

**Isolation and characterization of the cellulose  
synthase promoters of *Eucalyptus* trees**

by

**Nicky Creux**

Submitted in partial fulfillment of the requirements for the degree

***Magister Scientiae***

In the Faculty of Natural and Agricultural Sciences

Department of Genetics

University of Pretoria

Pretoria

November 2006

Under the Supervision of Prof. Alexander A. Myburg and Co-supervision  
of Prof. Dave K. Berger



## Declaration

I, the undersigned, hereby declare that the dissertation submitted herewith for the Degree M.Sc. to the University of Pretoria, contains my own independent work and has not been submitted for any degree at any other university.

A handwritten signature in black ink, appearing to read 'Nicky Creux', written over a horizontal line.

**Nicky Creux**

**November 2006**

## Preface

Cellulose is produced by all plant species and is the most abundant biopolymer in the world. The cellulose fibers produced in plants are used by many industries to produce valuable products such as paper, fabric and food additives. Fast growing, plantation tree species such as *Eucalyptus* and *Populus* are a major source of industrial cellulose. A more complete understanding of the cellulose biosynthetic pathway will be highly beneficial to industries that use cellulose fibers as a raw material. Cellulose synthase (*CesA*) genes have been identified in a number of plant species and can be split into two distinct groups, those associated with primary cell wall formation and those associated with secondary cell wall formation. Despite a growing number of studies on the molecular events underlying cellulose biosynthesis, there are no detailed studies on the transcriptional regulation of the genes involved in the cellulose biosynthetic pathway. The identification of cis-regulatory elements that regulate the *CesA* genes will also be useful for genetic modification and the construction of synthetic promoters to confer highly specific gene expression to transgenes. The **Aim** of this M.Sc. research project was to investigate the transcriptional regulation of different members of the *CesA* gene family in *Eucalyptus*. In order to achieve the aim six cellulose synthase promoter regions were isolated from *Eucalyptus grandis*. The promoter regions were comparatively analysed with the orthologous regions in *Arabidopsis* and *Populus* using bioinformatics tools to identify putative regulatory motifs that play a role in the *CesA* genes expression patterns.

**Chapter 1** of this dissertation is comprised of a brief review of the literature on the analysis and application of promoter sequences involved in wood formation. This review focuses on previously identified cis-regulatory elements of genes involved in

wood formation and the tools available for the *in silico* analysis of plant promoter regions.

Cellulose is produced by a complex of membrane bound enzymes, which deposit the cellulose on the outside of the plasma membrane. The cellulose synthase complex is made up of six catalytic subunits embedded in the membrane in a rosette-like structure. Each catalytic subunit is comprised of a number of cellulose synthase (CESA) proteins, which are encoded by different *CesA* genes. We have recently cloned seven cellulose synthase genes from *Eucalyptus grandis* (Ranik and Myburg 2006). **Chapter 2** of this dissertation describes the isolation and cloning of the promoter regions of six *CesA* genes from *Eucalyptus grandis*. This chapter includes the preliminary identification of core promoter elements and the predicted transcriptional start sites (TSS).

**Chapter 3** discusses the results of a comparative bioinformatics study of the orthologous *CesA* promoters from *Eucalyptus*, *Populus* and *Arabidopsis*. This chapter identifies a number of motifs that may play a role in *CesA* gene regulation. This section also discusses putative functions of the motifs identified some of which are conserved among the different species.

At the end of the dissertation a brief concluding remark is provided in a section titled **Concluding Remarks** which the results of the dissertation are put into perspective and conclusions drawn on the value of this study on both an academic and industrial level.

The findings presented in this M.Sc. dissertation represent the outcomes of a study undertaken from March 2004 to October 2006 in the Department of Genetics, University of Pretoria, under the supervision of Prof. A.A. Myburg and Prof. D.K. Berger. Chapters 2 and 3 have been prepared in the format of independent manuscripts for peer reviewed research journals. A certain degree of redundancy may therefore exist between the introductory sections of these chapters and Chapter 1. Although the chapters have been prepared in the format of journal manuscripts, more supporting data are included in the thesis chapters than would normally be included in a manuscript for a research journal. To submit the results for publication it is likely Chapter 2 and 3 will be combined. The preliminary results of this study have not yet been published or presented in any form as sections of this work form part of a provisional patent filed in June of 2006.

