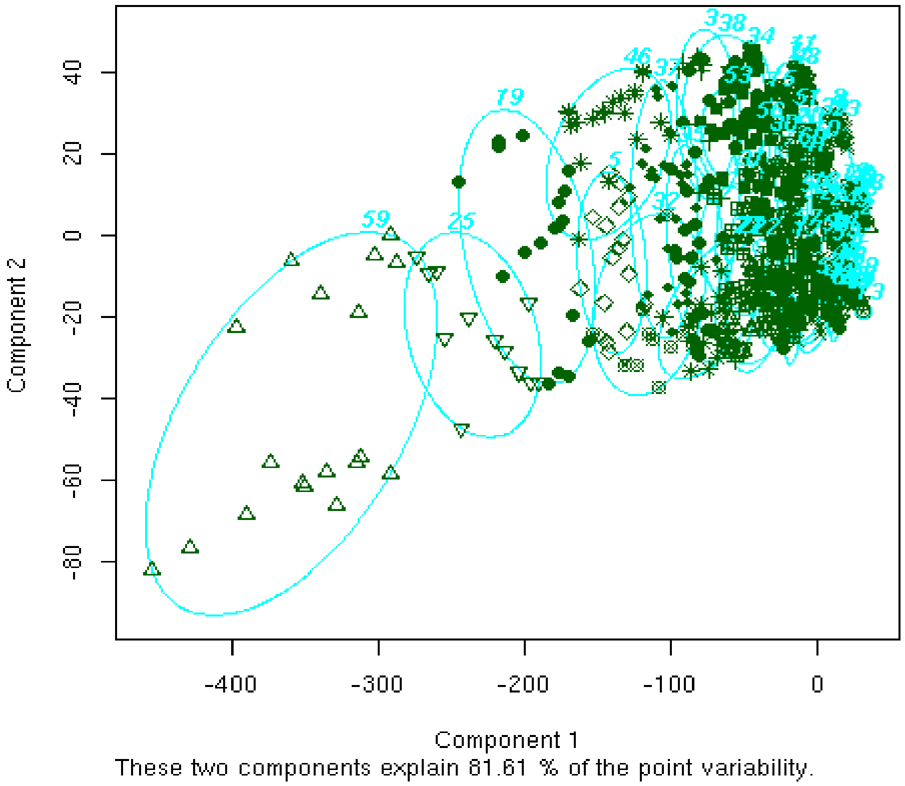


**Figure 6-10: Principal component analysis.** Clustering of sequences positively predicted by CISTER. The number that corresponds to each ellipse indicates the cluster number, and the symbols represent the member of each cluster. Similarity measurement: Euclidean distance. Number of clusters k=60

## 6.5 Conclusions

1. Some pairs of ABA-related *cis*-elements were found significantly over-represented in 1 kb upstream sequences, compared with background expectations. Almost all combinations of ABRE with other ABA-related *cis*-elements were over-represented. However, the pairs ABRE-CE3, CE3-ABRE were not observed in the set of sequences evaluated, and the pair ABRE-As1 was not over-represented.

2. Among the over-represented pairs, the homologous pair ABRE-ABRE showed the highest $S_{ij}$ score computed, indicating that the elements preferentially occur together than independently. Furthermore, around 60% of the pairs showed a distance between *cis*-elements <50 bp. Considering that ABRE needs a coupling element to activate the transcription of its gene, it seems that ABRE is the most important coupling element of itself in *A. thaliana.*

3. Another over-represented pair with a high $S_{ij}$ score was the pair MYB-ABRE. The distance between the two *cis*-elements showed a tendency to find both *cis*-elements of the pair separated by the following distances: 150 to 250 bp, or 400 to 450. These relatively large distances between *cis*-elements suggested that MYB does not function as a coupling element of ABRE. Instead, it might be involved in the slow induction of ABA-regulated genes, after accumulation of ABA in plants subjected to osmotic stress.

4. The reciprocal pair ABRE – MYB was also significantly over-represented. However, a $S_{ij}$ score of zero clearly indicated that ABRE and MYB (in that order) do not preferentially occur together. Additionally, the distances between the *cis*-elements were not largely different from results observed in random datasets. These results underlined the importance of the order of the elements in regulatory sequences, as has been outlined previously by Shen *et al.*2004[103].

5. Pairs of the *cis*-elements ABRE and CE1 were over-represented in 1 kb upstream sequences. However, $S_{ij}$ scores calculated for the reciprocal pairs ABRE-CE1 and CE1-ABRE were relatively low. The large distances between the *cis*-elements indicate that CE1 is perhaps not such a relevant coupling element of ABRE in *A. thaliana*, in contrast with results observed in monocotyledoneous species. Furthermore, the other coupling element studied here (CE3) was not found in association with ABRE, and no CE has been experimentally characterized in *A. thaliana*. Therefore, it appears that in contrast to observations in monocots, coupling

elements might not play a major role in the regulation of ABA-responsive genes in *A. thaliana*.

6. The *cis*-elements ABRE and DRE were significantly over-represented in both spatial arrangements. The absence of short distances between the elements of the pair DRE–ABRE might suggest a less relevant role of DRE as a coupling element of ABRE in *A. thaliana*. This characteristic, together with the avoidance of large distances between *cis*-elements, and positive but low $S_{ij}$ scores, support the observations made by Narusaka *et al.* 2003[75] that ABRE and DRE in regulatory sequences allowed the cross-talk between osmotic stresses (such as drought, cold and high salinity), making genes responsive to such stresses in ABA-dependent and ABA-independent signalling pathways.

7. The poor performance of the program MotifScanner, where the predictions could not be differentiated from results for random sequences, clearly demonstrates that the identification of putative *cis*-elements must not rely merely on sequence similarity between the motif model and the query sequence. Other features, like probability of the combination of *cis*-elements compared with a suitable background model might improve the performance of the program enormously.

8. The results gained with the program CISTER, where the number of upstream sequences predicted to have ABA-related *cis*-elements was visibly smaller than with the other programs used, underlining the importance of including additional features for the prediction of putative transcription factor binding sites.

# Chapter 7: Expression profiling in *A. thaliana* leaves after abscisic acid treatment

*In-house performed macroarray data were analysed to establish ABA regulated genes in leaves[78]. Raw expression data were analysed to determine if observed gene expression changes were attributable to ABA treatment. Genes regulated as a consequence of the treatment were clustered according to their temporal expression pattern. Promoters of the genes regulated were analysed, to establish putative regulatory sequences. The responsiveness of some genes was independently confirmed via RT-PCR. Finally, the cross-talk with other hormone signals was assayed comparing the results observed in this study with results observed in microarray and northern blot experiments for auxins, jasmonate, brassinosteroids or ethylene treatment.*

## *7.1 General strategy*

In a previous research project macroarray experiments were carried out in house for the large-scale detection of *A. thaliana* genes differentially regulated by ABA[78]. Briefly, nylon membranes were spotted with about 16.000 single-strand cDNAs from a collection of cDNAs originally generated at Michigan State University – MSU[76]. The collection includes expressed sequence tags (ESTs) from different tissues and developmental stages, including seedlings, rosettes, stems, flowers, and roots from plants of different ages.

Experiments were performed in duplicates, to account for technical and biological variation. Nylon membranes were hybridised with two kinds of probes. The first hybridisation, called reference hybridisation, was performed to determine the amount of cDNA spotted at each spot onto the nylon membranes. The probes were short oligonucleotides, which were homologous to a part of the vector sequence surrounding the ESTs. The second hybridisation, called complex hybridisation, was performed with a set of radiolabeled cDNAs derived from total RNA isolated from ABA-treated or untreated plants.

After hybridisation with radiolabeled probes, macroarray filters were scanned with a phosphorimager (Fuji, Japan) and raw images were analysed with the AIS Image Processing System. The program semi-automatically identifies each spot of the filter and integrates their signal intensity. Data were normalized using the software Haruspex, developed at the Max-Planck Institute of Molecular Plant Physiology, Golm, Germany (S. Kloska, B. Essigmann and T. Altmann, unpublished data).

## 7.2 Biological / technical variation

Five time points were measured (30, 60, 90, 120 and 300 min after ABA treatment). For each of the time points, two independent hybridisation experiments were conducted. To evaluate the biological and/or technical variability, the magnitude of the radioactivity (gene activity according to the Haruspex software) was compared with scatter plots (Figure 7-1A – E). The gene activity of each clone was compared with itself on membranes hybridised with cDNA from plants subjected to the same treatment. In Figure 7-1 the x-axis represents the logarithm of gene activity measured in membrane H2, the y-axis represents the logarithm of gene activity measured in membrane H1, both membranes were hybridised with cDNA from plants treated with ABA. For each time point, the variability between hybridisations was not above or below the threshold of a 3-fold change. This threshold was chosen according to the recommendations given by Thimm *et al.* 2001[116]. Thus, for each time point the technical or biological variability inherent to the experimental procedures (e.g. biological variability, efficiency of cDNA synthesis, amount of cDNA bound to the membrane) was not significantly different.

Membranes subjected to different treatments were also compared with scatter plots. For each clone subjected to the same treatment the average gene activity was calculated, and average results were plotted. In Figure 7-2A – E, the x-axis represents the average gene activity for clones hybridised with cDNA from untreated plants, logarithmically transformed. The y-axis represents the average gene activity for clones hybridised with cDNA from ABA-treated plants, logarithmic transformed.

Clear changes in expression levels could be observed for all time points. For some clones the average gene activity was below or above the chosen threshold (3-fold change). Clones that showed a change in gene activity above the threshold were considered as up-regulated by the treatment, and clones that showed a change in gene activity below the threshold were considered as down-regulated by the treatment.

The scatter plot analysis clearly showed that the observed differences in gene expression can be attributed to the ABA treatment. In contrast, biological variation (RNA was obtained from independently grown plants) or technical variation (membranes were hybridised independently) did not largely result in differences in gene activity above or below the 3-fold change threshold.

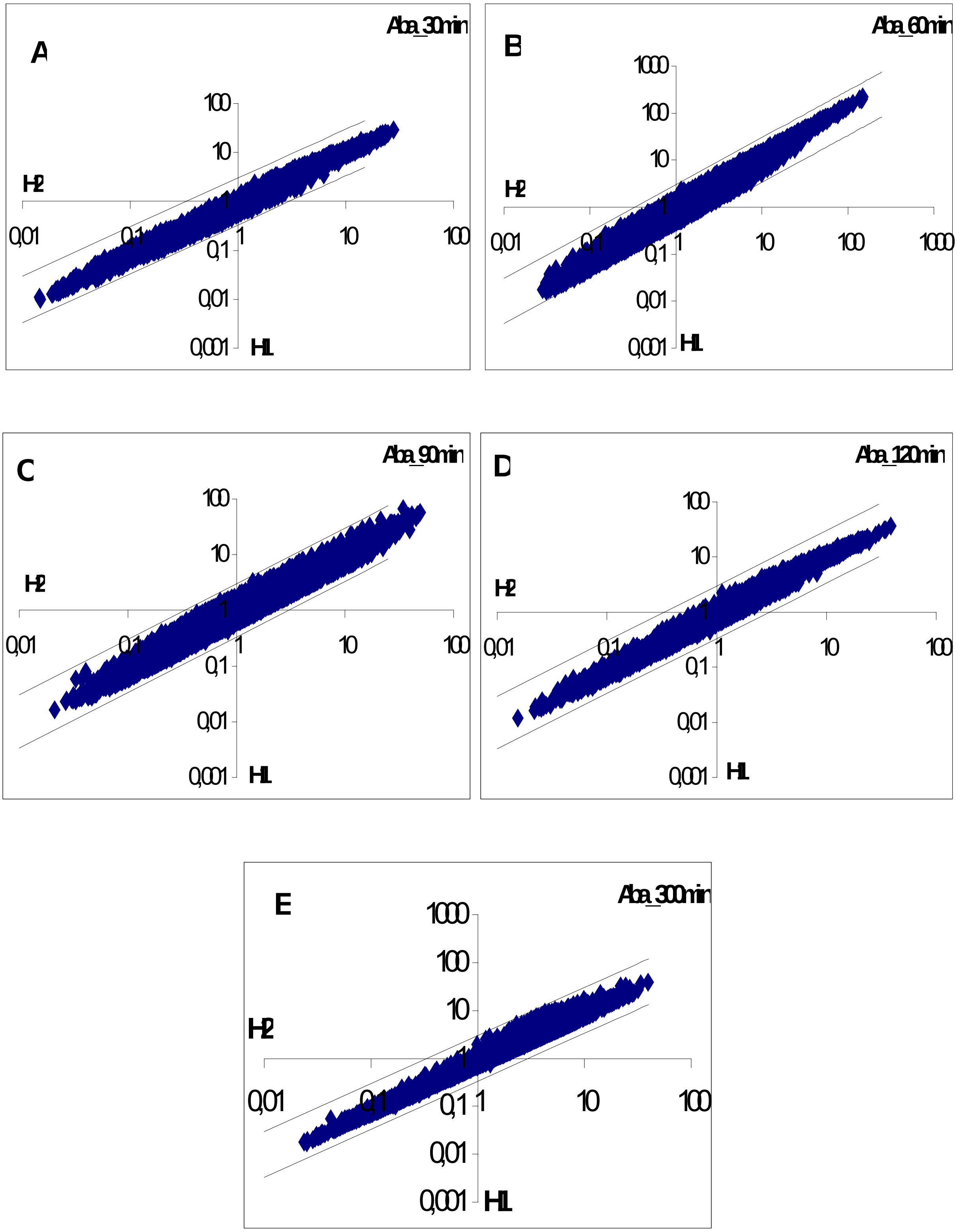


**Figure 7-1: Gene activity for membranes H1 versus H2 hybridised with cDNA from ABA-treated plants (Scatter plots).** A. 30. B. 60. C. 90. D. 120. E. 300 min after treatment. To indicate the significance threshold the guide lines y=3x and y=x/3 were added to the figures

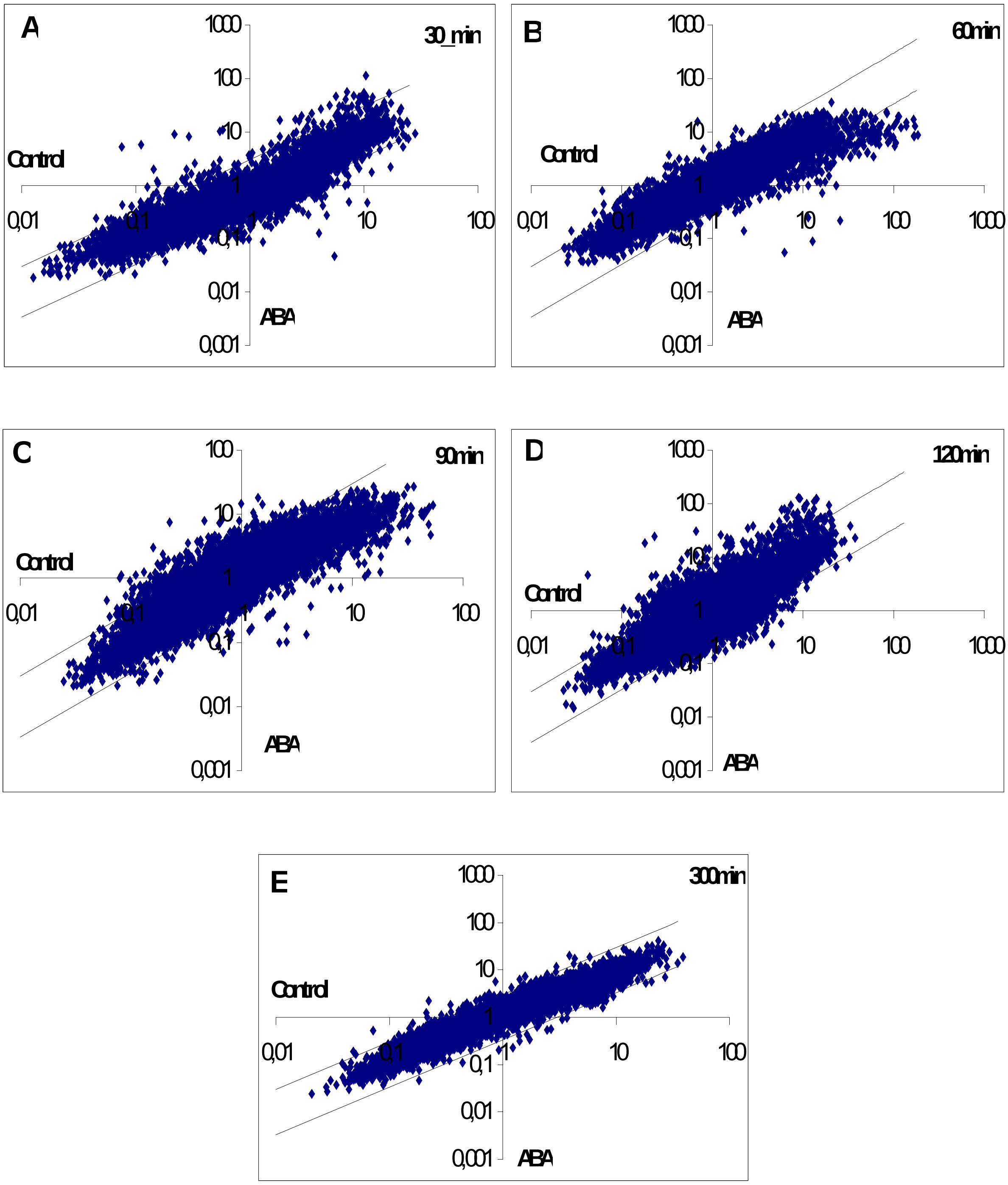


**Figure 7-2: Gene activity for membranes hybridised with cDNA from plants treated with ABA vs. membranes hybridised with cDNA from control plants (Scatter plots).** Average gene activity for two membranes hybridised with either cDNA. x-axis membranes hybridised with cDNA from control plants, y-axis membranes hybridised with cDNA from ABA-treated plants. A. 30.. B. 60. C. 90. D. 120. E. 300 after treatment. To indicate the significance threshold the guide lines y=3x and y=x/3 were added to the figures

## 7.3 ABA-regulated genes

In the previous section it was shown that the observed differences in gene activity were due to the ABA treatment, and not to technical or biological variations. For the identification of genes affected by the treatment, only clones that were successfully measured at each time point, and in each treatment were kept for further analysis.

This procedure reduced the dataset by about 38%, since only around 10.000 from the initial 16.129 clones spotted obtained a gene activity above twice the local background. For the other clones spotted, not enough product was spotted on at least one membrane.

MSU clone names were translated into gene numbers, according to *A. thaliana* gene identifiers established in the Arabidopsis Genome Initiative[1]. The 10.000 measured clones correspond to 4757 genes, meaning that not each clone corresponded to a different *A. thaliana* gene.

For each membrane and time point, gene activity of clones that corresponded to the same gene were merged by their mean. Importantly, before merging the values, it was checked that each single value was not larger than twice the standard deviation of the other measurements for the same gene. If the gene activity was larger than twice the standard deviation, the corresponding value was deleted and re-calculated using the KNN method[120]. The inspection of replicates of a gene was done automatically, with GEPAS version 1.0 (http://gepas.bioinfo.cnio.es/cgi-bin/preprocess).

A ratio of expression was calculated as the mean gene activity of gene *i* measured in treated membranes, divided by the mean gene activity of gene *i* measured in control membranes. Ratios were logarithmically transformed on a $\log_2$ basis ($\mathrm{Log}_2$). Based on the ratio of gene expression ($\mathrm{Log}_2$U), the genes regulated by the ABA treatment were those with a ratio $\geq$1.58 $\mathrm{Log}_2$U (3-fold change) in the case of up-regulation, or with a ratio $\leq$-1.58 $\mathrm{Log}_2$U in the case of down-regulation. From the 4757 genes measured, 680 were regulated in at least one time point. The complete list of genes regulated with their corresponding ORF annotation can be consulted on-line under:

www.bio-uni-potsdam.de/jgomez/abaregulated.html

In Figure 7-3 the percentage of genes regulated at each time point is presented. It was observed that after 30 and 60 min of ABA treatment most of the regulated genes showed an increase in transcript level, while the number of down-regulated genes was very low (less than 5 % 30 min after treatment, and 10 % 60 min after treatment). Ninety min after treatment the proportion of down-regulated genes was higher than that of up-regulated genes. After 120 min of treatment the number of down-regulated genes increased again, and reached its maximum. Five hours after treatment the proportion of regulated genes was the smallest for all time points tested, and most of the genes were down-regulated.