## Acknowledgements

I would like to give many thanks to the following people, organizations and institutes for assisting me in the completion of this study:

- To Prof. Zander Myburg for his dependable and creative leadership of this project and for adding great insight and thorough reviews with an unsurpassable work ethic.
- To Prof. Dave Berger for excellent advice, insights, guidance and the comprehensive reviewing of this dissertation.
- To Martin Ranik for his help in the isolation of the promoters of *EgCesA2* and *EgCesA7* and without his previous work in identifying the *EgCesA* genes this work would not be possible. Also for a very high standard of work for which to strive for.
- To Joanne Bradfield for all the time and effort she spent on the isolation of *EgCesA7* and her help in the genome walking, cloning and sequencing of *EgCesA2* and *EgCesA 7* promoters.
- To all the past and present member of the Forest Molecular Genetics Laboratory: Elna Cowley (Lab mom), Adrene Laubscher, Dr Yoseph Beyene, Dr Solomon Fekybelu, Minique de Castro (Ouma), Honghai Zhou, Kitt Payn, Frank Maleka, Michelle Victor, John Kemp, Luke Solomon, Marja O'Neil, Grant McNair, Alisa Postma, Mmoledi Mphalele, Tracey-Leigh Hatherell and Eshchar Mizrachi for maintaining a stimulating and productive environment in which to study and work.
- To my family for all the emotional and financial support, and for always believing in me.

- To my friends (especially Ouma, Vinet, Bronwyn and Jana) for always being ready to listen even when they didn't understand and to Juan Jager for all his patience, understanding and many weekend lifts to varsity.
- To the genetics department of the University of Pretoria and the Forestry and Agricultural Biotechnology Institute (FABI), for providing a sound academic environment and exposure to all aspects of science.
- To the sequencing facility at the University of Pretoria, Renate Zipfel, Gladys Shabangu, Mia Bolton, for fast and efficient service, dedication and determination.
- Mondi Business Paper South Africa, for supplying the plant materials used in the study, excellent support and funding contributions
- To the National Research foundation of South Africa (NRF), for funding supplied by the grant holder-linked scholarship
- To The Human Resources and Technology for Industry Programme (THRIP), for financial support of the research.

# Table of Contents

<b>DECLARATION</b> .....	<b>ii</b>
<b>PREFACE</b> .....	<b>iii</b>
<b>ACKNOWLEDGEMENTS</b> .....	<b>vi</b>
<b>TABLE OF CONTENTS</b> .....	<b>viii</b>
<b>CHAPTER 1</b> .....	<b>1</b>
LITERATURE REVIEW .....	1
ANALYSIS AND APPLICATION OF PROMOTER SEQUENCES INVOLVED IN WOOD FORMATION .....	1
1.1 Introduction .....	2
1.2. Xylogenesis .....	4
1.3. Plant gene regulation .....	13
1.4. Promoter analysis.....	23
1.5. Conclusion .....	34
1.6 Aim of the study .....	35
1.7 References.....	37
<b>CHAPTER 2</b> .....	<b>51</b>
ISOLATION AND SEQUENCE CHARACTERIZATION OF PROMOTER REGIONS OF SIX CELLULOSE SYNTHASE GENES IN <i>EUCALYPTUS GRANDIS</i> .....	51
2.1 Abstract.....	52
2.2 Introduction .....	53
2.3 Materials and Methods .....	57
2.4 Results.....	62
2.5 Discussion.....	68
2.6 Acknowledgments .....	76
2.7 Tables.....	77
2.8 Figures.....	84
2.9 References.....	92
<b>CHAPTER 3</b> .....	<b>98</b>
<i>IN SILICO</i> ANALYSIS OF CIS-ACTING ELEMENTS IN THE CELLULOSE SYNTHASE PROMOTERS OF <i>EUCALYPTUS</i> , <i>POPULUS</i> AND <i>ARABIDOPSIS</i> .....	98
3.1 Abstract.....	99
3.2 Introduction .....	100
3.3 Materials and Methods .....	104
3.4 Results.....	110
3.5 Discussion.....	120
3.6 Acknowledgements.....	135
3.7 Tables.....	137
3.8 Figures.....	151
3.9 References.....	158
<b>CONCLUDING REMARKS</b> .....	<b>163</b>
References.....	169
<b>SUMMARY</b> .....	<b>171</b>
<b>APPENDIX 1</b> .....	<b>174</b>
CELLULOSE SYNTHASE PROMOTER SEQUENCES .....	174
<b>APPENDIX 2</b> .....	<b>182</b>
PRIMARY AND SECONDARY ASSOCIATED CESA MOTIF DATASHEETS .....	182
Motifs identified in more than one dataset.....	186



Motifs identified in CesA set 1.....	193
Motifs identified in CesA set 2.....	210



# **CHAPTER 1**

## **LITERATURE REVIEW**

### **ANALYSIS AND APPLICATION OF PROMOTER**

### **SEQUENCES INVOLVED IN WOOD**

### **FORMATION**

## 1.1 Introduction

A number of important developmental processes occur during the life cycle of a plant, from seed germination through root, stem and leaf development to flower formation and seed production. The processes of vascular development and xylogenesis (wood formation) are of great importance as they are the mechanisms by which aerial plants acquire mechanical strength and the ability to transport water and nutrients. On a molecular level vascular development is controlled by a large network of genes expressed in specific spatial or temporal patterns. Many of the key structural genes have been identified and certain processes of wood formation such as lignin deposition have been extensively described. However the transcriptional regulation of the genes involved in vascular development and xylogenesis has not yet been well characterized. Greater insight into the regulatory mechanisms that underlie developmental processes such as wood formation will enhance our understanding of plant development.

The process of xylogenesis is characterized by four main events: cell division from the cambial initials, cell enlargement and differentiation, secondary cell wall deposition and finally, programmed cell death. Fibers and vessels are the main products of xylogenesis. The vessels are long hollow cells that join together to form continuous channels that extend through the length of the plant stem and carry water from the roots to the shoots. A network of many genes governs each of the four steps in xylogenesis. Each gene is in-turn controlled by its own promoter and a set of transcription factors, but the information on these regulatory mechanisms is very limited and requires more in-depth studies.

RNA polymerase II promoters control the transcription of protein-coding genes such as the genes involved in wood formation (Sawant et al. 2005). These promoters can have many different forms, but in general they can be divided into the proximal and distal regions. The proximal promoter is the region to which the transcriptional initiation complex binds and the distal region where the transcription factor binding sites are located. Transcription is also controlled by the methylation of the DNA and its chromatin structure.

With the ever-increasing knowledge on the different promoters a number of algorithms have been created in order to identify and analyse promoters via faster *in silico* methods. Because of the complexity of gene transcription, no *in silico* promoter prediction tool will be 100% accurate, although the accuracy has been increased by a number of different advancements (Tompa et al. 2005). The genome sequences of *Arabidopsis*, rice and poplar have helped in forming a more accurate plant promoter model and this has increased the accuracy of *in silico* predictions greatly. The number of available tools has also increased the accuracy, as now it is possible to use different bioinformatics tools and then compare their results (Rombauts et al. 2003; Tompa et al. 2005). The aim of this review is to briefly summarize what is known about wood formation (Gunnerås 2005) focusing on the regulation of the key genes involved. Transcription factors that have been identified by different studies that indicate their role in wood formation (Torres-Schumann et al. 1996; Sessa et al. 1998) and the software packages that can be used for the *in silico* analysis of plant promoters are also discussed here.

## **1.2. Xylogenesis**

Meristems are points of growth within the plant that are located in plant organs such as the shoot tips, root tips and stems. The lateral meristem found in the stem is also called the vascular cambium and it is involved in the thickening of woody plant stems by the production of the secondary vascular tissues. The vascular cambium is a thin layer of cells that differentiates to form the vascular tissues in the tree stem and makes perennial growth of the tree possible by replacing the xylem and phloem regularly so they are always in an optimal working condition (Plomion et al. 2001). Within the stem the cambium divides to the outside to produce phloem initials. The cambium also divides towards the inside of the stem to produce cells that differentiate into tracheary elements. These elements carry water from the roots of the tree to the shoots and form part of the xylem tissue, which constitutes wood (Samuels et al. 2006). This process of tracheary element formation is known as xylogenesis. Xylogenesis proceeds through four phases: cell division from the cambial initials, elongation and differentiation of the xylem cells, strengthening of the cell walls and finally programmed cell death (Roberts and McCann 2000). This process will be described with an emphasis on the regulatory genes and mechanisms that have been reported to play a role in xylogenesis (Cosgrove 2005; Groover 2005; Hughes 2006).