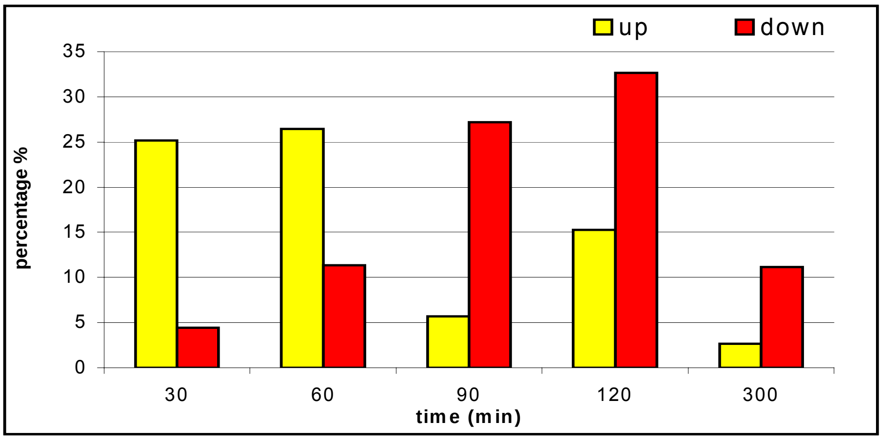


**Figure 7-3: Percentage of genes differentially regulated after ABA treatment**

### 7.3.1 Time-dependent analysis of genes regulated by ABA

Different patterns of expression were observed during the time course of the experiment. For the 680 genes regulated by ABA in at least one time point, genes exhibiting a similar expression pattern were grouped together. Based on the observations in Figure 7-3, the 90 min time point was chosen as transition time point in gene expression, because before 90 min of treatment most of the ABA-responsive genes were up-regulated, and 90 min after treatment most responsive genes were down-regulated. Taking this into account, the following groups of expression profiles were established:

Group 1: Genes that are predominantly **down-regulated**. This group comprises genes that were down-regulated in at least four time points, and two genes down-regulated at three time points (60, 90 and 300 min or 30, 60 and 90 min after treatment respectively). At the other time points the expression of these genes remained stable. Thirteen genes belonged to this group (see Figure 7-4A). Among them, six were down-regulated during the course of the experiment, five were down-regulated in four out of five time points, and two at three out of five time points. Importantly, other genes down-regulated at three time points were included in groups 4 or 5, because they were down-regulated at the late phase of the experiment (starting 90 min after treatment), or transiently down-regulated at the beginning of the experiment (30 min), and at the end (120 and 300 min after treatment).

Group 2: Genes that are predominantly **up-regulated**. This group comprises the genes that were up-regulated in at least four time points. At the other time points the expression of these genes remained stable. Seven genes belonged to this group (see Figure 7-4B). Among them, six were up-regulated during the course of the experiment, and one was up regulated in four out of five time points. Up-regulated genes at three time points were included in group 8, because they were found transiently up-regulated at the beginning of the experiment (30 and 60 min), and at the end (300 min).

Group 3: **Early responsive genes**. This category includes genes that responded 30 and/or 60 min after treatment. For the other measured time points the expression remained stable. Hundred eighty seven genes belonged to this group (see Figure 7-4C). Regarding the up-regulated genes, 75 were up-regulated 30 min after treatment. Fifty one genes were up-regulated 60 min after treatment. Twenty genes were up-regulated at both time points (30 and 60 min after treatment). Regarding down-regulated genes, 9 genes were down-regulated 30 min after treatment, 27 genes were down-regulated 60 min after treatment, and 2 genes were down-regulated at both time points. Finally, 3 genes were up-regulated 30 min after treatment, and then down-regulated 60 min after treatment.

Group 4: **Late responsive genes**. This category includes genes responding 90 min after treatment. This is the biggest group with three hundred and one genes (Figure 7-4D - E). Among these genes, 180 were down-regulated at one time point (either 90, 120 or 300 min after treatment), 45 genes were up-regulated at one time point (either 90, 120 or 300 min after treatment), 70 genes were regulated at two time points, and the most common profile was down-regulation 90 min after treatment and up-regulation 120 min after treatment (37 genes). Finally, only 6 genes were regulated at three time points, three of them were down-regulated.

Group 5: **Transient down-regulation**. As described at the beginning of the section, 90 min was chosen as the transition time point. In this category genes down-regulated in the early phase of the ABA treatment (30 and/or 60 min after treatment), and again down-regulated in the late phase of the stimulus (either time point 90 min after treatment) were included. Twenty five genes belonged to this group (see Figure 7-4F). Patterns can be sub-divided into two groups, as shown in Figure 7-4F. Genes down-regulated 30 and 90 min after treatments are shown in blue, and genes down-regulated at 60 and 120 min are indicated in black. Most genes included in this group were down-regulated only at two time points (19 genes).

Group 6: **Transient down/up regulation**. This group includes genes that were down-regulated in the early phase of the experiment (30 and/or 60 min after treatment), but up-regulated in the late phase of the experiment (90 min or more after treatment). Eleven genes belonged to this group (Figure 7-4G). Most genes were down-regulated 60 min after treatment and up-regulated 120 min after treatment (7 genes).

Group 7: **Transient up/down regulation.** This group includes genes up-regulated in the early phase of the experiment (30 and/or 60 min after treatment), but down-regulated in the late phase of the experiment (90 min or more after treatment). Hundred seven genes belonged to this group (see Figure 7-4H). About sixty percent of them were up-regulated 60 min after treatment, and down-regulated 120 min after treatment (68 genes).

Group 8: **Transient up-regulation**. This group includes up-regulated genes in the early phase of the experiment (30 and/or 60 min after treatment), that are mostly stable 90 min after treatment, and then up regulated again in the late phase of the experiment (120 or 300 min after treatment). Eight genes belonged to this group (see Figure 7-4I). Only one gene was up-regulated 90 min after treatment, and the expression pattern is shown in blue. Most genes were up-regulated 30 and 120 min after treatment.

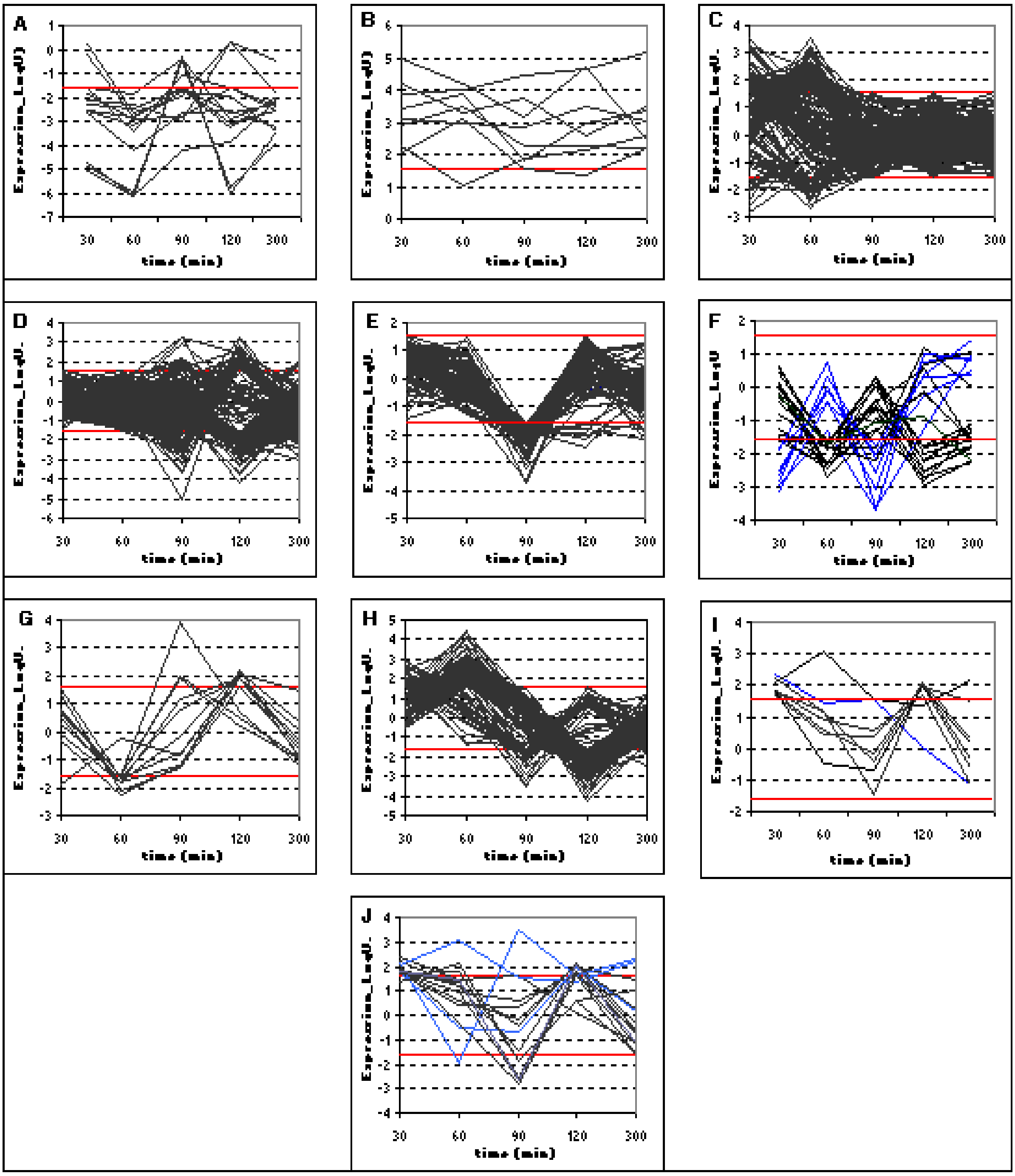


**Figure 7-4. Time-dependent expression patterns, genes regulated by ABA.** Superimposed patterns of expression. A. group 1, genes predominantly down-regulated. B. group 2, genes predominantly up-regulated. C. group 3, early-responsive genes (30 and/or 60 min after treatment). D. group 4, late-responsive genes (up/down regulated or both 90 min or more after treatment). E. group 4, late-responsive genes down-regulated (90 min or more after treatment). F. group 5, transient down-regulated genes. G. group 6, transient down/up-regulated genes. H. group 7, transient up/down-regulated genes. I. Group 8, transient up regulated genes. J. group 9, oscillating patterns. Red lines mark the threshold of up- or down-regulation (±1.58Log$_2$U)

Group 9: **Oscillating patterns**. This group comprises twenty-one genes (see Figure 7-4J). These genes generally showed the following time course of expression: up-regulation 30 and 120 min after treatment, and down regulation 90 min after treatment (11 genes), and are shown in black in Figure 7-4J. The other patterns of expression were very variable from time point to time point.

Overall results showed that most of the genes were regulated by ABA at one time point, either at the beginning or at the end of the experiment (groups 3 and 4). Most of the genes regulated at two time points were transiently up/down regulated. Only about 10% of the genes regulated by ABA were regulated in more than two time points, showing complex patterns of expression. Within the genes regulated at more than two time points, 2% corresponded to genes regulated at each time point, and belonged to the groups 1, 2 or 9.

### 7.3.2 Functional groups

To analyse the physiological relevance of genes differentially expressed after ABA treatment and to gain some insights in the physiological processes affected by the treatment, the functional categories of the genes regulated by ABA were retrieved from the Gene Ontology annotation (GO)[7]. The categories employed referred to the third level of the GO hierarchy.

A gene product could have been assigned to one or more of 107 functional groups (GO categories). Genes that were regulated by ABA were grouped into 60 GO categories, including one category that was regarded as NN (no functional assignment in the third level of hierarchy).

The first 25 most abundantly represented GO categories for genes regulated by ABA are presented in Table 7-1.

Most of the genes regulated by ABA belonged to the GO category “metabolism” (13%), the second and third most abundant categories were “intracellular” and “cell growth and/or maintenance”. The category “intracellular” comprises proteins that are connected to any cellular membrane, e.g. plastids, vesicle trafficking, endoplasmic reticulum, etc. The category “cell growth and/or maintenance” compromises proteins involved in cell cycle. Genes without any functional assignment at the third hierarchical level were the fifth most abundant. A closer look in the GO-annotation of genes regarded as NN showed that due to the lower similarity of the coding region of these genes with experimentally determined genes, they are classified in the category “molecular function”, one of the three basic GO-categories at the first level of hierarchy (the other categories are “biological process” or “cellular component”).

**Table 7-1: GO-functional categories for genes regulated by ABA.** No. Genes=number of genes belonging to the category.GO level of hierarchy 3

| GO_category | No. Genes | % |
|---|---|---|
| metabolism | 279 | 17,16 |
| intracellular | 272 | 16,73 |
| cell growth and/or maintenance | 180 | 11,07 |
| membrane | 139 | 8,55 |
| NN | 92 | 5,66 |
| nucleic acid binding | 92 | 5,66 |
| hydrolase activity | 70 | 4,31 |
| transferase activity | 57 | 3,51 |
| nucleotide binding | 54 | 3,32 |
| oxidoreductase activity | 44 | 2,71 |
| response to external stimulus | 42 | 2,58 |
| kinase activity | 29 | 1,78 |
| metal ion binding | 26 | 1,60 |
| response to stress | 24 | 1,47 |
| cell communication | 22 | 1,35 |
| carrier activity | 20 | 1,23 |
| lyase activity | 20 | 1,23 |
| protein binding | 16 | 0,98 |
| ion transporter activity | 13 | 0,80 |
| ligase activity | 13 | 0,80 |
| response to endogenous stimulus | 9 | 0,55 |
| isomerase activity | 8 | 0,49 |
| electron transporter activity | 7 | 0,43 |
| translation factor activity, nucleic acid binding | 7 | 0,43 |
| death | 6 | 0,37 |

Other categories observed between the 25 most abundantly represented were “response to external stimulus”, “responses to stress” and “kinase activity”, as well as a large range of enzymatic categories.

To evaluate whether the cDNA collection spotted onto the membranes has a bias towards one of these different categories, mainly towards metabolic gene products, the GO annotation at the third level of hierarchy for the genes spotted onto the membranes was investigated. Table 7-2 shows the 25 most abundantly represented GO categories for the genes spotted onto the membranes.

**Table 7-2: GO-functional categories for genes spotted onto the membrane.** No. Genes=number of genes belonging to the category.GO level of hierarchy 3

| GO_category | No. Genes | % |
|---|---|---|
| Metabolism | 2605 | 17,35 |
| Intracellular | 2439 | 16,24 |
| Cell growth and/or maintenance | 1566 | 10,43 |
| Membrane | 1248 | 8,31 |
| nucleic acid binding | 826 | 5,50 |
| hydrolase activity | 689 | 4,59 |
| Transferase activity | 637 | 4,24 |
| NN | 635 | 4,23 |
| nucleotide binding | 535 | 3,56 |
| response to external stimulus | 391 | 2,60 |
| oxidoreductase activity | 361 | 2,40 |
| cell communication | 295 | 1,97 |
| metal ion binding | 277 | 1,85 |
| kinase activity | 276 | 1,84 |
| response to stress | 233 | 1,55 |
| carrier activity | 197 | 1,31 |
| response to endogenous stimulus | 155 | 1,03 |
| lyase activity | 139 | 0,93 |
| protein binding | 112 | 0,75 |
| ion transporter activity | 108 | 0,72 |
| isomerase activity | 87 | 0,58 |
| ligase activity | 87 | 0,58 |
| translation factor activity, nucleic acid binding | 72 | 0,48 |
| receptor activity | 67 | 0,45 |
| morphogenesis | 63 | 0,42 |

Table 7-2 shows that the genes corresponding to the cDNAs spotted were classified basically into the same GO categories as genes regulated by ABA. Genes annotated into the category “metabolism” correspond to around 17% of the genes, and this was the most abundant category.

When comparing the 25 most abundant GO categories of genes regulated by ABA with those of genes spotted onto the membranes, the largest differences were found for the categories "electron transporter activity" and "death", which were within the group of the 25 most common categories in the set of regulated genes, but were not found in the group of the 25 most abundant categories of the genes spotted onto the membranes. *Vice versa* from the genes spotted onto the membranes, the categories "receptor activity" and "morphogenesis" were found within the group of the 25 most common categories, whereas they were not observed within the group of the 25 most common categories of the genes regulated by ABA. The analysis of the GO categories for the genes successfully measured confirms that the collection of cDNAs used (MSU collection) has a bias towards genes classified into the GO categories "metabolism", "intracellular", "membrane" and "cell growth and/or maintenance". Furthermore, the ten most abundant GO categories were exactly the same as for genes regulated by ABA or spotted onto the membranes.

The comparison of the datasets revealed that about the same percentage of genes of each GO category was found in either data set. The largest absolute difference was found for genes annotated as NN (no annotation at that level of hierarchy), where 5.7% of the ABA-regulated genes belonged to this category, whereas only 4.2% of the spotted genes belonged to it.

The distribution of GO categories for the groups of expression patterns described in section 7.3.1 was investigated to test whether specific GO categories could be associated to each pattern. It was established that for the groups 1, and 3 - 9 the five most common GO categories were those listed on the top in Table 7-1, i.e., "metabolism", "intracellular", "membrane", "NN" and "cell growth and maintenance". In addition the following categories were particular for each group:

**Group 1** (genes predominantly down-regulated). Were classified in the categories "ion transporter activity", "metal ion binding" and "response to external stimulus". Some examples of transporters predominantly down-regulated by ABA were: "$K^+$ efflux antiporter, putative (KEA4)" (*At2g19600*) and "sulfate transporter family protein" (*At5g13550*).

**Group 2** (genes predominantly up-regulated). This group comprised genes of the categories "metabolism", "intracellular", but also "nucleotide binding", "cell communication" and "hydrolase activity". Examples of predominantly up-regulated "nucleotide binding" genes were: "calcium-dependent protein kinase 19 (CDPK19)" (*At5g19450*) and "expressed protein" (*At5g55540*).

**Group 3** (early responsive genes). Were classified in the categories "hydrolase activity", "oxidoreductase activity" and "transferase activity". Examples of early responsive genes with hydrolase activity were: "protein phosphatase 2C, putative/PP2C" (*At2g25070*) and "serine/threonine protein phosphatase 2A (PP2A) regulatory subunit B' (B'beta)" (*At3g09880*).

**Group 4** (late responsive genes). In addition to the five most common GO categories, the category “response to external stimulus” was one of the most prominent ones. Examples of late responsive genes belonging to the category “response to external stimulus” were “disease resistance protein (TIR-NBS class), putative” (*At1g72890*), “superoxide dismutase (Cu-Zn), chloroplast (SODCP)/copper/zinc superoxide dismutase (CSD2)” (*At2g28190*), and “avirulence- responsive protein, putative/avirulence induced gene (AIG) protein, putative” (*At3g28940*) .

**Group 5** (genes transiently down-regulated). Were classified in the categories “channel/pore class transporter activity” and “carrier activity”. An example of a gene transiently down-regulated by ABA belonging to the category “carrier activity” was “calcium-transporting ATPase1, plasma membrane-type/Ca(2+)-ATPase isoform 1 (ACA1)/plastid envelope ATPase1 (PEA1)” (*At1g27770*).

**Group 6** (genes transiently down/up regulated). Considering the physiological role of ABA in seed maturation and dormancy[135], it was interesting to observe that in this group the GO categories “germination”, “post embryonic development” and “response to external stimulus” were prominent. An example of a gene found transiently down/up regulated that belonged to this group is “Dof zinc finger protein DAG2/Dof affecting germination 2 (DAG2)” (*At2g46590*).

**Group 7** (genes transiently up/down regulated). In addition to the five most common GO categories, the GO categories “kinase activity”, “nucleotide binding” and “response to external stimulus” were observed. Examples of genes classified as kinases were “leucine-rich repeat transmembrane protein kinase, putative” (*At2g02220*), “protein kinase, putative” (*At2g05940*), “serine/threonine/tyrosine kinase, putative” (*At2g24360*) and “CBL-interacting protein kinase 7 (CIPK7)” (*At3g23000*).

**Group 8** (genes transient up-regulated). Genes belonging to this group were also classified in the GO category “oxidoreductase activity”. An example was “NADH-ubiquinone oxidoreductase B8 subunit, putative” (*At5g47890*).

**Group 9** (oscillating patterns). Genes of this group were also classified in the GO categories “kinase activity” and “nucleotide binding”. Some examples of kinases were “S-locus protein kinase, putative” (*At4g27300*) and “protein kinase family protein” (*At5g11850*).

So far the GO annotations found for genes regulated by ABA do not seem to be statistically different from the annotations found for genes of the cDNA collection used. To verify that there are no statistically significant differences in GO annotations, the program GOSSIP was used[10]. The algorithm uses all categories in the gene ontology (GO) to tests if any enrichment of terms is found in a test group compared to the annotations in a reference group. The program calculates a P-value using the one-side Fisher exact test. The null hypothesis is that the annotations of the test group are sampled randomly from the reference group. The significance is measured according to the false discovery rate (FDR) that is calculated for each P-value, and quantifies the expected number of false discoveries in relation to the total number of positives at a given P-value. The FDR is kept below an $\alpha$ threshold (in this case 5 percent or 0.05). Categories significantly enriched are those with a FDR below the $\alpha$ threshold[10].

For the first test, the set of spotted genes was set as the reference group, and the set of ABA regulated genes was set as the test group. It was observed that the FDR was close to 0.98 for all GO categories. Clearly, none of the categories had a FDR below the threshold of 0.05. Thus, it was confirmed that there is not a significant enrichment of categories in the set of regulated genes, compared with spotted genes.

The same test was used to analyse whether there is an enrichment of GO categories in the set of ABA-regulated genes compared with the GO categories for all genes in the genome. To perform the test, the reference group was set as the whole genome, and the test group was set as the ABA-regulated genes.