### **1.2.1. Cell division from cambial initials**

During xylogenesis the vascular cambium maintains a few layers of undifferentiated cells during cell division and differentiation. As soon as the cambium cells divide, differentiation begins and immature phloem or xylem cells are produced (Mellerowicz et al. 2001). The vascular cambium is formed from the procambium, which is originally derived from the shoot apical meristem (SAM) during the development of

the primary plant body. A number of studies have investigated the genes that play a role in SAM maintenance and cell fate. It was found that two transcription factors, a Class III HD-ZIP protein and *KANADI*, were involved in the spatial development of the lateral organs of *Arabidopsis* (Bowman et al. 2002) and these same transcription factors have been shown to play a similar role in the radial patterning of the *Arabidopsis* vascular cambium (Emery et al. 2003). An expression study performed on poplar cambium revealed that the poplar homologs to Class III HD-Zip and *KANADI* were up regulated during wood formation, with the *Class III HD-Zip* gene expression increasing on the phloem side of the cambium while the *KANADI* gene expression increased on the xylem side. These results suggest that *KANADI* is involved in the differentiation of cambium cells to xylem cells while the Class III HD-Zip proteins play a role in the differentiation of the cambium cells to phloem cells (Schrader et al. 2004).

There are a number of genes involved in the maintenance of the shoot apical meristem and the differentiation of the cells to form phloem and xylem. Two transcription factors belonging to the KNOX family, SHOOTMERISTEMLESS (*STM*) and BREVIPEDICELLUS (*BP*), regulate cell fate in the shoot apical meristem of *Arabidopsis* (Long et al. 1996; Mele et al. 2003). The *STM* and *BP* genes in *Arabidopsis* (Ko et al. 2004) and their orthologs in poplar (Schrader et al. 2004) were up regulated during secondary growth also suggesting a role in cambium cell fate (Groover 2005). It has been suggested that while *STM* and *BP* maintain the undifferentiated cambium cells, cell-to-cell communication rather than direct lineage is required for cell differentiation (Bannan 1957). The *CLAVATA* family of genes play an integral role in mediating the cell-to-cell communication via receptor ligand

interactions, but this mechanism will not be discussed in detail here (for detailed review see: Groover 2005).

### 1.2.2. Cell differentiation and elongation

Once cell division has occurred, the cells allocated for xylem cell fate will begin to differentiate and expand both longitudinally and radially in order to form the tracheary elements. These xylem mother cells will differentiate into fusiform and ray initials and these further differentiate into vessel elements, fibers and ray parenchyma cells. Each of these steps in the cell differentiation process will be under the control of a regulatory gene network. There will be genes involved in the differentiation to xylem cells such as *NST1* and *NST2* (*NAC SECONDARY WALL THICKENING PROMOTING FACTOR 1/2*) involved in proto- and metaxylem formation (Mitsuda et al. 2005). Another gene involved in differentiation to xylem is *REVOLUTA/INTERFASCICULAR FIBERLESS1* (*REV/IFL1*), which is a member of the class III HD-ZIP. If this gene is mutated it disrupts the formation of interfascicular fibers that normally provide support to the stem of the plant thus weakening these structures (Zhong et al. 1997).

The differentiating plant cells undergo cell elongation but plant cell walls are a hindrance in this process. The cell wall polymers in the primary cell wall must be loosened by specialised proteins such as expansins. Expansins disrupt the hydrogen bonds formed between the cellulose microfibrils and their interlinking glycans, as well as the covalent bonds within the cross-linking glycans, therefore allowing the cell wall to extend (McQueen-Mason and Cosgrove 1995). Expansins are produced by a large gene family, which has been divided into  $\alpha$ -expansins and  $\beta$ -expansins

(Cosgrove et al. 2002). The  $\alpha$ -expansins are of interest as they have been found to play a role in plant cell wall development, while the  $\beta$ -expansins are used during pollen tube growth. Twenty-two expansin genes have been identified in the *Arabidopsis* genome (Cosgrove 1998), but it has been found that there is little or no redundancy and that these expansins mostly have differential expression patterns suggesting they have many different regulatory mechanisms (Rose et al. 1997). In the poplar cDNA library produced by Sterky et al. (1998) a large number of expansin ESTs were identified, indicating a role for expansins in wood forming tissues. Recently a membrane steroid binding protein (MSBP1) was found to negatively regulate hypocotyl elongation in *Arabidopsis* and in knock out mutants the expression levels of expansins were altered (Yang et al. 2005). The regulation of these genes is of interest as fibre length is an important wood characteristic and understanding the control of these genes could be essential for manipulating fibre length and maximizing secondary growth.