This comparison showed that there are some GO categories that are enriched in the set of ABA-regulated genes compared to annotations for the whole genome. These results are schematically shown in Figure 7-5.

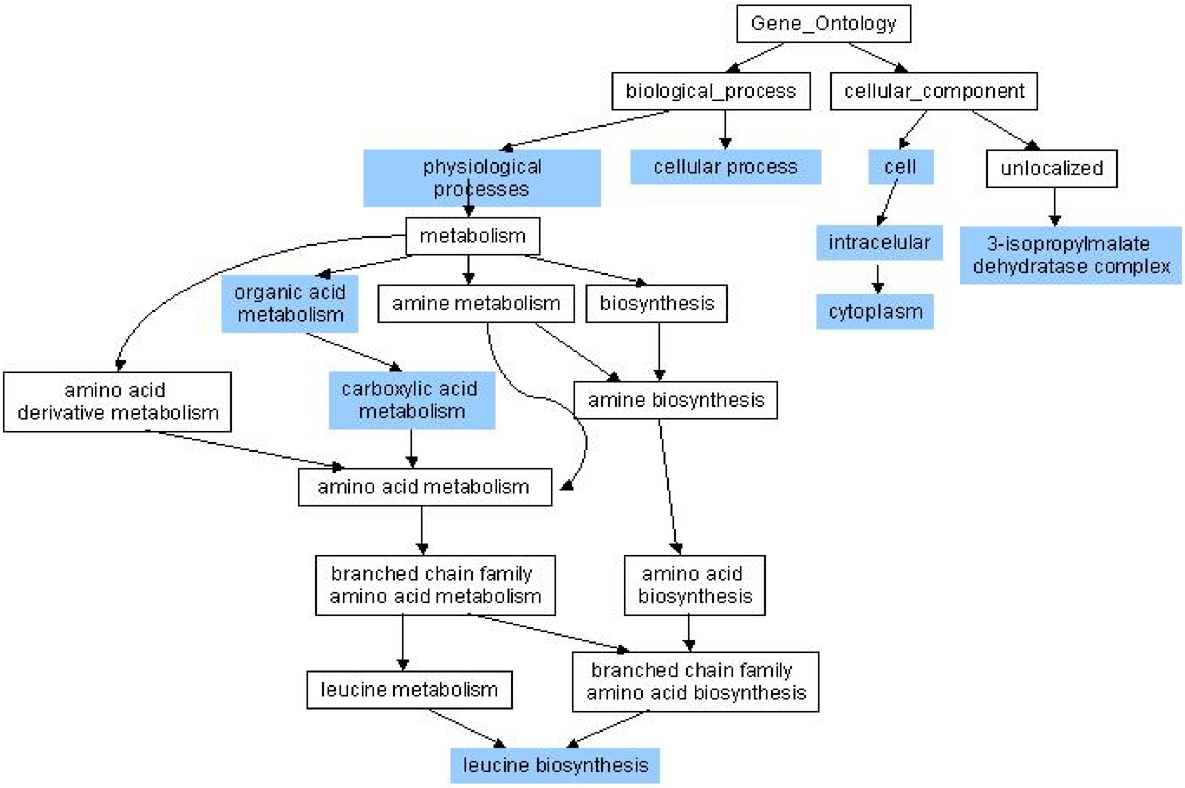


**Figure 7-5: Enrichment of GO categories in the set of ABA-regulated genes compared with GO categories for the whole genome.** Enrichment was tested using the program GOSSIP[10]. Enriched categories are marked as blue boxes

As explained earlier, GO categories are organized hierarchically. The first hierarchical level is divided into three groups: “biological process”, “cellular component” and “molecular function”.

In Figure 7-5 the second line of boxes shows this hierarchical level at the top of the figure. Two from the three categories of this hierarchical level are shown. The following hierarchies descend down to the ninth level at the bottom of the figure. Categories with a FDR below the threshold of 0.05, and hence significantly enriched in the set of ABA-regulated genes are shown as blue boxes.

At the third hierarchical level the GO categories “intracellular” and “3-isopropylmalate dehydratase complex” were significantly enriched in the set of ABA-regulated genes compared to the whole genome (FDR below the $\alpha$ threshold of 0.05).

In general, nine GO categories at different hierarchical levels were found to be significantly enriched in the set of ABA-regulated genes compared with annotations for the whole genome. Considering the role of ABA in osmotic stress responses, the over-representation of the category “intracellular” is not surprising. The physiological changes associated with osmotic stress responses include regulation of membrane-associated enzymes and proteins like dehydrins and LEA proteins[135,138].

### 7.3.3 *Cis*-elements in the upstream regions of ABA-regulated genes

To search for over-represented motifs in upstream regions of ABA-regulated genes, 1 kb upstream sequences were extracted for all genes within a pattern of expression (section 7.3.1), including those genes for which the intergenic region was shorter than 1 kb (333 genes). These sequences were analysed without masking regions of low complexity.

On each set of upstream regions all oligonucleotides of size w=8 were counted on both strands, and the statistical significance of the number of occurrences was computed. Significance was computed using the methodology proposed by van Helden[123]. The computations are based on the binomial distribution. Briefly, the probability to observe exactly the number of occurrences for each oligonucleotide of size w=8 was computed. Then, the probability to observe less or the same number of occurrences was computed. Finally, the probability to observe the same number of occurrences or more is calculated according to Equation (7-1). A significance index (sig) was calculated as the negative logarithm of the computed probability. High values for the significance index correspond to the most exceptional motifs and are considered as over-represented oligonucleotides in the set of sequences evaluated.

$$P(\geq occ)=1-P(\leq obs)+P(obs) \tag{7-1}$$

Oligonucleotides with the highest significance index were compared with the list of *cis*-elements generated (section 4.1.3) to investigate if they have been already described as regulatory sequences in plants.

None of the oligonucleotides with the highest significance index (sig ≥ 6,5) has been previously described as *cis*-regulatory motif. These over-represented oligonucleotides that might be unknown *cis*-elements are an important resource for further research, towards the identification of new *cis*-elements involved in ABA-mediated gene regulation.

From the list of already described *cis*-elements, independent of the pattern of expression (groups 1-9), the motifs MYB, ABRE, -300 element (related to Dof binding sites)[49] and ERELE (Ethylene responsive element)[49] were the most over-represented motifs of size w=8 in 1 kb upstream sequences of ABA-regulated genes.

For every group the following already defined *cis*-elements were significantly over-represented in 1 kb upstream sequences (significance index ≥0,8):

**Group 1** (predominantly down-regulated genes): *Cis*-elements CCA (recognized by MYB related transcription factors)[49], ABRE, LREN (light regulatory element)[121] and MYB binding sites.

**Group 2** (predominantly up-regulated genes): *Cis*-elements CACGCAATGMGH3 (confers auxin inducibility)[49], MYB binding site and SV40.

**Group 3** (early responsive genes): *Cis*-elements ABRE, -300 element, SV40 and TE2F (involved in transcriptional activation in actively dividing cells and tissues)[49].

**Group 4** (late responsive genes): *Cis*-elements MYB binding site, ABRE, -300 element, PIATGAPB (involved in light-activated gene expression) [49] and RY (*cis*-element widely distributed in seed-specific gene promoters)[87].

**Group 5** (transiently down-regulated genes): *Cis*-elements ABRE, CACGCAATGMGH3, OCTAMOTIF2 (observed in plant histone genes)[49], ERELE and SV40.

**Group 6** (transiently down/up regulated genes): *Cis*-elements SV40, MYB binding site, -300 element and CBF1 (responsible for the induction of COR genes by ABA)[49].

**Group 7** (transiently up/down regulated genes): *Cis*-elements –300 element, ABRE and EREL.

**Group 8** (transiently up-regulated genes): *Cis*-elements ERELE, MYB binding site, EREL, AMMORESIIUDCRNIA1 (involved in ammonium-response)[49] and –300 element.

**Group 9** (oscillating patterns): *Cis*-elements PIATGAPB, ABRE, MYB binding site and ERELE.

### 7.3.4 Transcription factors regulated by ABA

It has been experimentally proven that members of the following families of transcription factors are involved in the regulation of ABA-responsive genes: basic domain leucine zipper (bZIP proteins) that bind to the *cis*-element ABRE; basic helix-loop-helix (bHLH proteins) that bind to the *cis*-element MYB binding site; and the ERF/AP2 family of transcription factors, related to the APETALA2 family, that bind to the *cis*-element DRE[2,3,23,45,69,74,105,122,135].

Another family of transcription factors found to be involved in ABA-independent gene expression in response to abiotic stresses is the NAC family. These transcription factors function as transcriptional activators in cooperation with zinc-finger homeodomain proteins, or alone[135].

From the 680 genes regulated by ABA, 40 were classified as transcriptional activators according to the GO annotation. The transcription factors that were found to be regulated by ABA were classified into families, using a classification scheme prepared by Diego Riaño in-house (Riaño-Pachón, *unpublished data*). The results are shown in , it was observed that most transcription factors regulated by ABA in the present study belonged to the families APETALA2/EREBP (6 genes), bHLH family (5 genes) and bZIP transcription factor family (3 genes).

**Table 7-3: Members of different transcription factor families found to be regulated by ABA in the present study.** Classification of *A. thaliana* transcription factors into families made by Diego Riaño (Riaño-Pachón, *unpublished data)*

| Transcription Factor Family | Regulated |
|---|---|
| AP2/EREBP | 6 |
| bHLH | 5 |
| bZIP | 3 |
| Aux/IAA family | 2 |
| C2C2-CO-like | 2 |
| C2C2-Dof | 2 |
| G2-like transcription factor family, GARP | 2 |
| HB (Homeobox transcription factor family) | 2 |
| MYB | 2 |
| NAC domain transcription factor family (NAM) | 2 |
| TCP transcription factor family | 2 |
| Trihelix, Triple-Helix transcription factor family | 2 |
| TUB transcription factor family | 2 |
| Auxin-responsive factor (ARF) | 1 |
| C2C2-GATA | 1 |
| C2H2 | 1 |
| GRAS transcription factor family | 1 |
| JUMONJI family | 1 |
| Pseudo ARR transcription factor family | 1 |
| TOTAL | 40 |

In section 7.3.1 the time course expression of all 680 ABA-regulated genes was presented. ABA-regulated genes were grouped according to 9 different expression patterns. Regarding ABA-responsive transcription factors, none of them showed the expression patterns:

"predominantly up-regulated" (Group 2), "transient down-regulated" (Group 5) or "transient up-regulated" (Group 8).

Transcription factors found to be regulated by ABA belonged to the following gene expression patterns:

| | |
|---|---|
| Group 1: Predominantly down-regulated genes | 1 |
| Group 3: Early responsive genes | 12 |
| Group 4: Late responsive genes | 15 |
| Group 6: Transient down/up regulated genes | 2 |
| Group 7: Transient up/down regulated genes | 9 |
| Group 9: Oscillating patterns | 1 |

It was found that members of the families of transcription factors involved in the regulation of ABA-responsive genes belonged to the groups: "early responsive genes" (Group 3) or "late responsive genes" (Group 4).

Three out six members of the family AP2/EREBP (some members recognize DRE) were found up-regulated either 30 min (2 genes) or 120 min after treatment (1 gene). The transcription factor that was found up-regulated 120 min after treatment was also down regulated 90 min after treatment. The ORF annotation indicated that it is the transcription factor RAV2/AP2 (*At1g68840*). The other three members of the family found to be regulated by ABA were down-regulated 30, 90 or 300 min after treatment.

Three out of five members of the family bHLH were found to be up-regulated 30, 60 or 120 min after treatment. The gene up-regulated 120 min after treatment was down-regulated 90 min after treatment. The other two members of the family found to be regulated by ABA were down-regulated 60 or 90 min after treatment.

All regulated members of the bZIP family were regulated 60 or 90 min after treatment. One of them was up-regulated 60 min after treatment, the other two were down-regulated 90 min after treatment.

## *7.4 Independent confirmations via RT-PCR*

It was of interest to confirm ABA-responsiveness of the following transcription factors:

1. Dof zinc-finger protein DAG2 (Dof affecting germination 2) (*At2g46590*). Promoter analysis showed that *DAG2* is expressed throughout the *A. thaliana* life span, and activity is restricted to the vascular system in all organs[42]. In seeds, increasing concentrations of exogenous ABA prevent germination of wild-type and *dag2* mutants[42]. In this study, expression profiling was made using whole leaves. Independent RT-PCR confirmation intended to probe whether the expression of *DAG2* is affected by increasing concentrations of exogenous ABA.

2. No apical meristem (NAM) family protein (*At5g39610*). For the members of this transcription factor family, only the NAC member *RD26* has been extensively documented as being induced by ABA[35]. The independent confirmation via RT-PCR was intended to probe the regulation by ABA of another member of the family.
3. Two members of the bZIP transcription factor family were selected, i.e. genes *At1g42990* and *OBF4* (*At5g10030*). Members of the bZIP transcription factor family bind to the *cis*-element ABRE, the most relevant *cis*-element in ABA mediated gene regulation[23,32,45,59,135].

For these experiments, wild-type *A. thaliana* cv. C24 plants were grown hydroponically in half-concentrated MS medium supplemented with 2% sucrose. Four-week old plants were stimulated with 100 μM ABA mixed into the fresh medium. Leaves were harvested for the isolation of RNA 30, 60, 90, 120 and 300 min after the addition of ABA. Control plants were stimulated with an equivalent amount of 1 N NaOH, used to dissolve ABA.

Total RNA was digested with DNaseI to avoid the presence of traces of genomic DNA in the samples that would be used for RT-PCR. RNA was reverse transcribed. The generated cDNA was used for the RT-PCR experiments. Expression of the target genes was quantified relative to the expression of a reference gene (*ubiquitin 10* - *At4g05320*). For the analysis, first the expression level of each target gene was compared with respect to the expression of the reference gene by calculation of the $\Delta C_t$, which is defined as $C_{t\text{-target}} - C_{t\text{-reference}}$. $C_t$ is the cycle number were the fluorescence achieved an arbitrary threshold value of 0.2. Secondly, to compare treated and untreated samples, each $\Delta C_t$ value obtained for the treatment ($\Delta C_{t\text{-T}}$) was subtracted from the $\Delta C_t$ values obtained for the corresponding control ($\Delta C_{t\text{-C}}$). Considering an amplification efficiency of 1, relative changes in expression are given by $2^{-\Delta\Delta Ct}$, where the $\Delta\Delta$Ct value is a logarithmic measurement of the relative changes in the expression level upon treatment. A $2^{-\Delta\Delta Ct}$ equal to one indicates no differences in the expression level upon treatment. A $2^{-\Delta\Delta Ct}$ larger than one indicates up-regulation of the gene expression upon treatment, whereas a $2^{-\Delta\Delta Ct}$ smaller than 1 indicates down-regulation of the gene expression upon treatment. In order to assess the robustness of the results, a second reference gene was used (*actin* 2 – *At3g18780*). Considering that no differences in gene expression are expected for this housekeeping gene after ABA-treatment, a $2^{-\Delta\Delta Ct} \approx 1$ was expected. Experimentally, a slight deviation of this trend was observed for the time points 60 and 90 min after treatment ($2^{-\Delta\Delta Ct} = 1.4$). For these specific time points, a $2^{-\Delta\Delta Ct} = 1.4$ was taken as no differences in expression level upon treatment, values larger than 1.4 as up-regulation, and smaller than 1.4 as down-regulation.

Figure 7-6 shows the results observed in the array and in RT-PCR experiments for every target gene. It was observed that all transcription factors assayed are definitively regulated by ABA. However, the expression patterns observed in RT-PCR studies were different from the patterns observed in macroarray experiments. This disagreement between the two experiments was unexpected. However, there are some plausible explanations for the divergences. The plants used for macroarray experiments were grown in soil and ABA was applied by spraying onto the leaves. In contrast, for RT-PCR experiments, the hormone was mixed into the medium, to be absorbed by the roots. These changes were made to provide better controlled experimental conditions.

Nevertheless, it is worth mentioning that also in the literature disagreements between macro and microarray experiments, and RT-PCR or Northern blot experiments have been reported[25,78,85]. Rajeevan *et al.* 2001[85] explains that differences might be due to the fact that in array experiments expression differences of closely related members of gene families may be masked by cross-hybridisation. Holland *et al.* 2002 (cited in Czechowski *et al.* 2004)[25] pointed out that array technologies are qualitative, and that there is not strict linear correlation between signal strength and transcript abundance. Czechowski *et al.* 2004 compared the results obtained with Affymetrix and RT-PCR experiments and found that the agreement observed between both datasets was quite poor[25]. Furthermore, Raheevan *et al.* 2001, Czechowski *et al.* 2004[25,85] and other authors strongly recommend to perform the confirmation experiments with the same pool of RNA, to avoid differences inherent to the quality of the RNA used for either experiments, or the progressive degradation of unstable RNAs[17,25,85].

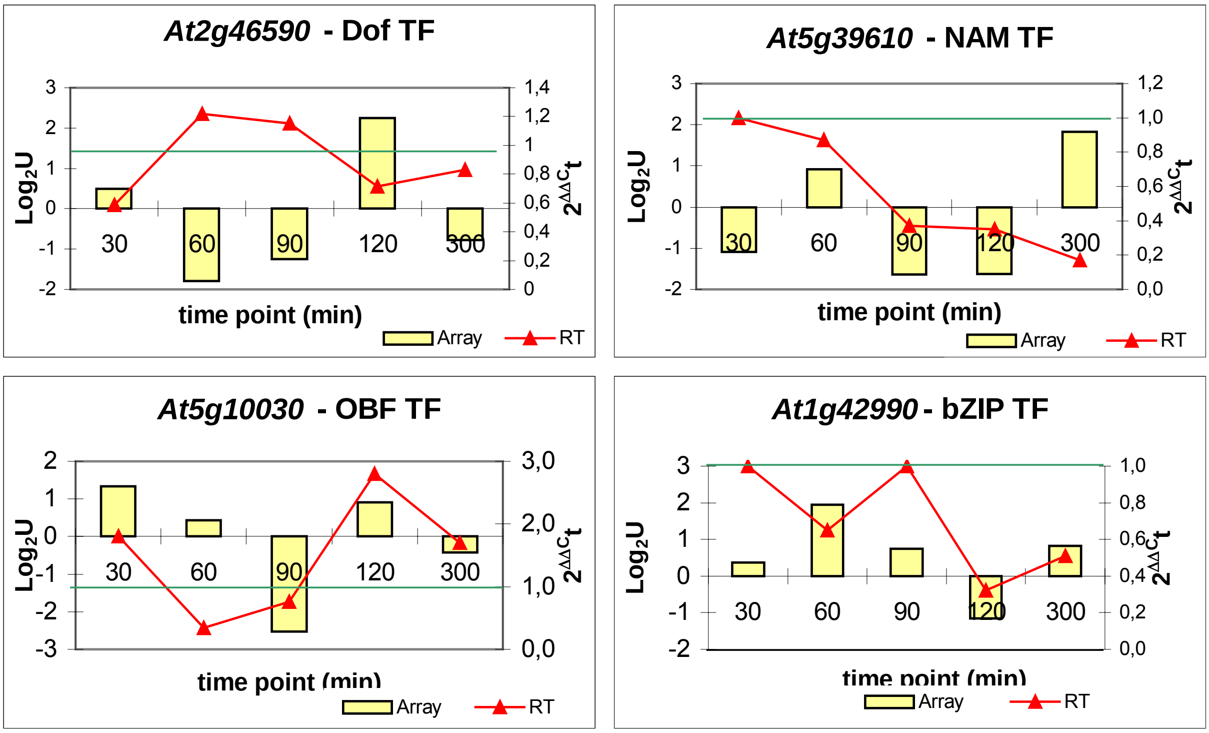


**Figure 7-6: Macroarray and RT-PCR results.** Left y-axis Log$_2$U, right y-axis $2^{-\Delta\Delta Ct}$ values. Bars denote the average expression values observed in macroarray experiments, lines denote transcript abundance relative to ubiquitin observed in RT-PCR. The green line added to each plot show indicates the point where there are no changes in gene expression in RT-experiments. The $2^{-\Delta\Delta Ct}$ values above the green line indicate up-regulation of the target gene, the $2^{-\Delta\Delta Ct}$ values below the green line indicate down-regulation of the target gene

## 7.5 Cross-talk with other hormone signals

The analysis of mutants impaired in the response to a certain hormone had revealed that is not unusual that these plants are also impaired in the response to other hormones. To investigate which of the ABA-regulated genes found in this study have been found to be regulated by other phytohormones, the genes regulated by ethylene, jasmonate, indole acetic acid (IAA) and brassinosteroids treatment (according to expression profiling or northern blots) were compared with ABA-responsive genes found in this study. Results for the other hormone signals were downloaded from www.scri.sari.ac.uk. The number of genes regulated by each treatment is shown in Table 7-4.