### **1.2.3. Cell wall biosynthesis**

#### *Cellulose biosynthesis*

After cell differentiation and cell elongation have occurred, the secondary cell wall is deposited in order to strengthen the cell. Strengthening of the cell wall is achieved by the deposition of additional cellulose, hemicellulose and lignin. During the deposition of cellulose there are a number of cell wall textures that can be formed depending on the angles at which the cellulose microfibrils are deposited (Emons and Mulder 2000). The microfibril angles will give the xylem different properties such as more strength or elasticity and thus is important for the pulp and timber industries. Lignin deposited



in xylem cells provides support to the cell walls and has the secondary function of providing waterproofing to the cell walls.

The model for cellulose biosynthesis is as follows: Each individual cellulose chain is synthesised by cellulose synthase (CESA) proteins embedded in the cell membrane. The CESA proteins associate to form large cellulose synthase complexes, arranged into rosette shaped structures with a six by six symmetry (i.e. 36 CESA proteins arranged into six complexes of six CESA proteins each). Each rosette therefore synthesises 36 cellulose chains that associate through H-bonding into para-crystalline microfibrils (Brown and Saxena 2000). Joshi et al. (2004) state that there are three main steps to the production of cellulose within the plant cell wall. The first step occurs with a protein known as sucrose synthase (SuSy), which, channels UDP-glucose to the CESA protein complex. During the second step, the CESA proteins polymerise the glucose to form glucan chains. The final step of this process is managed by a membrane bound enzyme known as KORRIGAN, which edits the glucan chains and monitors their convergence during the production of microfibrils.

CESA proteins are encoded by a family of cellulose synthase (*CesA*) genes. *CesA* genes have been isolated from different plant species. Eight *CesA* genes have been isolated from barley (Burton et al. 2004) and seven have been identified in *Eucalyptus* (Ranik and Myburg 2006). The *Arabidopsis* genome contains 10 different cellulose synthase genes, while rice genome has 12 (Richmond and Somerville 2000: <http://cellwall.stanford.edu/>) *CesA* genes and 18 *CesA* genes have been identified in the poplar genome (Djerbi et al. 2004). There is a strong relationship that has been conserved among these genes and their expression patterns. In *Arabidopsis* three genes play a role in cellulose production during the primary cell wall formation and

these are *AtCesA1*, 3 and 6 (Burn et al. 2002; Taylor et al. 2003). Three other genes, *AtCesA4*, 7 and 8, are involved in secondary cell wall formation (Desprez et al. 2002). The functions of the *CesA* genes were obtained by studying various *Arabidopsis* mutant phenotypes (Fagard et al. 2000). A mutation in *AtCesA1* produced plants with reduced cellulose content and disrupted growth in all organs. This stunted phenotype was referred to as *rsw1* (Radial Swelling) (Arioli et al. 1998). The mutation in *AtCesA6* caused the plants to have reduced cellulose, reduced cell elongation and collapsed xylem and these mutants are known as *prc1* (Procuste) mutants (Desprez et al. 2002). The same phenotype was produced when a point mutation occurs in *korrigan* (*kor*) (Desprez et al. 2002), although not one of the *CesA* family members, *kor* is a key player in cellulose production as discussed above. The *kor* mutant phenotype is known as *irregular xylem2* (*irx2*) and has now been used in a number of different cell wall studies (Szyjanowicz et al. 2004). Other *irregular xylem* mutants have also been identified, *irx1* (Taylor et al. 2000), *irx3* (Taylor et al. 1999) and *irx5* (Turner and Somerville 1997; Taylor et al. 2003), which are due to mutations in *AtCesA7*, 4 and 8. The phenotypes of these mutants were similar to *irx2* also displaying reduced cellulose and collapsed xylem. The homologue genes for *AtCesA7*, 4 and 8 have recently been identified in poplar (Samuga and Joshi 2002) and *Eucalyptus* (Ranik and Myburg 2006) further enforcing their role in secondary wall formation (Taylor et al. 2000).