**Table 7-4: Genes regulated by different phytohormones.** Results downloaded from DRASTIC (www.scri.sari.ac.uk). Overlap=Genes found to be regulated by ABA and by other phytohormone separately. $\% = \frac{overlap}{regulated} * 100$

| Treatment | Regulated | Up | Down | Overlap | % |
|---|---|---|---|---|---|
| Brassinosteroids | 125 | 94 | 31 | 2 | 1.6 |
| Ethylene | 326 | 208 | 118 | 14 | 4.3 |
| IAA | 61 | 49 | 12 | 5 | 8.2 |
| Methyl jasmonate | 356 | 251 | 105 | 19 | 5.3 |

Considering the number of genes regulated by each stimulus, the overlap between ABA and IAA regulated genes was larger than the overlap between ABA and other hormones, including ethylene, the phytohormone where more interactions has been reported[6,30,37,98,99]. As it would be expected, not always genes that were up- or down-regulated by ABA were similarly up- or down-regulated by the other hormone. In the case of ABA and brassinosteroids, 2 genes were found to be regulated by both hormones separately. One gene was up-regulated by ABA and by brassinosteroids, and one was up-regulated by brassinosteroids and down-regulated by ABA. In the case of ethylene and ABA, 14 genes were found to be regulated by both hormones separately. Seven genes were up-regulated by ethylene, and only one of them was similarly up-regulated by ABA. The other 7 genes were down-regulated by ethylene, and 6 of them were similarly down-regulated by ABA. However, these genes were also up-regulated by ABA 30 and/or 60 min after ABA-treatment. Genes found to be regulated by ABA and ethylene were primordially enzymes. Five genes were regulated by ABA and IAA separately. Only one gene was down-regulated by IAA, the other genes were up-regulated by IAA, and all five genes were up-regulated by ABA. One of them was up/down regulated by ABA. Interestingly, two of the four genes up-regulated by both hormones are annotated as putative protein kinases, the other 3 genes (including the one down-regulated by IAA and up-regulated by ABA) are annotated as peroxidases. In the case of jasmonate and ABA, 19 genes were regulated separately by both hormones. Ten genes were down-regulated by jasmonate, 9 of them also down-regulated by ABA (including 4 up/down or down/up regulated). The other 9 genes were up-regulated by jasmonate. Only 2 of the 9 genes up-regulated by jasmonate were similarly up-regulated by ABA. Some of the genes regulated by ABA and jasmonate were transcription factors of the bZIP and NAC family.

In general, a very small percentage of ABA-regulated genes was found to be regulated by other phytohormones. Remarkably (i) nearly all genes regulated by IAA and ABA were up-regulated by both hormones, and some of them are annotated as putative kinases, that might be involved in signal transduction, and (ii) nearly all genes up- or down-regulated by jasmonate were down-regulated by ABA, including some transcription factors.

## 7.6 Conclusions

1. A group of 680 genes of *A. thaliana* was found to be regulated by ABA in leaves. Most of these genes are newly defined as regulated by ABA. Among them are the transcription factors *OBF-4* (*At5g10030*), *bZIP* (*At1g42990*), and the Dof factor *DAG2* (*At2g46590*). The ABA-responsiveness of these genes was independently confirmed via RT-PCR.

2. Most of the genes found to be regulated by ABA were regulated at a single time point, and complex patterns of expression (genes regulated in at least 3 out of 5 time points) were seldomly found.

3. Only 3 out of 40 transcription factors found to be regulated by ABA have been previously reported[53,97]. The responsiveness to ABA of one of them (NAC family protein – *At5g39610*) common for 3 expression profiling experiments: this study, Seki *et al.* 2002 and Hoth *et al.* 2002[53,97] was independently confirmed via RT-PCR.

4. Only few of the genes found to be regulated by ABA in this study were regulated by other phytohormones such as ethylene, brassinosteroids, jasmonic acid or IAA. The overlap between genes regulated by IAA and ABA showed that all genes regulated by both hormones are up-regulated by ABA, and nearly all up-regulated by IAA. In contrast, nearly all genes regulated by ABA and jasmonic acid are down-regulated by ABA.

# Chapter 8: Biological relevance of computational predictions

*The relevance of the genome-wide computational predictions was validated by comparison with experimental results[78]. The genes found to be regulated by ABA using macroarrays (this study), together with published results on expression profiling in leaves upon ABA treatment[53,97] were taken as the reference group. Genes predicted to be regulated by ABA by computational analysis were compared with genes experimentally determined to be regulated by ABA.*

## *8.1 Reference group –ABA regulated genes found experimentally*

Two articles have been published reporting expression profiling results of wild type *A. thaliana* leaves stimulated with ABA. In this study, the expression profiling of around 10000 cDNAs corresponding to 4757 genes was analysed, representing about 16% of the genome. Seki *et al.* 2002 presented the expression profiling of around 7,000 genes[97], representing about 24% of the genome. Hoth *et al.* 2002 detected 29,475 unique signatures using MPSS[12] and covering ideally 100% of the genome of *A. thaliana*[53]. The resulting number of genes regulated by ABA was different for each approach. In this study 680 genes were found to be regulated by ABA, while Hoth *et al.* 2002 reported 1,400 genes [53], and Seki *et al.* 2002 reported 245 genes. Displaying the results of the three studies in a Venn diagram showed that 3 genes were regulated by ABA in both macro and microarrays, 46 genes were regulated by ABA in both macroarray and MPSS, 38 genes were regulated by ABA in both microarrays and MPSS, and only one gene was regulated by ABA independently of the screening methodology (Figure 8-1). This was the gene *At5g59320*, a “lipid transfer protein 3 (LTP3)”.

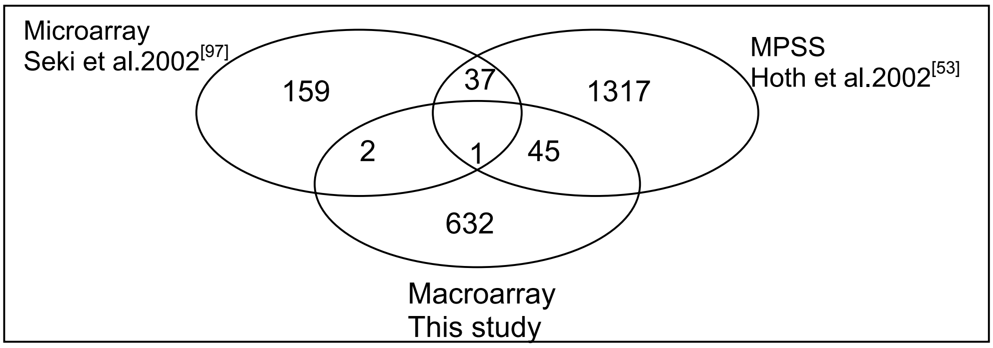


**Figure 8-1: Comparison of genes regulated by ABA in different profiling experiments.** ABA-regulated genes identified by Seki *et al.* 2002[97] using cDNA microarrays, by Hoth *et al.* 2002[53] with MPSS, and in this study with nylon filters

[12] MPSS stands for massively parallel signature sequencing

Reasons for the poor overlap between experiments could be manifold, and might stem from differences in plant culture conditions, treatments, or the expression profiling method used. In this study six-week old plants of *A. thaliana* cv C24 were sprayed with 100 μM ABA, a and the cDNA library used was obtained from different tissues and plants at different developmental stages[76]. Seki *et al.* 2002 used three-week old plants of *A. thaliana* cv Columbia-0, grown hydroponically and stimulated with the same concentration of ABA used in this study. The full-length cDNA library was obtained from plants or seeds subjected to ABA treatment, drought and cold stress, and included different developmental stages[96]. Hoth *et al.* 2002 used four-week old *A. thaliana* plants cv Landsberg, grown hydroponically and stimulated with 50 μM ABA. Harvesting time points were also different for each experiment, in this study the changes in gene expression were monitored at 30, 60, 90, 120 and 300 min after treatment. Seki *et al.* 2002 monitored the changes in gene expression 1, 2, 5, 10 and 24 hours after ABA-treatment. Finally, Hoth *et al.* 2002 pooled the RNA isolated 3 and 5 hours after ABA-treatment. Despite the described differences between experiments, MPSS data and macroarray data showed more common genes than macro and microarrays. None of the three genes found to be regulated by ABA in both macro and microarrays have been described before as being regulated by this phytohormone. One of these genes (*At5g42530*) was annotated as "expressed protein".

Interestingly, Seki *et al.* 2002[97] reported the use of a library constructed specifically from plants kept under stress conditions. In comparison with the results reported here and by Hoth *et al.* 2002[53], less genes were found to be regulated by ABA when such a specialized library was used. In addition to that, the careful analysis of the results presented by Seki *et al.* 2002[97] showed that, at least in the case of genes regulated by ABA, not each cDNA clone corresponded necessarily to a different gene. The degree of redundancy of the library could not be established with the supplementary data provided by the authors. Extrapolating from the results observed in the case of ABA-regulated genes, it seems that the real number of genes represented in the library is actually about one third of the number specified (in total 2,500 genes instead of 7,000).

The set of reference genes was complemented with a list of published ABA-regulated genes[8,9,20,29,31,36,38,40,41,52,55,57,62,63,70,82,92,104,106,109,110,122,125-128,131,133,136]. Finally, the experimentally confirmed reference group was composed of 2,174 genes. The expression of these genes has been reported in different tissues and at different developmental stages, nevertheless most of the expression profiling experiments were made using leaves.

## 8.2 Accuracy of computational predictions

The sensitivity of the computational predictions, defined as the number of true positive genes predicted *in silico*, was assessed first as the number of genes used to construct the consensus sequences and the frequency matrices predicted subsequently computationally. In total 50 TFBSs of *A. thaliana* were used to generate matrices and consensus sequences. Seventeen out of 50 genes were subsequently reported by the pattern-based search (consensus sequences), which is 34% of the genes used, or a sensitivity of the approach of 0.34. In the case of the matrix-based search (CISTER) 8 out of 50 genes were subsequently reported, representing 16% of the genes used, or a sensitivity of 0.16. The number of genes predicted by both computational programs was 1056. Only 4 out of 50 genes were initially used to generate the frequency matrices and the consensus sequences. These were the genes *RD29A*[132], *LEA14*, "alcohol dehydrogenase *ADH*"[29], and the expressed protein *At1g16850*.

It is important to note that in some cases, in the promoters of the genes selected to construct the consensus sequences and the frequency matrices, only one ABA-related *cis* element was documented. These genes will not be detected by any of the computational approaches used, where the combination of ABA-related *cis* elements was evaluated.

A further test of the sensitivity of the computational predictions was carried out. The genes experimentally determined to be regulated by ABA in this study and in other published studies were regarded as positive genes (regulated by ABA). This list of genes was compared with the genes predicted to be regulated by ABA by computational analysis. Sensitivity was defined as the number of true positive genes (TP), divided by the number of true positive and false negative genes (FN) (Equation (8-1)).

$$Sn = \frac{TP}{TP + FN} \tag{8-1}$$

In this case, true positives are predicted by computational analysis and experimentally confirmed as ABA-regulated genes. False negatives are genes that were determined to be regulated by ABA by experimentation, but not predicted by any of the computational approaches used.

When the macroarray results presented in this study were compared with the computational predictions, it was observed that:

1. The number of true positives for predictions made using consensus sequences was 143 genes. The sensitivity was 0.21, meaning that 21% of the genes regulated by ABA were predicted by this method. Comparing the number of genes regulated by ABA and predicted computationally at the different time points, it was found that 120 min after treatment, a higher proportion of genes found to be regulated by ABA were

also predicted correctly. In addition, for the time point 300 min after treatment, most of the genes found to be regulated by ABA were not computationally predicted (false negatives).

2. The number of true positives for predictions made by CISTER was 73 genes. The sensitivity was 0.11. As above, the comparison between the number of true positive and false negative genes at different time points showed that 300 min after treatment more false negatives were observed, while 120 min after treatment more true positives were observed.
3. The number of common genes between both computational methods was 1056. The number of true positives predicted by both methods was 23. The sensitivity in this case was the lowest (0.03). At each time point only few genes found to be regulated by ABA were predicted by computational analysis.

Using the reference list of experimentally confirmed ABA-regulated genes (2174 genes), the sensitivity of the predictions showed that:

1. Using consensus sequences, the number of true positives was 381 genes. The sensitivity was 0.22, indicating that around 78% of the genes regulated by ABA experimentally, were not predicted computationally.
2. Using CISTER, the number of true positives was 268 genes. The sensitivity was 0.12, indicating that the number of genes regulated by ABA but not predicted by computational analysis increases to 88%.
3. Taking the genes predicted by both computational methods, 97 true positive genes were found. In this case the sensitivity was 0.04. Results are graphically shown in a Venn diagram in Figure 8-2.

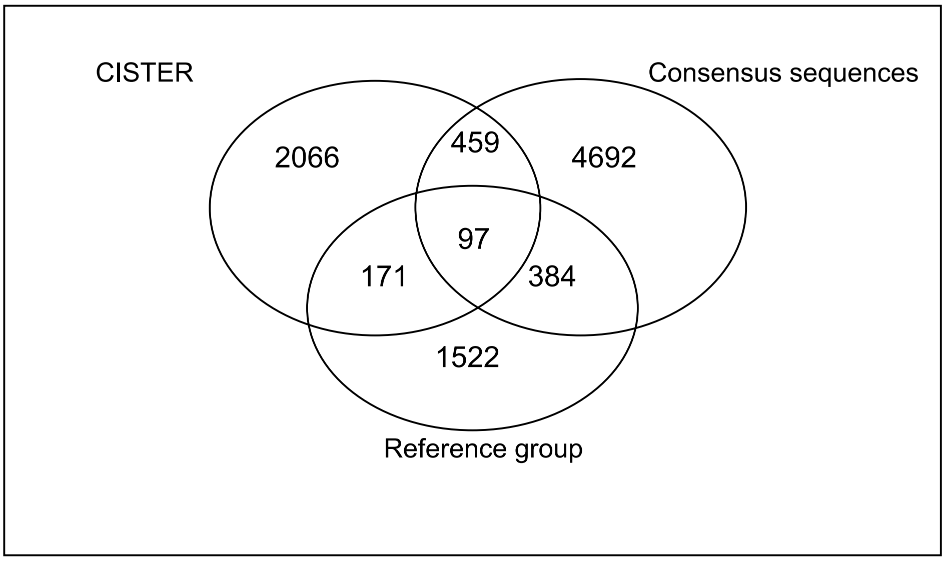


**Figure 8-2: Accuracy of computational predictions.** Comparison of the number of genes predicted by CISTER, Consensus sequences, and genes experimentally confirmed to be regulated by ABA (Reference group)

Overall, the use of consensus sequences delivers more sensitive results than frequency matrices. The overlap between computational methods was very small, and the number of true positive genes even smaller.

The number of experimentally confirmed ABA-regulated genes that were not predicted by computational analysis ranged from 78 to 97 percent. The analysis of the number of true positive and false negative genes at different time points in the macroarray experiment showed that during the late phase of the experiment (300 min after treatment), where only few genes are regulated by ABA, also only few genes were predicted by computational analysis. This result, together with the result obtained 120 min after ABA treatment, where more true positives were found, together with the observation that 40 different transcription factors were regulated by ABA in macroarray experiments, suggest that multiple regulatory networks are at work during the time course of the experiment. Computational predictions identify genes regulated by bZIP transcription factors (carrying ABRE, As1 or CE3 *cis* elements), bHLH transcription factors (carrying MYB binding sites) or AP2 transcription factors (carrying CE1 or DRE *cis* elements). Genes regulated by other transcription factors after ABA treatment will not be predicted. This result reflects the poor understanding of the temporal regulation mechanisms following ABA treatment, and the interplay of different transcription factors and binding sites during the signal transduction process.

Computational predictions were made over the whole genome, whereas the measurements were made over 16% of the genome, represented in the cDNA collection used. Therefore, genes predicted computationally but not represented in the experiment, or genes predicted but not confirmed to be regulated by ABA experimentally were also investigated.

Regarding genes predicted by any computational method, but not covered by the macroarray experiment, the following statements can be made:

1. From 6132 genes predicted using consensus sequences, 5182 were not measured by macroarrays (84%).
2. From 3293 genes predicted using CISTER, 2725 were not measured by macroarrays (83%).
3. From 1056 genes that overlap between both computational methods, 875 were not measured by macroarrays (83%)

The specificity of the computational approaches was measured taking into account the number of true negatives (genes not predicted computationally and not confirmed to be regulated by ABA experimentally). For this calculation, only results observed in the macroarray experiment were considered. Specificity was calculated as:

$$Sp = \frac{TN}{TN + FP} \tag{8-2}$$

Here, TN stands for true negatives, and FP stands for false positives. It was found that:

1. From 4077 genes not regulated by ABA in macroarray experiments, 808 were positively predicted using consensus sequences. In this case the specificity was 0.80.
2. In the case of CISTER, the number of false positives was 495; the specificity was 0.88.
3. In the case of the overlap between computational methods, the number of false positives was 158; the specificity was 0.96.

Overall results showed that the number of false positives (genes predicted computationally but not confirmed to be regulated by ABA experimentally) was low, indicating a high specificity of the computational approaches (between 0.96 and 0.80). However, the number of false negative genes (not predicted by computational analysis, but confirmed to be regulated by ABA experimentally) was also very high, indicating a low sensitivity.

The high specificity of the computational approaches is reflected in the few genes predicted to be putative ABA-regulated genes. Between 10 and 20 percent of the genes of *A. thaliana* were predicted to be regulated by ABA in this study, whereas other approaches to localize transcription factor binding sites identified between 1 and 11 MYB-binding sites, 2 and 3 G-boxes, and 6 to 7 AP2 binding sites per gene in the *A. thaliana* genome[107].

Despite the high specificity, the low selectivity of the approaches (reflected in the high number of false negatives) makes it impossible to determine whether every gene predicted to be ABA-regulated would be experimentally confirmed. Unless a better model of interaction between the different *cis* elements and transcription factors involved in the ABA-induced transcriptional network became available, the selectivity of the computational predictions of ABA-regulated genes will not increase.

The analysis of the true positive genes (predicted computationally and experimentally confirmed) revealed that the genes *At1g21760*, *At1g77000*, *At1g51550*, *At2g40920*, *At3g06380* and *At3g61060*, that encode proteins that contain the F-box domain—involved in the re-direction of proteins to the ubiquitin pathway—[1] were experimentally confirmed to be regulated by ABA, and they have putative TFBSs for transcription factors activated upon ABA treatment. This result confirmed the relevance of the regulation of components of the proteosome in ABA signalling[30].

Other genes classified as true positives included the DRE-binding protein DREB2A (*At5g05410*) and the putative DRE-binding protein *At4g16750*. The presence of ABA-related TFBSs indicates that the protein DREB2A, known to bind to DRE-binding sites[75], is directly regulated by ABA.

Different transcription factors were found to be regulated by ABA and predicted by computation, including members of the NAC family (Non Apical Meristem), e.g. *At1g01720*, *At1g32870*, *At1g77450*, *At2g22290*, *At4g27410*, *At5g39610*, *At5g52880* and *At5g61430*. This result is a clear indication of the relevance of the involvement of the NAC transcription factor family in the ABA transcriptional network. Further work, intended to localize putative target genes for these transcription factors and to clarify the possible relationship between *cis*-elements, will certainly improve the understanding of the ABA-mediated gene regulatory network.

A large number of transcription factors belonging to the bHLH and MYB families were also found to be regulated by ABA and predicted computationally, as well as putative ABRE-binding transcription factors.

Different genes involved in the auxin signalling pathway were also found to be regulated by ABA and predicted computationally to be putative targets of transcription factors activated during ABA-signalling. These included the genes *At1g17350*, *At1g48690*, *At2g28350*, *At3g03850*, *At3g23030*, *At4g17280*, *At4g48690* and *At5g25890*. Furthermore, the results presented in section 7.5 showed a high level of phytohormonal cross-talk between auxin and ABA. The putative regulation of components of the auxin signalling pathway by ABA-activated transcription factors revealed an emerging network interconnection between components of the ABA transcriptional machinery that has not been studied in any detail yet.

## *8.3 Conclusions*

1. The comparison between different expression profiling experiments, conducted using different *A. thaliana* accessions (Col-0, C24 and Landsberg) and using different methodologies and growing conditions, showed only one gene in common. Additionally, only few genes were comparably regulated by ABA when pairs of experiments were examined. The large differences between experiments might stem from differences in plant culture conditions, treatments, or the expression profiling method used, but might also be connected in some extend with sequence polymorphisms between accessions of *A. thaliana*, a common observation in coding and non-coding genes.

2. Computational predictions of ABA-related *cis* elements showed a high specificity, reflected in the fact that only few genes were predicted to be regulated by ABA. Only between 4-12 percent of the genes predicted by computational analysis corresponded to false positives. Despite the high specificity, the low selectivity of the approaches (i.e. a large number of false negatives) clearly indicates that only a small proportion of the ABA-regulated genes might be directly regulated by the *cis* elements found in ABA-responsive genes. Other genes experimentally determined to be regulated by ABA might be activated by other transcription factors that, in turn, could directly or indirectly be regulated by ABA.

3. The analysis of the ORF and GO annotations of the genes experimentally confirmed to be regulated by ABA revealed that different members of the NAC family of transcription factors are regulated by ABA. Furthermore, computational predictions revealed that the ABA-mediated transcriptional regulation of these genes might be achieved through the ABA-related *cis*-elements found in their upstream sequences. The activation/repression of these transcription factors could play a crucial role in the transduction of the ABA-signal.

4. Genes belonging to the auxin signalling pathway experimentally confirmed to be regulated by ABA and which are at the same time putative targets of transcription factors involved in the ABA signalling pathway confirmed the close interaction between both signalling pathways, at least at the transcriptional level.

## Chapter 9: Discussion

It is known that genes in genomes of eukaryotes are regulated by means of multiple regulatory proteins (transcription factors), acting through specific regulatory sequences (TFBSs). Normally, the complex networks formed by these interactions are dissected by laborious perturbation analyses[95]. Having complete genome sequence data available on one side, and techniques capable of monitoring simultaneously the expression of hundreds or thousand genes on the other side, the challenge is to understand regulatory mechanisms of all and every single gene in the genome.

Computational tools for the analysis of gene regulation are designed to speed up the process of understanding gene regulatory mechanisms. However, the reliable prediction of TFBSs and in particular the prediction of individual binding sites has proven to be a very difficult task. In this study, the simultaneous identification of different ABA-related *cis*-elements was used to increase predictive reliability.

The idea of using *cis*-elements to identify genes that respond to certain stimuli is not new. In *Drosophila melanogaster*, Dorsal recognition sites were used to identify genome-wide clusters of binding sites. The accuracy of predictions was around 34% (5 positives out of 15 genes predicted)[67]. In *A. thaliana* Zhang *et al.* 2005 used expression-profiling data to generate ABRE and Coupling Element position weight matrices to identify ABA and abiotic stress-responsive genes. Considering only the top scoring predictions, the accuracy was about 67%[137]. It has been found that the rate of success strongly depends on the correctness of the modelled interactions between *cis*-elements.

Experimentally identified *cis*-elements involved in ABA-regulated gene expression were used to find genes putatively regulated by ABA over the whole genome. The outcome of this analysis indicates that 10 to 20 percent of the genes in the *A. thaliana* genome might be regulated by ABA.

Screening of upstream sequences with consensus sequences and matrices suggested that some regulatory sequences identified in monocots might not play any regulatory role in ABA-mediated gene regulation in *A. thaliana*. This finding complicates the eventual identification of ABA-regulated genes by means of phylogenetic footprinting on ortologous genes of these two plant species. Phylogenetic footprinting has shown to be relatively successful to reduce the number of falsely predicted sites in drosophila[90], yeast[22]and humans and mouse[86].

The two regulatory sequences that were found to play a dubious role in the regulation of ABA-responsive genes in *A. thaliana* were the coupling elements CE1 and CE3. None of these elements has been identified yet in upstream sequences of ABA-regulated genes in *A. thaliana*. In this study, sequences corresponding to CE3 were seldomly found in *A. thaliana* upstream sequences. Additionally, CE3 was not found forming pairs with other ABA-related *cis*-elements (using consensus sequences to identify such occurrences). The identification of putative sequences of CE3 using frequency matrices showed the largest amount of putative occurrences (Table 6-10). A closer look at the identified sequences showed that they have little similarity with the consensus sequence of CE3, hence they had low scores.

CE3 is an ABA-related *cis*-element reported in monocots. Hobo *et al.* 1999[50] proposed to consider these binding sites as non-ACGT binding sites. In rice, the binding site is recognized by a transcription factor belonging to the bZIP class of transcription factors, and the sequence is relatively similar to ABRE (ABRE: ACGTGGC, CE3: GCGTGTC[103]). This sequence was under-represented in 1 kb upstream sequences (-0.46). Considering that so far none of the ABA-responsive genes in *A. thaliana* has been reported to have active CE3 sequences and taking into account its absence in 1 kb upstream sequences, it is proposed here that CE3 does not possess any regulatory function in *A. thaliana,* unlike in monocots.

Sequences corresponding to the coupling element CE1 were found very often in *A. thaliana* upstream sequences. Additionally, the element was found to form pairs with other ABA-related *cis*-elements. However, the *cis*-elements of a pair were separated by long distances. At least in the case of the pairs between ABRE and CE1 the distance between *cis*-elements was normally larger than 100 bp. In monocots, and particularly in barley, it was shown that the interaction between ABRE and CE1 strongly depends on the spacer distance and the orientation of the elements[101]. It should be noted that ABRE-binding proteins are bZIPs, capable of forming homo- and heterodimers to bind to their target sequence and thereby induce transcription[16,101,135]. The ABRE-CE1 complex must be in a particular orientation and distance to confer ABA-induction, and the distance is always relatively short (maximum 30 bp) suggesting a direct interaction between the transcription factors[103]. It is very unlikely that large distances between ABRE and CE1 binding sites may lead to a direct interaction between transcription factors, unless in *A. thaliana* unlike in monocots, the interaction between ABRE and CE1 is mediated by other protein(s) linking the transcription factors. Given the large distance found between ABRE and CE1, a direct interaction between transcription factors can be discarded.

The results of the computational analyses led to the conclusion that in *A. thaliana* the coupling element of ABRE is ABRE itself. This conclusion is based on the following findings: (i) pairs of ABRE *cis*-elements were over-represented in upstream sequences compared with random sequences; (ii) the frequency of the pairs was larger than the frequency of the *cis*-element alone; and (iii) the distance between *cis*-elements was in more than 50% of the cases shorter than 50 bp. According to these results, it is plausible that homodimers instead of heterodimer complexes bind to ABRE *cis*-elements to confer ABA responsiveness in *A. thaliana*.

DRE has been described as a coupling element of ABRE only in *A. thaliana*[132]. Here, it was found that very small and large distances between *cis*-elements were consistently avoided in upstream sequences. In addition, computational analysis revealed that the gene *DREB2A*, known to bind DRE binding sites might be target of transcription factors involved in the ABA-signalling pathway.

MYB binding sites were significantly over-represented in *A. thaliana* upstream sequences, and the distance between elements lead to the conclusion that the transcription factors that recognize both kinds of sequences do not interact directly but might very likely be part of a larger regulatory complex. Different transcription factors belonging to the bHLH transcription factor family (which binds MYB binding sites) and to the MYB family were found to be regulated by ABA, and to be targets of ABA-regulated transcription factors (the latter according to computational results).

According to the results observed in the computational screenings, the following interactions between ABA-related *cis*-elements are proposed:

1. Homodimers of ABRE-binding proteins recognize ABRE binding sites in ABA-responsive genes. Observations in mammals pointed out that proteins similar to ABRE-binding proteins (CREB) are constitutively bind to their target promoters and being activated by phosphorylation. Phosphorylation does not change the DNA binding properties, but stimulates the interaction with other proteins, including the transcriptional machinery[16]. A similar mechanisms might be found in *A. thaliana*, and it is proposed that the phosphorylation of ABRE-binding proteins might be achieved by kinases of the ABA-signalling pathway.
2. DRE and MYB binding proteins transcriptionally regulated by components of the ABA signalling pathway bind to their target promoters after transcriptional induction by ABA-regulated transcription factors. In the case of DRE, DRE-binding proteins have been shown to be induced independently of ABA after osmotic stress. ABA-mediated regulation of genes carrying DRE and MYB binding sites requires protein synthesis and perhaps also the accumulation of transcriptional activators and other signalling components such as kinases and phosphatases. Once MYB or DRE-binding proteins have been synthesized, it is very likely that MYB or DRE binding proteins form higher-order protein complexes together with ABRE-binding proteins. The distance found between *cis*-elements might give some indications of the size of the proteins complexes form.

Some experimental evidence from this study and from previous studies suggest that ABRE might interact with other ABA-related *cis*-elements in high-order protein complexes, that induction of ABA-regulated genes mediated by MYB and DRE requires protein synthesis, and that ABRE-binding proteins might be constitutively bound to their target sequences. This is the evidence mentioned above:

1. The activation of the ABA-induced gene *RD29A* from *A. thaliana* requires ABRE and DRE *cis*-elements. Activation mediated by ABA is a slow activation, suggesting that protein synthesis is involved[75]. Both ABRE-binding proteins and DRE-binding proteins are synthesised after ABA accumulation[75]. In this study, it was found that the expression of the gene *DREB2A* is up-regulated after ABA-treatment
2. The expression of the gene *RD22* from *A. thaliana* is mediated by the interaction of MYB and MYC binding sites, where the recognized MYC-binding site is identical to the ABRE-binding site. The induction of this gene, and other genes with MYB binding sites in their promoter sequences requires protein synthesis[2,3].
3. In this study, different members of the bHLH and MYB family of transcription factors were found to be regulated by ABA and are putative targets of ABA-regulated transcription factors. Once these transcription factors are induced and transcribed, they can bind to their target promoter sequences, and activate the expression of other ABA-regulated genes.
4. Very few members of the bZIP family of transcription factors (that bind to ABRE) were induced after ABA treatment and/or were computationally predicted to be targets of ABA-activated transcription factors, suggesting that ABRE-binding proteins are bound to their target promoters, and are activated by components of the ABA signalling pathway such as kinases, without the involvement of protein synthesis.

## 9.1 Expression profiling of leaves upon ABA stimulus

The action of ABA can be grouped into two categories: (i) avoidance mechanisms, which are activated very early, trying to minimize the exposure of the plant to stress, and (ii) tolerance mechanisms, which allow the plant to withstand the stress. During the second phase the plant accommodates to the new environmental conditions[30].

The avoidance mechanisms result in changes at the cytoplasmic level and includes an increased expression of specific genes to overcome the new adverse situation. In the case of osmotic stress or increasing concentrations of ABA, among the genes induced are those that encode chaperones, LEA proteins, enzymes for osmolyte biosynthesis, detoxification enzymes, and gene products involved in the transduction of the signal such as protein kinases, phosphatases, transcription factors and enzymes in phospholipid metabolism[30,135,138]. In parallel, the expression of some genes that need to be inhibited or which are not necessary in the ABA response are repressed.

In this study it was observed that at the beginning of the experiment (30 and 60 min after treatment) most of the genes were up-regulated, and only a small percentage was repressed. The large number of up-regulated genes might be related to the "avoidance mechanism" described.

After 90 min of treatment, most of the regulated genes were down-regulated, including some previously up-regulated genes. Five hours after treatment, only slight changes in gene expression were detected, and only about 15% of all genes found to be regulated by ABA were regulated at that specific time point. It was considered that early responsive genes where those regulated 30 and 60 min after treatment, whereas late responsive genes were those regulated after 90 min of treatment. It might be possible that the transition from avoidance to tolerance mechanisms starts 90 min after treatment.

From the 680 ABA-regulated genes, about one third was annotated as expressed proteins (177). Their potential function, according to the GO annotation[7], could be in "cell communication", "ion transporter activity", "kinase activity", "metabolism", "nucleic acid binding", "hydrolase activity", "secretion" or "protein transport". Further detailed characterisation of these genes is needed.

The analysis of the upstream sequences of genes found to be regulated by ABA showed that ABRE and MYB binding site were among the most highly over-represented *cis*-elements. Both *cis*-elements have been experimentally confirmed as being important in the regulation of gene expression mediated by ABA in *A. thaliana*[2,3,16,75,77,78,101,102,105,118,135,137].

Other over-represented *cis*-elements in upstream sequences were some light-dependent regulatory elements, like PIATGAPB and LREN. This suggests that genes putatively regulated by light (involved in the process of photosynthesis[121]) might also be regulated by ABA. Some of these genes were down-regulated (Group 1), and some up- or down-regulated in the late phase of the experiment (Group 4). This result is in good agreement with the results presented by Wu *et al.* 2001[130], who showed that although photosynthesis is inhibited under stress, proteins involved in the process might be up- or down-regulated in response to stress.

The over-representation of ethylene-responsive elements and *cis*-elements conferring auxin inducibility was not surprising, and underlined the importance of hormone cross-talk. *Cis*-elements conferring responsiveness to ethylene and auxin were found to be over-represented in genes grouped as "transiently down-regulated" (Group 5), "transiently up-regulated" (Group 8) and genes that showed oscillating patterns (Group 9).

The interaction between auxin, ethylene and ABA has been already described, mainly the interaction between ethylene and ABA. It is already evident that interactions among signalling pathways for different hormones may occur through kinases (e.g.MAPKs) or other signalling components[30]. Some signalling components could have different targets in different signalling pathways, so that certain components could act as nodes for information transfer between various pathways, functioning as integrators of different signals[37,98]. However, little is known about transcriptional regulation of genes that act in different signalling pathways. It is becoming evident that genes that act in different pathways might have binding sites for transcription factors activated by different hormones in their promoter sequences, to allow the integration of the signals at multiple levels. If the interaction between different *cis*-elements involved in apparently separate hormone responses could be established, and the current knowledge about the relations between transcription factors and binding sites improves, it would be possible in the future to distinguish direct from indirect interactions of signalling pathways in hormone responses.

With the present data the exact mechanism of regulation could hardly be established, but it can be speculated that:

1. The bZIP transcription factor up-regulated after 60 min of treatment (*At1g42990*) could lead to the differential expression of some late responsive genes, including the down-regulation of the other two bZIP transcription factors (*At4g36730* and *At5g10030*) 90 min after treatment. The over-representation of ABRE *cis*-elements[13] in the upstream sequences of genes belonging to groups 4 and 5 supports this hypothesis.
3. The expression pattern found for the Dof transcription factors *At2g46590* (*DAG2*) and *At3g61850* (*DAG1*), and the over-representation of Dof-binding sites in genes of the groups 6 and 7 suggest that one factor may down-regulate the other. When one transcription factor was up-regulated, the other was down-regulated and *vice versa*. Likewise, the Dof *At3g61850* up-regulated 60 min after treatment might be responsible for the up-regulation of the genes in group 7, 90 min after treatment. On the one hand, plants that constitutively over-express or had little or no expression (due to RNA interference) showed no altered expression of the other transcription factor in leaves (Dr. Maria Ines Zanor, personal communication). However, on the other hand, it was proposed that during seed germination the expression of both genes in maternal tissue plays opposite regulatory roles, since mutations on these genes caused opposite phenotypes[42,79].

## *9.2 Comparative analysis of experimental and computational data*

The comparison of *in silico* predictions with experimental results showed a small overlap of genes predicted by either computational method, and regulated by ABA. It was not expected

[13] ABRE elements are recognized by bZIP transcription factors

that *in silico* predictions and experimental data would coincide perfectly. However, it should be mentioned that expression profiling experiments showed also a very small overlap. Only one gene was found to be regulated by ABA in three expression profiling experiments. Other genes showed contradictory results, and others were found to be regulated in only one experiment.

It was found that the selectivity of the computational methods used was very high (between 0.8 and 0.9). Only a fraction of genes of *A. thaliana* were predicted to be putative regulated by ABA. However, the sensitivity of the approach (identification of genes experimentally regulated by ABA) was very low. The problems to identify more genes regulated by ABA seems to be directly related to the fact that ABA signalling involves the activation of a large and diverse amount of transcription factors that recognize different kind of transcription factors binding sites. In this study, only a fraction of genes regulated by ABA was detected *in silico*, since only few transcription binding sites were represented by either consensus sequences or frequency matrices.

So far the identification of ABA-regulated genes has been focused on the identification of two *cis*-elements in promoter regions, ABRE and CE1. The variety of *cis*-elements over-represented in the upstream region of ABA-regulated genes, together with the wide range of transcription factors up-regulated as result of the stimulation with ABA, showed that the network of interactions is far more complex. The lack of information about recognition sequences of *A. thaliana* transcription factors, together with the inherent difficulties to precisely define their target genes, allows hypothesizing that there are many functionally unconfirmed connections between *cis*-elements.

According to the results of the overlap between computational predictions and experimental results, it is proposed that genes putatively regulated by ABA might also contain in their promoter regions binding sites for the NAC family of transcription factors, apart from the sites considered here.

Another interaction between *cis*-elements that should be studied in more detail in the future is the possible interaction between ABA-related and auxin-related *cis*-elements, since a large overlap between both signalling pathways is becoming evident.

Based on the observed results, it becomes evident that the regulation of ABA-induced/repressed genes in *A. thaliana* is mediated by some of the *cis*-elements considered here. However, other *cis*-elements included in this study, such as the *cis*-elements As1 and CE3 might affect the results critically reducing the specificity. In addition, some *cis*-elements not included here might improve the detection of ABA-related *cis*-elements (e.g. NAC-binding sites or auxin-binding sites).

## Chapter 10: Conclusions

The genome-wide screening for the identification of *cis*-regulatory elements that confer ABA-responsiveness revealed that between 10 and 20 percent of the annotated genes in *A. thaliana* might be regulated by ABA.

It was possible to identify some pairs of ABA-related *cis*-elements significantly over-represented in *A. thaliana* upstream sequences. The *in silico* analysis revealed that in the case of ABA-mediated gene regulation in *A. thaliana*, some interactions between *cis*-elements observed in monocotyledonous species might not be active.

It was predicted computationally that the most prominent *cis*-elements found in ABA-responsive genes in *A. thaliana* are ABRE, DRE and MYB binding sites. Plausible interactions between these *cis*-elements might involve the formation of homodimers between ABRE binding proteins. Between ABRE and DRE, and ABRE and MYB might involve the formation of protein complexes. It is speculated in this study that the distance between interacting *cis*-elements give some hints about the size of the protein complexes involved.

The combination of *in silico* and experimental approaches revealed that with the present knowledge about the putative interactions between *cis*-elements, only a small fraction of the interconnections between *cis*- and trans- acting elements are recovered. Some guidelines about putative new interactions between *cis*-elements might be interactions between ABA- and auxin-responsive elements. As well as between ABA-related *cis*-elements and binding sites of the NAC family of transcription factors.

In future work, the *in silico* analysis of the rice genome will provide some insights about differences between the most important ABA-related *cis*-elements if any.

# Chapter 11: References

[1] Analysis of the genome sequence of the flowering plant Arabidopsis thaliana *Nature* 2000; 408(6814):796-815.

[2] Abe H, Urao T, Ito T *et al.* Arabidopsis AtMYC2 (bHLH) and AtMYB2 (MYB) function as transcriptional activators in abscisic acid signaling. *Plant Cell* 2003; 15(1):63-78.

[3] Abe H, Yamaguchi-Shinozaki K, Urao T *et al.* Role of arabidopsis MYC and MYB homologs in drought- and abscisic acid-regulated gene expression. *Plant Cell* 1997; 9(10):1859-68.

[4] Aerts S, Van Loo P, Thijs G, Moreau Y, De Moor B. Computational detection of cis -regulatory modules. *Bioinformatics* 2003; 19 Suppl 2:II5-II14.

[5] Alberts B, Bray D, Lewis J *et al.* Molecular Biology of the cell. 3rd Edition. New York: Garland Publishing, 1994.

[6] Anderson JP, Badruzsaufari E, Schenk PM *et al.* Antagonistic interaction between abscisic acid and jasmonate-ethylene signaling pathways modulates defense gene expression and disease resistance in Arabidopsis. *Plant Cell* 2004; 16(12):3460-79.

[7] Ashburner M, Ball CA, Blake JA *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 2000; 25(1):25-9.

[8] Baker SS, Wilhelm KS, Thomashow MF. The 5'-region of Arabidopsis thaliana cor15a has *cis*-acting elements that confer cold-, drought- and ABA-regulated gene expression. *Plant Mol Biol* 1994; 24(5):701-13.

[9] Beaudoin N, Serizet C, Gosti F, Giraudat J. Interactions between abscisic acid and ethylene signaling cascades. *Plant Cell* 2000; 12(7):1103-15.

[10] Bluthgen N, Kielbasa SM, Herzel H. Inferring combinatorial regulation of transcription in silico. *Nucleic Acids Res* 2005; 33(1):272-9.

[11] Bonnet Eric, Van de Peer Yves. zt: A Sofware Tool for Simple and Partial Mantel Tests. *Journal of Statistical Software* 2002; 7(10).

[12] Brocard IM, Lynch TJ, Finkelstein RR. Regulation and role of the Arabidopsis abscisic acid-insensitive 5 gene in abscisic acid, sugar, and stress response. *Plant Physiol* 2002; 129(4):1533-43.

[13] Burge Chris CAMKS. Over- and under-representation of short oligonucleotides in DNA sequences. *Proc Natl Acad Sci USA* 1992; 89:1358-62.

[14] Busk PK, Jensen AB, Pages M. Regulatory elements in vivo in the promoter of the abscisic acid responsive gene rab17 from maize. *Plant J* 1997; 11(6):1285-95.

[15] Busk PK, Pages M. Protein binding to the abscisic acid-responsive element is independent of VIVIPAROUS1 in vivo. *Plant Cell* 1997; 9(12):2261-70.

[16] Busk PK, Pages M. Regulation of abscisic acid-induced transcription. *Plant Mol Biol* 1998; 37(3):425-35.

[17] Bustin SA. Quantification of mRNA using real-time reverse transcription PCR (RT-PCR): trends and problems. *J Mol Endocrinol* 2002; 29(1):23-39.

[18] Calzone FJ, Theze N, Thiebaud P *et al.* Developmental appearance of factors that bind specifically to *cis*-regulatory sequences of a gene expressed in the sea urchin embryo. *Genes Dev* 1988; 2(9):1074-88.

[19] Cao X, Aufsatz W, Zilberman D *et al.* Role of the DRM and CMT3 methyltransferases in RNA-directed DNA methylation. *Curr Biol* 2003; 13(24):2212-17.

[20] Chen CN, Chu CC, Zentella R, Pan SM, Ho TH. AtHVA22 gene family in Arabidopsis: phylogenetic relationship, ABA and stress regulation, and tissue-specific expression. *Plant Mol Biol* 2002; 49(6):633-44.

[21] Chessel D, Dufour Anne B, Thiolousse Jean. The ade4 package -I:One-table methods. *Rnews* 2004; (4): 2-3..

[22] Chiang DY, Moses AM, Kellis M, Lander ES, Eisen MB. Phylogenetically and spatially conserved word pairs associated with gene-expression changes in yeasts. *Genome Biol* 2003; 4(7):R43.

[23] Choi H, Hong J, Ha J, Kang J, Kim SY. ABFs, a family of ABA-responsive element binding factors. *J Biol Chem* 2000; 275(3):1723-30.

[24] Cooper GM. The cell. A molecular approach. 2nd Edition. 2000.

[25] Czechowski T, Bari RP, Stitt M, Scheible WR, Udvardi MK. Real-time RT-PCR profiling of over 1400 Arabidopsis transcription factors: unprecedented sensitivity reveals novel root- and shoot-specific genes. *Plant J* 2004; 38(2):366-79.

[26] Davuluri RV, Sun H, Palaniswamy SK *et al.* AGRIS: Arabidopsis gene regulatory information server, an information resource of Arabidopsis *cis*-regulatory elements and transcription factors. *BMC Bioinformatics* 2003; 4(1):25.

[27] Denekamp M, Smeekens SC. Integration of wounding and osmotic stress signals determines the expression of the AtMYB102 transcription factor gene. *Plant Physiol* 2003; 132(3):1415-23.

[28] Dervan PB. Molecular recognition of DNA by small molecules. *Bioorg Med Chem* 2001; 9(9):2215-35.

[29] Dolferus R, Jacobs M, Peacock WJ, Dennis ES. Differential interactions of promoter elements in stress responses of the Arabidopsis Adh gene. *Plant Physiol* 1994; 105(4):1075-87.

[30] Fedoroff NV. Cross-talk in abscisic acid signaling. *Sci STKE* 2002; 2002(140):RE10.

[31] Forment J, Naranjo MA, Roldan M, Serrano R, Vicente O. Expression of Arabidopsis SR-like splicing proteins confers salt tolerance to yeast and transgenic plants. *Plant J* 2002; 30(5):511-9.

[32] Foster R, Izawa T, Chua NH. Plant bZIP proteins gather at ACGT elements. *FASEB J* 1994; 8(2):192-200.

[33] Frech K, Quandt K, Werner T. Finding protein-binding sites in DNA sequences: the next generation. *Trends Biochem Sci* 1997; 22(3):103-4.

[34] Frith MC, Hansen U, Weng Z. Detection of *cis*-element clusters in higher eukaryotic DNA. *Bioinformatics* 2001; 17(10):878-89.

[35] Fujita M, Fujita Y, Maruyama K *et al.* A dehydration-induced NAC protein, RD26, is involved in a novel ABA-dependent stress-signaling pathway. *Plant J* 2004; 39(6):863-76.

[36] Gaubier P, Raynal M, Hull G *et al.* Two different Em-like genes are expressed in Arabidopsis thaliana seeds during maturation. *Mol Gen Genet* 1993; 238(3):409-18.

[37] Gazzarrini S, McCourt P. Cross-talk in plant hormone signalling: what Arabidopsis mutants are telling us. *Ann Bot (Lond)* 2003; 91(6):605-12.

[38] Gilmour SJ, Artus NN, Thomashow MF. cDNA sequence analysis and expression of two cold-regulated genes of Arabidopsis thaliana. *Plant Mol Biol* 1992; 18(1):13-21.

[39] Goda H, Sawa S, Asami T *et al.* Comprehensive comparison of auxin-regulated and brassinosteroid-regulated genes in Arabidopsis. *Plant Physiol* 2004; 134(4):1555-73.

[40] Gong D, Gong Z, Guo Y, Zhu JK. Expression, activation, and biochemical properties of a novel Arabidopsis protein kinase. *Plant Physiol* 2002; 129(1):225-34.

[41] Gosti F, Bertauche N, Vartanian N, Giraudat J. Abscisic acid-dependent and -independent regulation of gene expression by progressive drought in Arabidopsis thaliana. *Mol Gen Genet* 1995; 246(1):10-8.

[42] Gualberti G, Papi M, Bellucci L *et al.* Mutations in the Dof zinc finger genes DAG2 and DAG1 influence with opposite effects the germination of Arabidopsis seeds. *Plant Cell* 2002; 14(6):1253-63.

[43] Guasconi V, Hakima Y, Ait-si-ali S. Transcription factors. Atlas of Genetics and Cytogenetics in Oncology and Haematology, 2002.

[44] GuhaThakurta D, Stormo GD. Identifying target sites for cooperatively binding factors. *Bioinformatics* 2001; 17(7):608-21.

[45] Guiltinan MJ, Marcotte WR, Jr., Quatrano RS. A plant leucine zipper protein that recognizes an abscisic acid response element. *Science* 1990; 250(4978):267-71.

[46] Hannah MA, Heyer AG, Hincha DK. A Global Survey of Gene Regulation during Cold Acclimation in Arabidopsisthaliana. *PLoS Genet* 2005; 1(2):e26.

[47] Hattori T, Terada T, Hamasuna S. Regulation of the Osem gene by abscisic acid and the transcriptional activator VP1: analysis of *cis*-acting promoter elements required for regulation by abscisic acid and VP1. *Plant J* 1995; 7(6):913-25.

[48] Herrero J, Diaz-Uriarte R, Dopazo J. Gene expression data preprocessing. *Bioinformatics* 2003; 19(5):655-6.

[49] Higo K, Ugawa Y, Iwamoto M, Korenaga T. Plant *cis*-acting regulatory DNA elements (PLACE) database: 1999. *Nucleic Acids Res* 1999; 27(1):297-300.

[50] Hobo T, Asada M, Kowyama Y, Hattori T. ACGT-containing abscisic acid response element (ABRE) and coupling element 3 (CE3) are functionally equivalent. *Plant J* 1999; 19(6):679-89.

[51] Hong RL, Hamaguchi L, Busch MA, Weigel D. Regulatory elements of the floral homeotic gene AGAMOUS identified by phylogenetic footprinting and shadowing. *Plant Cell* 2003; 15(6):1296-309.

[52] Hong SW, Jon JH, Kwak JM, Nam HG. Identification of a receptor-like protein kinase gene rapidly induced by abscisic acid, dehydration, high salt, and cold treatments in Arabidopsis thaliana. *Plant Physiol* 1997; 113(4):1203-12.

[53] Hoth S, Morgante M, Sanchez JP *et al.* Genome-wide gene expression profiling in Arabidopsis thaliana reveals new targets of abscisic acid and largely impaired gene regulation in the abi1-1 mutant. *J Cell Sci* 2002; 115(Pt 24):4891-900.

[54] Iwasaki T, Yamaguchi-Shinozaki K, Shinozaki K. Identification of a *cis*-regulatory region of a gene in Arabidopsis thaliana whose induction by dehydration is mediated by abscisic acid and requires protein synthesis. *Mol Gen Genet* 1995; 247(4):391-8.

[55] Kaldenhoff R, Kolling A, Richter G. A novel blue light- and abscisic acid-inducible gene of Arabidopsis thaliana encoding an intrinsic membrane protein. *Plant Mol Biol* 1993; 23(6):1187-98.

[56] Kel A, Wingender E. In silico analysis of gene regulatry sequences. Towards target gene identification. [10 th International Conference - Intelligent Systems for Molecular Biology (ISMB)]. 2. Edmonton, Canada. 3-8-2002.

[57] Kiyosue T, Yamaguchi-Shinozaki K, Shinozaki K. Characterization of two cDNAs (ERD10 and ERD14) corresponding to genes that respond rapidly to dehydration stress in Arabidopsis thaliana. *Plant Cell Physiol* 1994; 35(2):225-31.

[58] Knight H, Knight MR. Abiotic stress signalling pathways: specificity and cross-talk. *Trends Plant Sci* 2001; 6(6):262-7.

[59] Ko.S, Kamada.H. isolation of carrot basic leucine zipper transcription factor using yeast one-hybrid screening. *Plant Molecular Biology Reporter* 2002; 20:301a-h.

[60] Kovarik A, Matyasek R, Leitch A *et al.* Variability in CpNpG methylation in higher plant genomes. *Gene* 1997; 204:25-33.

[61] Krawczyk S, Thurow C, Niggeweg R, Gatz C. Analysis of the spacing between the two palindromes of activation sequence-1 with respect to binding to different TGA factors and transcriptional activation potential. *Nucleic Acids Res* 2002; 30(3):775-81.

[62] Lang V, Palva ET. The expression of a rab-related gene, rab18, is induced by abscisic acid during the cold acclimation process of Arabidopsis thaliana (L.) Heynh. *Plant Mol Biol* 1992; 20(5):951-62.

[63] Lee YH, Chun JY. A new homeodomain-leucine zipper gene from Arabidopsis thaliana induced by water stress and abscisic acid treatment. *Plant Mol Biol* 1998; 37(2):377-84.

[64] Lewin B. Genes VI. 6th Edition edn. New York: Oxford University Press, Inc., 1997.

[65] Manke T, Bringas R, Vingron M. Correlating protein-DNA and protein-protein interaction networks. *J Mol Biol* 2003; 333(1):75-85.

[66] Mantel N. The detection of disease clustering and a generalized regression approach. *Cancer Res* 1967; 27(2):209-20.

[67] Markstein M, Markstein P, Markstein V, Levine MS. Genome-wide analysis of clustered Dorsal binding sites identifies putative target genes in the Drosophila embryo. *Proc Natl Acad Sci U S A* 2002; 99(2):763-8.

[68] Maruyama K, Sakuma Y, Kasuga M *et al.* Identification of cold-inducible downstream genes of the Arabidopsis DREB1A/CBF3 transcriptional factor using two microarray systems. *Plant J* 2004; 38(6):982-93.

[69] Menkens AE, Schindler U, Cashmore AR. The G-box: a ubiquitous regulatory DNA element in plants bound by the GBF family of bZIP proteins. *Trends Biochem Sci* 1995; 20(12):506-10.

[70] Mikami K, Katagiri T, Iuchi S, Yamaguchi-Shinozaki K, Shinozaki K. A gene encoding phosphatidylinositol-4-phosphate 5-kinase is induced by water stress and abscisic acid in Arabidopsis thaliana. *Plant J* 1998; 15(4):563-8.

[71] Mundy J, Yamaguchi-Shinozaki K, Chua NH. Nuclear proteins bind conserved elements in the abscisic acid-responsive promoter of a rice rab gene. *Proc Natl Acad Sci U S A* 1990; 87(4):1406-10.

[72] Munnik T, Meijer HJ. Osmotic stress activates distinct lipid and MAPK signalling pathways in plants. *FEBS Lett* 2001; 498(2-3):172-8.

[73] Murashige T, Skoog F. A revised medium for rapid growth and bioassays with tobacco tissue cultures. *Physiol Plant* 1962; 15:473-97.

[74] Nakashima K, Shinwari ZK, Sakuma Y *et al.* Organization and expression of two Arabidopsis DREB2 genes encoding DRE-binding proteins involved in dehydration- and high-salinity-responsive gene expression. *Plant Mol Biol* 2000; 42(4):657-65.

[75] Narusaka Y, Nakashima K, Shinwari ZK *et al.* Interaction between two *cis*-acting elements, ABRE and DRE, in ABA-dependent expression of Arabidopsis rd29A gene in response to dehydration and high-salinity stresses. *Plant J* 2003; 34(2):137-48.

[76] Newman T, de Bruijn FJ, Green P *et al.* Genes galore: a summary of methods for accessing results from large-scale partial sequencing of anonymous Arabidopsis cDNA clones. *Plant Physiol* 1994; 106(4):1241-55.

[77] Ono A, Izawa T, Chua NH, Shimamoto K. The rab16B promoter of rice contains two distinct abscisic acid-responsive elements. *Plant Physiol* 1996; 112(2):483-91.

[78] Ornatowska M. Gene expression profiling in epidermal fragments and leaves of Arabidopsis thaliana after Abscisic Acid treatment. 2003.

[79] Papi M, Sabatini S, Bouchez D *et al.* Identification and disruption of an Arabidopsis zinc finger gene controlling seed germination. *Genes Dev* 2000; 14(1):28-33.

[80] Pavy N, Rombauts S, Dehais P *et al.* Evaluation of gene prediction software using a genomic data set: application to Arabidopsis thaliana sequences. *Bioinformatics* 1999; 15(11):887-99.

[81] Pfaffl MW. A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res* 2001; 29:2002-7.

[82] Piao HL, Pih KT, Lim JH *et al.* An Arabidopsis GSK3/shaggy-like gene that complements yeast salt stress-sensitive mutants is induced by NaCl and abscisic acid. *Plant Physiol* 1999; 119(4):1527-34.

[83] Pla M, Gomez J, Goday A, Pages M. Regulation of the abscisic acid-responsive gene rab28 in maize viviparous mutants. *Mol Gen Genet* 1991; 230(3):394-400.

[84] R Development Core Team. R: A Language and Environment for Statistical Computing. 2005. R Foundation for Statistical Computing.

[85] Rajeevan MS, Vernon SD, Taysavang N, Unger ER. Validation of array-based gene expression profiles by real-time (kinetic) RT-PCR. *J Mol Diagn* 2001; 3(1):26-31.

[86] Rateitschak K, Muller T, Vingron M. Annotating significant pairs of transcription factor binding sites in regulatory DNA. *In Silico Biol* 2004; 4(4):479-87.

[87] Reidt W, Wohlfarth T, Ellerstrom M *et al.* Gene regulation during late embryogenesis: the RY motif of maturation-specific gene promoters is a direct target of the FUS3 gene product. *Plant J* 2000; 21(5):401-8.

[88] Riano-Pachon DM, Dreyer I, Mueller-Roeber B. Orphan transcripts in Arabidopsis thaliana: identification of several hundred previously unrecognized genes. *Plant J* 2005; 43(2):205-12.

[89] Rice P, Longden I, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* 2000; 16(6):276-7.

[90] Richards S, Liu Y, Bettencourt BR *et al.* Comparative genome sequencing of Drosophila pseudoobscura: chromosomal, gene, and *cis*-element evolution. *Genome Res* 2005; 15(1):1-18.

[91] Roe JL, Nemhauser JL, Zambryski PC. TOUSLED participates in apical tissue formation during gynoecium development in Arabidopsis. *Plant Cell* 1997; 9(3):335-53.

[92] Rouse DT, Marotta R, Parish RW. Promoter and expression studies on an Arabidopsis thaliana dehydrin gene. *FEBS Lett* 1996; 381(3):252-6.

[93] Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math* 1987; 20:53-65.

[94] Sambrok JF, Maniatis T. Molecular cloning: A laboratory manual. 2nd Edition edn. Cold Spring Harbor Laboratory Press, 1989.

[95] Schlitt T, Palin K, Rung J *et al.* From gene networks to gene function. *Genome Res* 2003; 13(12):2568-76.

[96] Seki M, Carninci P, Nishiyama Y, Hayashizaki Y, Shinozaki K. High-efficiency cloning of Arabidopsis full-length cDNA by biotinylated CAP trapper. *Plant J* 1998; 15:707-20.

[97] Seki M, Narusaka M, Ishida J *et al.* Monitoring the expression profiles of 7000 Arabidopsis genes under drought, cold and high-salinity stresses using a full-length cDNA microarray. *Plant J* 2002; 31(3):279-92.

[98] Sharp RE. Interaction with ethylene: changing views on the role of abscisic acid in root and shoot growth responses to water stress. *Plant Cell Environ* 2002; 25(2):211-22.

[99] Sharp RE, LeNoble ME. ABA, ethylene and the control of shoot and root growth under water stress. *J Exp Bot* 2002; 53(366):33-7.

[100] Shen Q, Chen CN, Brands A, Pan SM, Ho TH. The stress- and abscisic acid-induced barley gene HVA22: developmental regulation and homologues in diverse organisms. *Plant Mol Biol* 2001; 45(3):327-40.

[101] Shen Q, Ho TH. Functional dissection of an abscisic acid (ABA)-inducible gene reveals two independent ABA-responsive complexes each containing a G-box and a novel *cis*-acting element. *Plant Cell* 1995; 7(3):295-307.

[102] Shen Q, Zhang P, Ho TH. Modular nature of abscisic acid (ABA) response complexes: composite promoter units that are necessary and sufficient for ABA induction of gene expression in barley. *Plant Cell* 1996; 8(7):1107-19.

[103] Shen QJ, Casaretto JA, Zhang P, Ho TH. Functional definition of ABA-response complexes: the promoter units necessary and sufficient for ABA induction of gene expression in barley ( Hordeum vulgare L.). *Plant Mol Biol* 2004; 54(1):111-24.

[104] Shi H, Zhu JK. Regulation of expression of the vacuolar Na+/H+ antiporter gene AtNHX1 by salt stress and abscisic acid. *Plant Mol Biol* 2002; 50(3):543-50.

[105] Shinozaki K, Yamaguchi-Shinozaki K, Seki M. Regulatory network of gene expression in the drought and cold stress responses. *Curr Opin Plant Biol* 2003; 6(5):410-7.

[106] Soderman E, Mattsson J, Engstrom P. The Arabidopsis homeobox gene ATHB-7 is induced by water deficit and by abscisic acid. *Plant J* 1996; 10(2):375-81.

[107] Steffens NO, Galuschka C, Schindler M, Bulow L, Hehl R. AthaMap: an online resource for in silico transcription factor binding sites in the Arabidopsis thaliana genome. *Nucleic Acids Res* 2004; 32 Database issue:D368-D372.

[108] Stormo GD. DNA binding sites: representation and discovery. *Bioinformatics* 2000; 16(1):16-23.

[109] Strizhov N, Abraham E, Okresz L *et al.* Differential expression of two P5CS genes controlling proline accumulation during salt-stress requires ABA and is regulated by ABA1, ABI1 and AXR2 in Arabidopsis. *Plant J* 1997; 12(3):557-69.

[110] Takahashi S, Katagiri T, Yamaguchi-Shinozaki K, Shinozaki K. An Arabidopsis gene encoding a Ca2+-binding protein is induced by abscisic acid during dehydration. *Plant Cell Physiol* 2000; 41(7):898-903.

[111] Thijs G. Probabilistic methods to search for regulatory elements in sets of coregulated genes. 2003. Katholieke Universiteit Leuven.

[112] Thijs G. Probabilistic methods to serach for regulatory elements in sets of coregulated genes. 2003. Katholieke Universiteit Leuven.

[113] Thijs G, Lescot M, Marchal K *et al.* A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics* 2001; 17(12):1113-22.

[114] Thijs G, Marchal K, Lescot M *et al.* A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. *J Comput Biol* 2002; 9(2):447-64.

[115] Thijs G, Moreau Y, De Smet F *et al.* INCLUSive: integrated clustering, upstream sequence retrieval and motif sampling. *Bioinformatics* 2002; 18(2):331-2.

[116] Thimm O, Essigmann B, Kloska S, Altmann T, Buckhout TJ. Response of Arabidopsis to iron deficiency stress as revealed by microarray analysis. *Plant Physiol* 2001; 127(3):1030-43.

[117] Thomas T.I, Chung H.J, Nunberg A.N. ABA signaling in plant development and growth. Series MCBU. Basel: Birkhäuser Verlag, 1997: 23-44.

[118] Thomas T.I, Chung H.J, Nunberg A.N. ABA signaling in plant development and growth. Series MCBU. Basel: Birkhäuser Verlag, 1997: 23-44.

[119] Tjian R. Molecular machines that control genes. *Sci Am* 1995; 272(2):54-61.

[120] Troyanskaya O, Cantor M, Sherlock G *et al.* Missing value estimation methods for DNA microarrays. *Bioinformatics* 2001; 17(6):520-5.

[121] Tyagi A.K, Gaur T. Light Regulation of Nuclear Photosynthetic Genes in Higher Plants. *Critical Reviews in Plant Sciences* 2003; 22(5):417-52.

[122] Uno Y, Furihata T, Abe H *et al.* Arabidopsis basic leucine zipper transcription factors involved in an abscisic acid-dependent signal transduction pathway under drought and high-salinity conditions. *Proc Natl Acad Sci U S A* 2000; 97(21):11632-11637.

[123] van Helden J ABC-VJ. Extracting regulatory sites from the upstream regions of yeast genes by computational analysis of oligonucleotide frequencies. *Journal Mol Biol* 1998; 281:827-42.

[124] Vavouri T, Elgar G. Prediction of *cis*-regulatory elements using binding site matrices - the successes, the failures and the reasons for both. *Curr Opin Genet Dev* 2005.

[125] Wang H, Qi Q, Schorr P *et al.* ICK1, a cyclin-dependent protein kinase inhibitor from Arabidopsis thaliana interacts with both Cdc2a and CycD3, and its expression is induced by abscisic acid. *Plant J* 1998; 15(4):501-10.

[126] Wang ML, Belmonte S, Kim U *et al.* A cluster of ABA-regulated genes on Arabidopsis thaliana BAC T07M07. *Genome Res* 1999; 9(4):325-33.

[127] Wilhelm KS, Thomashow MF. Arabidopsis thaliana cor15b, an apparent homologue of cor15a, is strongly responsive to cold and ABA, but not drought. *Plant Mol Biol* 1993; 23(5):1073-7.

[128] Williams J, Bulman M, Huttly A, Phillips A, Neill S. Characterization of a cDNA from Arabidopsis thaliana encoding a potential thiol protease whose expression is induced independently by wilting and abscisic acid. *Plant Mol Biol* 1994; 25(2):259-70.

[129] Wingender E, Dietze P, Karas H, Knuppel R. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res* 1996; 24(1):238-41.

[130] Wu Y, Thorne ET, Sharp RE, Cosgrove DJ. Modification of expansin transcript levels in the maize primary root at low water potentials. *Plant Physiol* 2001; 126(4):1471-9.

[131] Xiong L, Gong Z, Rock CD *et al.* Modulation of abscisic acid signal transduction and biosynthesis by an Sm-like protein in Arabidopsis. *Dev Cell* 2001; 1(6):771-81.

[132] Yamaguchi-Shinozaki K, Shinozaki K. Characterization of the expression of a desiccation-responsive rd29 gene of Arabidopsis thaliana and analysis of its promoter in transgenic plants. *Mol Gen Genet* 1993; 236(2-3):331-40.

[133] Yamaguchi-Shinozaki K, Shinozaki K. The plant hormone abscisic acid mediates the drought-induced expression but not the seed-specific expression of rd22, a gene responsive to dehydration stress in Arabidopsis thaliana. *Mol Gen Genet* 1993; 238(1-2):17-25.

[134] Yamaguchi-Shinozaki K, Shinozaki K. A novel *cis*-acting element in an Arabidopsis gene is involved in responsiveness to drought, low-temperature, or high-salt stress. *Plant Cell* 1994; 6(2):251-64.

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

[136] Yoshiba Y, Kiyosue T, Katagiri T *et al.* Correlation between the induction of a gene for delta 1-pyrroline-5-carboxylate synthetase and the accumulation of proline in Arabidopsis thaliana under osmotic stress. *Plant J* 1995; 7(5):751-60.

[137] Zhang W, Ruan J, Ho TH *et al.* *Cis*-regulatory element based targeted gene finding: genome-wide identification of ABA- and abiotic stress-responsive genes in Arabidopsis thaliana. *Bioinformatics* 2005.

[138] Zhu JK. Salt and drought stress signal transduction in plants. *Annu Rev Plant Biol* 2002; 53:247-73.

# *In silico* Identification of Genes Regulated by Abscisic Acid in Arabidopsis thaliana (L.) Heynh.

## Summary

Abscisic acid (ABA) is a major plant hormone that plays an important role during plant growth and development. During vegetative growth ABA mediates (in part) responses to various environmental stresses such as cold, drought and high salinity. The response triggered by ABA includes changes in the transcript level of genes involved in stress tolerance. The aim of this project was the *In silico* identification of genes putatively regulated by ABA in *A. thaliana*. *In silico* predictions were combined with experimental data in order to evaluate the reliability of computational predictions.

Taking advantage of the genome sequence of *A. thaliana* publicly available since 2000, 1 kb upstream sequences were screened for combinations of *cis*-elements known to be involved in the regulation of ABA-responsive genes. It was found that around 10 to 20 percent of the genes of *A. thaliana* might be regulated by ABA.

Further analyses of the predictions revealed that certain combinations of *cis*-elements that confer ABA-responsiveness were significantly over-represented compared with results in random sequences and with random expectations. In addition, it was observed that other combinations that confer ABA-responsiveness in monocotyledonous species might not be functional in *A. thaliana*. It is proposed that ABA-responsive genes in *A. thaliana* show pairs of ABRE (abscisic acid responsive element) with MYB binding sites, DRE (dehydration responsive element) or with itself.

The analysis of the distances between pairs of *cis*-elements suggested that pairs of ABREs are bound by homodimers of ABRE binding proteins. In contrast, pairs between MYB binding sites and ABRE, or DRE and ABRE showed a distance between *cis*-elements that suggested that the binding proteins interact through protein complexes and not directly.

The comparison of computational predictions with experimental data confirmed that the regulatory mechanisms leading to the induction or repression of genes by ABA is very incompletely understood. It became evident that besides the *cis*-elements proposed in this study to be present in ABA-responsive genes, other known and unknown *cis*-elements might play an important role in the transcriptional regulation of ABA-responsive genes. For example, auxin-related cis elements, or the *cis*-elements recognized by the NAM-family of transcription factors (Non-Apical meristem).

This work documents the use of computational and experimental approaches to analyse possible interactions between *cis*-elements involved in the regulation of ABA-responsive genes. The computational predictions allowed the distinction between putatively relevant combinations of *cis*-elements from irrelevant combinations of *cis*-elements in ABA-responsive genes. The comparison with experimental data allowed to identify certain *cis*-elements that have not been previously associated to the ABA-mediated transcriptional regulation, but that might be present in ABA-responsive genes (e.g. auxin responsive elements). Moreover, the efforts to unravel the gene regulatory network associated with the ABA-signalling pathway revealed that NAM-transcription factors and their corresponding binding sequences are important components of this network.

## Acknowledgments

First and foremost, I would like to thank Prof. Dr. Bernd Müller-Röber for giving me the possibility to carry out my PhD project in his lab, and for his continuous support. I am grateful for the opportunity I received to work under his supervision, and for the scientific freedom I had during my PhD.

I would like to thank Prof. Dr. Joachim Selbig for his interest in my work, and his helpful support during my graduate studies. And also thanks to Peter Krüger and the crew of the Bioinformatic Group at the Max-Planck Institute for Molecular Plant Physiology.

I am especially grateful to Dr. Jorge Mayer (Freiburg) for his critical discussion and proof reading of my manuscript.

I would like to thank Dr. Edward Oakeley (Basel) for agreeing to examine my thesis, and Prof. Dr. Thomas Altmann for immeadiatly accepting the task of being “Vorsitzender der Prüfungskommission”.

I’d like to thank Dr. Ingo Dreyer and Diego Mauricio Riaño for their helpful comments and fruitful discussions throughout my graduate studies, for their enormous support and patience.

Further I wish to thank all my colleagues and friends at the University of Potsdam and at the Max-Planck Institute of Molecular Plant Physiology. Specially Dr. Magdalena Ornatowska who provided important experimental data. I am also grateful to Dr. Fernando Gómez, Dr. Maria Ines Zanor and Fernando Arana whose help in the lab were irreplaceable.

Last but not least, I am indebted to my German family in Laatzen, and my family in Colombia. They have been an enormous support and an incredible source of energy and happiness.

# Appendix 1: Binding sites and frequency matrices

## 1. ABRE

### 1.1. Binding sites

| Organism | Gene | Sequence | Reference |
|---|---|---|---|
| *H. vulgare* | *HVA22* | gccACGTacac | [16] |
| *H. vulgare* | *HVA1* | cctACGTggcg | [16] |
| *H. vulgare* | *HVA2* | cgcACGTgtcg | [15] |
| *Z. mays* | *RAB17* | cgtACGTgtac | [15] |
| *T. aestivum* | *EM* | cacACGTgccg | [15] |
| *C. plantagineum* | *CDET27-45* | ggcACGTatgt | [15] |
| *O. sativa* | *RAB16A* | cgtACGTggcg | [15] |
| *O. sativa* | *RAB16D* | cgtACGTggct | [15] |
| *L. esculentum* | *LE25* | aaaACGTgtca | [15] |
| *A. thaliana* | *RAB18* | attACGTgtcc | [15] |
| *O. sativa* | *RAB16B* | tacACGTccct | [15] |
| *O. sativa* | *RAB16C* | tacACGTaccc | [15] |
| *O. sativa* | *RAB16C* | cacACGTcctt | [15] |
| *O. sativa* | *RAB16C* | catACGTggcg | [15] |
| *Z. mays* | *RAB18* | gccACGTgggc | [4] |
| *Z. mays* | *RAB18* | tccACGTctct | [4] |
| *O. sativa* | *OSEM* | cgtACGTgtcg | [9] |
| *H. vulgare* | *HVA1* | cctACGTggcg | [9] |
| *O. sativa* | *RAB17* | gagACGTggcg | [3] |
| *O. sativa* | *RAB17* | cacACGTcccg | [3] |
| *O. sativa* | *RAB17* | cgtACGTgtac | [3] |
| *O. sativa* | *RAB17* | tgtACGTgctg | [3] |
| *G. hirsutum* | *LEA D-7* | gatACGTgttt | [2] |
| *G. hirsutum* | *LEA D-19* | cttACGTggat | [2] |
| *G. hirsutum* | *LEA D-34* | gttACGTgtta | [2] |
| *G. hirsutum* | *LEA D-1113* | tatACGTggca | [2] |
| *A. thaliana* | *RD29B* | taaACGTggac | [22] |
| *A. thaliana* | *RD29B* | cgtACGTgtca | [22] |
| *A. thaliana* | | cggACGTgtcg | [8] |
| *A. thaliana* | | gcgacACGTac | [8] |
| *A. thaliana* | *ATMYB74* | aggacACGTaa | [7] |
| *A. thaliana* | *ATMYB102* | gggacACGTat | [7] |

ABRE binding sites - 1

| | | | |
|---|---|---|---|
| *A. thaliana* | *AT1G51090* | atgACGTgtat | [12] |
| *A. thaliana* | *RD29A* | catACGTgtcc | [12] |
| *A. thaliana* | *COR15A* | tacACGTggcc | [12] |
| *A. thaliana* | *COR15A* | gccACGTgtaa | [12] |
| *A. thaliana* | *COR15A* | ttcACGTgtat | [12] |
| *A. thaliana* | *COR15A* | aatACGTgtaa | [12] |
| *A. thaliana* | *AT1G01470* | gtcACGTgttg | [12] |
| *A. thaliana* | *AT1G01470* | tatACGTgtct | [12] |
| *A. thaliana* | *AT1G01470* | tgtACGTgtga | [12] |
| *A. thaliana* | *HVA22D* | cacACGTggcg | [12] |
| *A. thaliana* | *HVA22D* | tcgACGTgtgg | [12] |
| *A. thaliana* | *XERO2* | aatACGTgttg | [12] |
| *A. thaliana* | *At4g01020* | ggaACGTgtaa | [12] |
| *A. thaliana* | *At4g01020* | agcACGTgtgt | [12] |
| *A. thaliana* | *At4g01020* | attACGTgtct | [12] |
| *A. thaliana* | *At2g15320* | ctaACGTgtta | [12] |
| *A. thaliana* | *At2g15320* | aagACGTggtg | [12] |
| *A. thaliana* | *At3g46640* | ccaACGTggac | [12] |
| *A. thaliana* | *At3g46640* | tccACGTggct | [12] |
| *A. thaliana* | *At3g46640* | atgACGTgttg | [12] |
| *A. thaliana* | *Hos1* | atcACGTgtcc | [12] |
| *A. thaliana* | *At3g50960* | aatACGTgttg | [12] |
| *A. thaliana* | *At1g27730* | gaaACGTgtac | [12] |
| *A. thaliana* | *At1g27730* | cacACGTgtac | [12] |
| *A. thaliana* | *ERD7* | ttaACGTggca | [12] |
| *A. thaliana* | *ERD7* | aagACGTggat | [12] |
| *A. thaliana* | *PDC1* | tatACGTggga | [12] |
| *A. thaliana* | *PDC1* | tctACGTgtat | [12] |
| *A. thaliana* | *At4g46768* | agaACGTgtca | [12] |
| CONSENSUS | | nryACGTgtm | |

### 1.2. ABRE_PSMF

$$\begin{array}{c} \mathbf{A} \\ \mathbf{C} \\ \mathbf{G} \\ \mathbf{T} \end{array} \begin{pmatrix} 14 & 23 & 8 & 61 & 0 & 3 & 0 & 3 & 0 & 18 & 13 \\ 20 & 11 & 19 & 0 & 61 & 0 & 3 & 4 & 7 & 27 & 14 \\ 12 & 16 & 10 & 0 & 0 & 58 & 0 & 54 & 18 & 6 & 19 \\ 15 & 11 & 24 & 0 & 0 & 0 & 58 & 0 & 36 & 10 & 15 \end{pmatrix}$$

## 2. As1

### 2.1. Binding sites

| Organism | Sequence | Reference |
|---|---|---|
| *A. thaliana* | tgACGtcA | [14] |
| *CaMV* | taACGtaa | [11] |
|  | tgACGaaa | [18] |
| *N. tabacum* | taAGCtaa | [11] |
| *Agrobacterium* | tgACG | [18] |
|  | tgACGtc | [18] |
| *N. tabacum* | ttACGcaa | [11] |
| *N. tabacum* | ttAGCtaa | [11] |
| *soybean* | ttACGtaa | [11] |
| *A. thaliana* | ttATGtca | [11] |
| *CaMV* | tgACGtaa | [11] |
| *Agrobacterium* | aaACGtaa | [11] |
| N. tabacum | taACGtca | [11] |
| *A. thaliana* | ctACGtca | [11] |
| CONSENSUS | tdACGtaa |  |

### 2.2. As1_PSFM

$$\begin{array}{cc} \begin{matrix} \mathbf{A} \\ \mathbf{C} \\ \mathbf{G} \\ \mathbf{T} \end{matrix} & \begin{pmatrix} 1 & 4 & 14 & 0 & 0 & 1 & 8 & 12 \\ 1 & 0 & 0 & 11 & 2 & 1 & 5 & 0 \\ 0 & 5 & 0 & 2 & 12 & 0 & 0 & 0 \\ 12 & 5 & 0 & 1 & 0 & 11 & 0 & 0 \end{pmatrix} \end{array}$$

## 3. CE3

### 3.1. Binding sites

| Organism | Gene | Sequence | Reference |
|---|---|---|---|
| *H. vulgare* | *HVA1* | cgCGTgtcctc | [16] |
| *Z. mays* | *RAB18* | cgCGCctcctc | [3] |
| *O. sativa* | *OSEM* | gaCGCgtgtcg | [9] |
| H. vulgare | *HVA1* | aaCGCgtgtcc | [9] |
| *O. sativa* | *RAB16B* | gcCGCgtggca | [9] |
| *O. sativa* | *mutant seq* | gaCGCgtggcc | [9] |
| *O. sativa* | *LPT2* | aCGCgtgg | [5] |
| *O. sativa* | *RAB16A* | ccCGCcgcgct | [13] |
| *O. sativa* | *RAB16B* | caCGGcgcgct | [13] |
| *O. sativa* | *RAB16C* | ccCGGcgcgct | [13] |
| *O. sativa* | *RAB16D* | ggCGCcgcgct | [13] |
| *O. sativa* | *RAB16D* | acCGCcgcgcc | [13] |
| O. sativa | *OSEM* | gcGGCctcgcc | [8] |
| CONSENSUS | | smCGCstcgcy | |

### 3.2. CE3_PSFM

$$\begin{array}{cc} \begin{matrix}\mathbf{A}\\ \mathbf{C}\\ \mathbf{G}\\ \mathbf{T}\end{matrix} & \begin{pmatrix} 2 & 5 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 5 & 5 & 12 & 0 & 10 & 7 & 0 & 8 & 2 & 10 & 6 \\ 5 & 3 & 1 & 13 & 2 & 6 & 5 & 5 & 9 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 8 & 0 & 2 & 2 & 4 \end{pmatrix} \end{array}$$

## 4. DRE

### 4.1. DRE binding sites

| Organism | Gene | Sequence | Reference |
|---|---|---|---|
| *A. thaliana* | *COR15A* | ggCCGACctc | [19] |
| *A. thaliana* | *COR15B* | ggCCGACctc | [19] |
| *A. thaliana* | *COR78-1* | taCCGACat | [19] |
| *Z. mays* | *RAB17* | aaCCGAGac | [3] |
| *Z. mays* | *RAB17* | caCCGACgc | [3] |
| *A. thaliana* | *RD29A* | taCCGACat | [22] |
| *A. thaliana* | *RD29A* | taCCGACat | [22] |
| *A. thaliana* | *RD17* | taCCGACtt | [22] |
| *H. vulgare* | *HVA1* | tgCCGACgc | [22] |
| *A. thaliana* | *COR15A* | ggCCGACat | [23] |
| *A. thaliana* | *KIN1* | taCCGACat | [23] |
| *A. thaliana* | *At1g16850* | tgCCGACtc | [23] |
| *A. thaliana* | *RD29A* | gaCCGACta | [23] |
| *A. thaliana* | *RD29A* | agCCGACac | [23] |
| *A. thaliana* | *RD17* | gaCCGACat | [23] |
| *A. thaliana* | *RD17* | agCCGACca | [23] |
| *A. thaliana* | *At2g42530* | ggCCGACct | [23] |
| *A. thaliana* | *At2g15970* | taCCGACat | [23] |
| *A. thaliana* | *KIN2* | taCCGACat | [23] |
| *A. thaliana* | *ERD10* | gaCCGACat | [23] |
| *A. thaliana* | *ERD10* | tgCCGACgt | [23] |
| *A. thaliana* | *At1g51090* | cgCCGACat | [12] |
| *A. thaliana* | *At1g51090* | agCCGACat | [12] |
| *A. thaliana* | *COR15A* | ggCCGACct | [12] |
| *A. thaliana* | *COR15A* | gaCCGACag | [12] |
| *A. thaliana* | *At1g01470* | caCCGACgt | [12] |
| *A. thaliana* | *At1g01470* | gaCCGACtt | [12] |
| *A. thaliana* | *At1g01470* | gaCCGACca | [12] |
| *A. thaliana* | *HVA22D* | caCCGACCg | [12] |
| *A. thaliana* | *HVA22D* | gaCCGACgt | [12] |
| *A. thaliana* | *HVA22D* | caCCGACct | [12] |
| *A. thaliana* | *XERO2* | caCCGACgt | [12] |
| *A. thaliana* | *XERO2* | gaCCGACgt | [12] |
| *A. thaliana* | *At1g01020* | caCCGACat | [12] |
| *A. thaliana* | *At1g01020* | caCCGACtt | [12] |

DRE binding sites - 1

| | | | |
|---|---|---|---|
| *A. thaliana* | *At1g01020* | aaCCGACaa | [12] |
| *A. thaliana* | *At2g15320* | agCCGACat | [12] |
| *A. thaliana* | *At2g15320* | agCCGACct | [12] |
| *A. thaliana* | *At3g46640* | cgCCGACgg | [12] |
| *A. thaliana* | *At3g46640* | taCCGACat | [12] |
| *A. thaliana* | *At3g46640* | aaCCGACct | [12] |
| *A. thaliana* | *HOS1* | tgCCGACtt | [12] |
| *A. thaliana* | *HOS1* | taCCGACtt | [12] |
| *A. thaliana* | *HOS1* | tgCCGACct | [12] |
| *A. thaliana* | *At3g50960* | gaCCGACgt | [12] |
| *A. thaliana* | *At3g50960* | ggCCGACat | [12] |
| *A. thaliana* | *At3g50960* | caCCGACgt | [12] |
| *A. thaliana* | *FP6* | caCCGACgt | [12] |
| *A. thaliana* | *FP6* | taCCGACct | [12] |
| *A. thaliana* | *FP6* | agCCGACct | [12] |
| *A. thaliana* | *ERD7* | gaCCGACcg | [12] |
| *A. thaliana* | *ERD7* | gaCCGACca | [12] |
| *A. thaliana* | *PDC1* | taCCGACat | [12] |
| *A. thaliana* | *At1g35300* | tgCCGACat | [12] |
| *A. thaliana* | *At4g14000* | gaCCGACtt | [12] |
| *A. thaliana* | *At4g14000* | taCCGACcg | [12] |
| *A. thaliana* | At4g14000 | agCCGACta | [12] |
| *A. thaliana* | *At4g14000* | agCCGACca | [12] |
| *A. thaliana* | *At4g14000* | taCCGACtg | [12] |
| *A. thaliana* | *At4g15910* | caCCGACct | [12] |
| *A. thaliana* | *At4g46768* | ggCCGACat | [12] |
| *A. thaliana* | *At4g46768* | gaCCGACct | [12] |
| CONSENSUS | | kaCCGACmt | |

### 4.2. DRE_PSFM

$$\begin{array}{c} \mathbf{A} \\ \mathbf{C} \\ \mathbf{G} \\ \mathbf{T} \end{array} \begin{pmatrix} 11 & 39 & 0 & 0 & 0 & 62 & 0 & 22 & 7 \\ 12 & 0 & 62 & 62 & 0 & 0 & 61 & 19 & 5 \\ 20 & 23 & 0 & 0 & 62 & 0 & 1 & 11 & 6 \\ 19 & 0 & 0 & 0 & 0 & 0 & 0 & 10 & 44 \end{pmatrix}$$

## 5. MYB

### 5.1. MYB binding sites

| Organism | Gene | Sequence | Reference |
|---|---|---|---|
| *A. thaliana* | *RD22* | tcAACCa | [1] |
| *A. thaliana* | | attAACTg | [10] |
| *A. thaliana* | | gtcTAACc | [10] |
| | | cccAACTg | [20] |
| *A. thaliana* | *At5g44420* | attAACTa | [1] |
| *A. thaliana* | *At5g44420* | gttAACTa | [1] |
| *A. thaliana* | *At1g75830* | gctAACCa | [1] |
| *A. thaliana* | *At1g75830* | ttaAACCa | [1] |
| *A. thaliana* | *At1g75830* | cctAACCa | [1] |
| *A. thaliana* | *At2g25510* | ttaAACCa | [1] |
| *A. thaliana* | *At3g03270* | agaAACCa | [1] |
| *A. thaliana* | *At3g03270* | ccaAACCa | [1] |
| *A. thaliana* | *At1g77120* | ataAACCa | [1] |
| *A. thaliana* | *At1g77120* | tatAACCa | [1] |
| *A. thaliana* | *At5g25980* | ctcAACGg | [1] |
| *A. thaliana* | *At5g25980* | actAACCa | [1] |
| *A. thaliana* | *At5g25980* | gcaAACCa | [1] |
| *A. thaliana* | *At3g14210* | cttAACTg | [1] |
| *A. thaliana* | *At5g19550* | attAACCa | [1] |
| *A. thaliana* | *At5g19550* | cctAACCa | [1] |
| *A. thaliana* | *At1g52400* | tctAACTg | [1] |
| *A. thaliana* | *At1g52400* | actAACCa | [1] |
| *A. thaliana* | *At4g21830* | ttaAACCa | [1] |
| *A. thaliana* | *At4g21830* | catAACCa | [1] |
| *A. thaliana* | *At4g08870* | ccaAACCa | [1] |
| *A. thaliana* | *At4g08870* | ataAACCa | [1] |
| *A. thaliana* | *At4g08870* | tctAAACt | [1] |
| *A. thaliana* | *At1g07920* | accAACGg | [1] |
| *A. thaliana* | *At1g07920* | ttaAACCa | [1] |
| *A. thaliana* | *At1g07920* | gtaAACCa | [1] |
| *A. thaliana* | *At1g07920* | gctAACCa | [1] |
| *A. thaliana* | *KIN2* | ctaAACCa | [1] |
| *A. thaliana* | *KIN2* | actAACCa | [1] |
| *A. thaliana* | *At1g31580* | gttAACCa | [1] |
| *A. thaliana* | *At1g31580* | cctAACTg | [1] |

MYB binding sites - 1

| | | | |
|---|---|---|---|
| *A. thaliana* | *At2g28000* | tctAACCa | [1] |
| *A. thaliana* | *At2g28000* | atcAACGg | [1] |
| *A. thaliana* | *At2g28000* | gatAACCa | [1] |
| *A. thaliana* | *At5g42530* | ccaAACCa | [1] |
| *A. thaliana* | *At5g42530* | ccaAACCa | [1] |
| *A. thaliana* | *At5g24770* | ttaAACCa | [1] |
| *A. thaliana* | *At5g20830* | ataAACCa | [1] |
| *A. thaliana* | *At5g20830* | gctAACCa | [1] |
| *A. thaliana* | *At2g19590* | cctAACGg | [1] |
| *A. thaliana* | *At2g25510* | ctaAACCa | [1] |
| *A. thaliana* | *At4g08870* | ttaAACCa | [1] |
| *A. thaliana* | *At4g08870* | agtAACCa | [1] |
| *A. thaliana* | *At4g08870* | cctAACCa | [1] |
| *A. thaliana* | *At5g25980* | gtaAACCa | [1] |
| *A. thaliana* | *At5g25980* | tctAACCa | [1] |
| *A. thaliana* | *At5g25980* | ataAACCa | [1] |
| *A. thaliana* | *At3g14210* | ccaAACCa | [1] |
| *A. thaliana* | *At1g31580* | ctcAACTg | [1] |
| *A. thaliana* | *At1g31580* | ataAACCa | [1] |
| *A. thaliana* | *At1g77120* | tcaAACCa | [1] |
| *A. thaliana* | *At1g77120* | gtaAACCa | [1] |
| *A. thaliana* | *At2g03770* | tctAACCa | [1] |
| *A. thaliana* | *At3g57050* | cttAACCa | [1] |
| *A. thaliana* | *At3g57050* | acaAACCa | [1] |
| *A. thaliana* | *At3g57050* | gcaAACCa | [1] |
| *A. thaliana* | *At3g29930* | acaAACCa | [1] |
| *A. thaliana* | *At3g57050* | cttAACTg | [1] |
| *A. thaliana* | *RD22* | tggTTAGc | [1] |
| *A. thaliana* | *MYB74* | cagTTGAc | [1] |
| *P. hybrida* | | acaTTTGa | [17] |
| *P. hybrida* | | aagTTAGt | [17] |
| | | tccAAACg | [21] |
| | | aatTAACt | [21] |
| *A. thaliana* | *RD29B* | gagCAACt | [20] |
| | | attTAACt | [6] |
| *CONSENSUS* | | mywAACCa | |

## 5.2. MYB_PSFM

$$\begin{array}{cc} \begin{array}{c} \mathbf{A} \\ \mathbf{C} \\ \mathbf{G} \\ \mathbf{T} \end{array} & \begin{pmatrix} 21 & 7 & 29 & 62 & 65 & 8 & 2 & 49 \\ 20 & 31 & 7 & 1 & 1 & 60 & 52 & 3 \\ 13 & 3 & 4 & 0 & 0 & 1 & 7 & 12 \\ 16 & 29 & 30 & 7 & 4 & 1 & 9 & 6 \end{pmatrix} \end{array}$$

## REFERENCES

[1] Abe H, Urao T, Ito T *et al.* Arabidopsis AtMYC2 (bHLH) and AtMYB2 (MYB) function as transcriptional activators in abscisic acid signaling. *Plant Cell* 2003; 15(1):63-78.

[2] Baker SS, Wilhelm KS, Thomashow MF. The 5'-region of Arabidopsis thaliana cor15a has *cis*-acting elements that confer cold-, drought- and ABA-regulated gene expression. *Plant Mol Biol* 1994; 24(5):701-13.

[3] Busk PK, Jensen AB, Pages M. Regulatory elements in vivo in the promoter of the abscisic acid responsive gene rab17 from maize. *Plant J* 1997; 11(6):1285-95.

[4] Busk PK, Pages M. Protein binding to the abscisic acid-responsive element is independent of VIVIPAROUS1 in vivo. *Plant Cell* 1997; 9(12):2261-70.

[5] Busk PK, Pages M. Regulation of abscisic acid-induced transcription. *Plant Mol Biol* 1998; 37(3):425-35.

[6] Chen CN, Chu CC, Zentella R, Pan SM, Ho TH. AtHVA22 gene family in Arabidopsis: phylogenetic relationship, ABA and stress regulation, and tissue-specific expression. *Plant Mol Biol* 2002; 49(6):633-44.

[7] Denekamp M, Smeekens SC. Integration of wounding and osmotic stress signals determines the expression of the AtMYB102 transcription factor gene. *Plant Physiol* 2003; 132(3):1415-23.

[8] Hattori T, Terada T, Hamasuna S. Regulation of the Osem gene by abscisic acid and the transcriptional activator VP1: analysis of *cis*-acting promoter elements required for regulation by abscisic acid and VP1. *Plant J* 1995; 7(6):913-25.

[9] Hobo T, Asada M, Kowyama Y, Hattori T. ACGT-containing abscisic acid response element (ABRE) and coupling element 3 (CE3) are functionally equivalent. *Plant J* 1999; 19(6):679-89.

[10] Iwasaki T, Yamaguchi-Shinozaki K, Shinozaki K. Identification of a *cis*-regulatory region of a gene in Arabidopsis thaliana whose induction by dehydration is mediated by abscisic acid and requires protein synthesis. *Mol Gen Genet* 1995; 247(4):391-8.

[11] Krawczyk S, Thurow C, Niggeweg R, Gatz C. Analysis of the spacing between the two palindromes of activation sequence-1 with respect to binding to different TGA factors and transcriptional activation potential. *Nucleic Acids Res* 2002; 30(3):775-81.

[12] Maruyama K, Sakuma Y, Kasuga M *et al.* Identification of cold-inducible downstream genes of the Arabidopsis DREB1A/CBF3 transcriptional factor using two microarray systems. *Plant J* 2004; 38(6):982-93.

[13] Mundy J, Yamaguchi-Shinozaki K, Chua NH. Nuclear proteins bind conserved elements in the abscisic acid-responsive promoter of a rice rab gene. *Proc Natl Acad Sci U S A* 1990; 87(4):1406-10.

[14] Narusaka Y, Nakashima K, Shinwari ZK *et al.* Interaction between two *cis*-acting elements, ABRE and DRE, in ABA-dependent expression of Arabidopsis rd29A gene in response to dehydration and high-salinity stresses. *Plant J* 2003; 34(2):137-48.

[15] Shen Q, Ho TH. Functional dissection of an abscisic acid (ABA)-inducible gene reveals two independent ABA-responsive complexes each containing a G-box and a novel *cis*-acting element. *Plant Cell* 1995; 7(3):295-307.

[16] Shen Q, Zhang P, Ho TH. Modular nature of abscisic acid (ABA) response complexes: composite promoter units that are necessary and sufficient for ABA induction of gene expression in barley. *Plant Cell* 1996; 8(7):1107-19.

[17] Solano R, Nieto C, Avila J *et al.* Dual DNA binding specificity of a petal epidermis-specific MYB transcription factor (MYB.Ph3) from Petunia hybrida. *EMBO J* 1995; 14(8):1773-84.

[18] Steffens NO, Galuschka C, Schindler M, Bulow L, Hehl R. AthaMap: an online resource for in silico transcription factor binding sites in the Arabidopsis thaliana genome. *Nucleic Acids Res* 2004; 32 Database issue:D368-D372.

[19] Stockinger EJ, Gilmour SJ, Thomashow MF. Arabidopsis thaliana CBF1 encodes an AP2 domain-containing transcriptional activator that binds to the C-repeat/DRE, a *cis*-acting DNA regulatory element that stimulates transcription in response to low temperature and water deficit. *Proc Natl Acad Sci U S A* 1997; 94(3):1035-40.

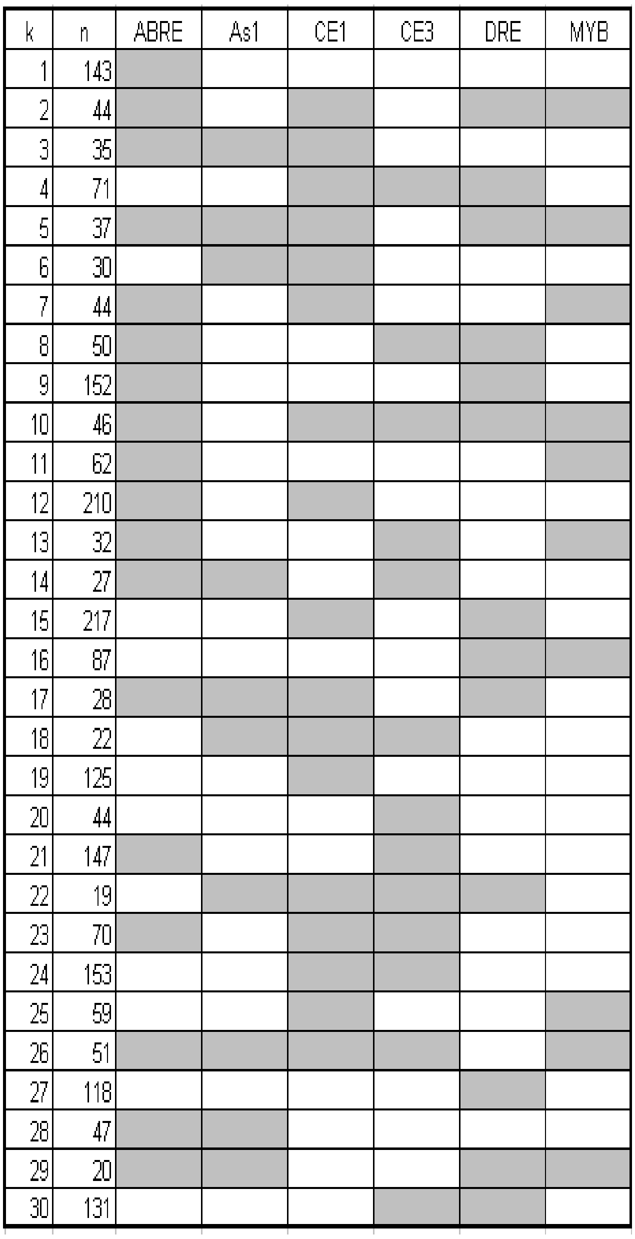
[20] Uno Y, Furihata T, Abe H *et al.* Arabidopsis basic leucine zipper transcription factors involved in an abscisic acid-dependent signal transduction pathway under drought and high-salinity conditions. *Proc Natl Acad Sci U S A* 2000; 97(21):11632-7.

[21] Urao T, Yamaguchi-Shinozaki K, Urao S, Shinozaki K. An Arabidopsis myb homolog is induced by dehydration stress and its gene product binds to the conserved MYB recognition sequence. *Plant Cell* 1993; 5(11):1529-39.

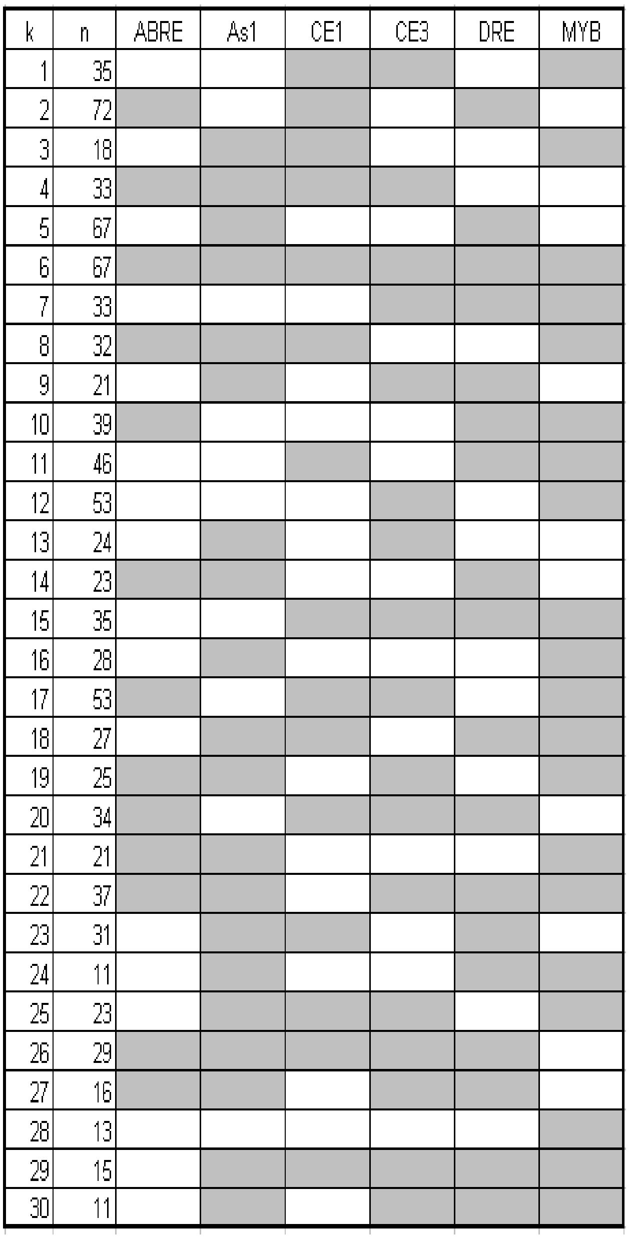
[22] Yamaguchi-Shinozaki K, Shinozaki K. A novel *cis*-acting element in an Arabidopsis gene is involved in responsiveness to drought, low-temperature, or high-salt stress. *Plant Cell* 1994; 6(2):251-64.

[23] Yoshiba Y, Kiyosue T, Katagiri T *et al.* Correlation between the induction of a gene for delta 1-pyrroline-5-carboxylate synthetase and the accumulation of proline in Arabidopsis thaliana under osmotic stress. *Plant J* 1995; 7(5):751-60.

**Appendix 2: Combinations of *cis*-elements.** CISTER results. *Cis*-elements per cluster. Number of clusters k=60. k=cluster number. n=*cis*-elements per cluster. Cells shaded in grey indicate that at least one *cis*-element of the type was observed in that cluster.is grey shaded.



| k | n | ABRE | As1 | CE1 | CE3 | DRE | MYB |
|---|---|---|---|---|---|---|---|
| 1 | 143 | ▨ | | | | | |
| 2 | 44 | ▨ | | ▨ | | ▨ | ▨ |
| 3 | 35 | ▨ | ▨ | ▨ | | | |
| 4 | 71 | | | ▨ | ▨ | ▨ | |
| 5 | 37 | ▨ | ▨ | ▨ | | ▨ | ▨ |
| 6 | 30 | | ▨ | ▨ | | | |
| 7 | 44 | ▨ | | ▨ | | | ▨ |
| 8 | 50 | ▨ | | | ▨ | ▨ | |
| 9 | 152 | ▨ | | | | ▨ | |
| 10 | 46 | ▨ | | ▨ | ▨ | ▨ | ▨ |
| 11 | 62 | ▨ | | | | | ▨ |
| 12 | 210 | ▨ | | ▨ | | | |
| 13 | 32 | ▨ | | | ▨ | | ▨ |
| 14 | 27 | ▨ | ▨ | | ▨ | | |
| 15 | 217 | | | ▨ | | ▨ | |
| 16 | 87 | | | | | ▨ | ▨ |
| 17 | 28 | ▨ | ▨ | ▨ | | ▨ | |
| 18 | 22 | | ▨ | ▨ | ▨ | | |
| 19 | 125 | | | ▨ | | | |
| 20 | 44 | | | | ▨ | | |
| 21 | 147 | ▨ | | | ▨ | | |
| 22 | 19 | | ▨ | ▨ | ▨ | ▨ | |
| 23 | 70 | ▨ | | ▨ | ▨ | | |
| 24 | 153 | | | ▨ | ▨ | | |
| 25 | 59 | | | ▨ | | | ▨ |
| 26 | 51 | ▨ | ▨ | ▨ | ▨ | | ▨ |
| 27 | 118 | | | | | ▨ | |
| 28 | 47 | ▨ | ▨ | | | | |
| 29 | 20 | ▨ | ▨ | | | ▨ | ▨ |
| 30 | 131 | | | | ▨ | ▨ | |



| k | n | ABRE | As1 | CE1 | CE3 | DRE | MYB |
|---|---|---|---|---|---|---|---|
| 1 | 35 | | | ▨ | ▨ | | ▨ |
| 2 | 72 | ▨ | | ▨ | | ▨ | |
| 3 | 18 | | ▨ | ▨ | | | ▨ |
| 4 | 33 | ▨ | ▨ | ▨ | ▨ | | |
| 5 | 67 | | ▨ | | | ▨ | |
| 6 | 67 | ▨ | ▨ | ▨ | ▨ | ▨ | ▨ |
| 7 | 33 | | | | ▨ | ▨ | ▨ |
| 8 | 32 | ▨ | ▨ | ▨ | | | ▨ |
| 9 | 21 | | ▨ | | ▨ | ▨ | |
| 10 | 39 | ▨ | | | | ▨ | ▨ |
| 11 | 46 | | | ▨ | | ▨ | ▨ |
| 12 | 53 | | | | ▨ | | ▨ |
| 13 | 24 | | ▨ | | ▨ | | |
| 14 | 23 | ▨ | ▨ | | | ▨ | |
| 15 | 35 | | | ▨ | ▨ | ▨ | ▨ |
| 16 | 28 | | ▨ | | | | ▨ |
| 17 | 53 | ▨ | | ▨ | ▨ | | ▨ |
| 18 | 27 | | ▨ | ▨ | | ▨ | ▨ |
| 19 | 25 | ▨ | ▨ | | ▨ | | ▨ |
| 20 | 34 | ▨ | | ▨ | ▨ | ▨ | |
| 21 | 21 | ▨ | ▨ | | | | ▨ |
| 22 | 37 | ▨ | ▨ | | ▨ | ▨ | ▨ |
| 23 | 31 | | ▨ | ▨ | | ▨ | |
| 24 | 11 | | ▨ | | | ▨ | ▨ |
| 25 | 23 | | ▨ | ▨ | ▨ | | ▨ |
| 26 | 29 | ▨ | ▨ | ▨ | ▨ | ▨ | |
| 27 | 16 | ▨ | ▨ | | ▨ | ▨ | |
| 28 | 13 | | | | | | ▨ |
| 29 | 15 | | ▨ | ▨ | ▨ | ▨ | ▨ |
| 30 | 11 | | ▨ | | ▨ | ▨ | ▨ |