Described above is the current molecular knowledge on cellulose biosynthesis and from this discussion it clear that while there is some information on the key genes and their proteins, there is little information on the transcriptional regulation of this process. There are no publicly available references to transcription factors directly

involved in cellulose biosynthesis in plants or to the binding sites of these transcription factors in promoters of *CesA* genes. A large number of expression studies have identified transcription factor genes that are co-expressed with the *CesA* genes, but there is no direct evidence of how, or if they play a role in the regulation of the *CesA* genes

### *Lignin biosynthesis*

Another important process to consider when discussing secondary cell wall formation is the production and deposition of lignin. Lignin is a phenolic polymer deposited during cell wall formation. In secondary cell wall formation, lignin plays two roles, as mentioned previously, that of waterproofing the xylem vessels and offering strength and support to the plant stem. Unlike cellulose, lignin is not made of one type of monomer but in fact has three main monomers. The three units of lignin are hydroxyphenyl (H), guaiacyl (G) and syringyl (S), which are, produced from three different precursors namely p-coumaryl, coniferyl and sinapyl alcohols (Mellerowicz et al. 2001). Genetic engineering of the lignin biosynthetic pathway has been a main focus of forest biotechnology because it is costly and difficult to remove lignin from the pulp during paper production. If lignin is not properly removed from the pulp it oxidizes over time, which discolours the paper and decreases the quality (Grima-Pettenati and Goffner 1999). A number of groups have manipulated genes in the lignin biosynthetic pathway to alter the production of lignin during secondary cell wall formation (Hu et al. 1998; Harding et al. 2002; Li et al. 2003; Goicoechea et al. 2005).

The substantial amount of knowledge on the lignin biosynthetic pathway and the genes involved has led to the isolation of the promoters of a number of lignin biosynthetic genes such as *phenylalanine ammonia-lyase (PAL)*, *Cinnamoyl-CoA Reductase (CCR)*, *Cinnamyl alcohol dehydrogenase (CAD)* and *Hydroxycinnamate-CoA-ligase (4CL)* (Hauffe et al. 1991; Leyva et al. 1992; Lacombe et al. 2000; Lauvergeat et al. 2002). The isolation of these promoter regions has led to the identification of transcription factor binding sites that play likely a role in the regulation of lignin deposition. A pine O-methyltransferase promoter was found to have binding sites for different transcription factors such as LIM, MYB (MYeloBlastosis virus) and bZIP (Moyle et al. 2002). The MYB and bZIP transcription factors will be discussed later in this review. LIM (Named after the genes first identified: *Lin-11*, *Isl-1* and *Mec-3* genes) is a transcription factor, first isolated from *Nicotiana tabacum*, that was found to bind to the AC-element identified in a number of lignin genes and drives the xylem-specific expression of these genes (Kawaoka et al. 2000; Kawaoka and Ebinuma 2001). Despite more than a decade of gene regulation research in lignin biosynthesis, much remains to be learned about the transcriptional regulation of this pathway and its co-ordination with other carbon-fixing pathways such as cellulose and hemi-cellulose biosynthesis.

#### **1.2.4. Programmed Cell Death**

The last step in tracheary element formation is programmed cell death. This process has been well documented (Roberts and McCann 2000) and has a number of interesting characteristics. One of the most interesting characteristics is the degradation of the nucleic acids. It has been hypothesised that secondary cell wall synthesis is coordinated with the triggering of vacuole collapse and finally cell death.

The model states that as secondary cell wall formation comes to an end, a protease is secreted into the extra cellular matrix. The protease may alter ligands involved in the  $\text{Ca}^{+2}$  uptake or may release a molecule from the extracellular matrix that activates the  $\text{Ca}^{+2}$  channels. The large uptake of calcium causes the vacuole to rupture, which results in cell death (Groover and Jones 1999). The influx of calcium and the production of proteases will be regulated by a number of genes of which few have been identified. The *Zinnia elegans* transdifferentiation model (Fukuda et al. 1980) has been useful in the identification of genes involved in this process. With this model, ZEN1 was identified to be one of the main enzymes involved in the degradation of DNA during programmed cell death (Fukuda 2000). Two xylem cysteine proteases (XCP1 and XCP2) were also identified in *Arabidopsis* xylem cDNA libraries and play a role in secondary xylem formation in *Arabidopsis* (Zhao et al. 2000).

The genetic regulation of programmed cell death in plants is poorly characterised and unlike the wealth of information in animals, few regulators of this process have been identified in plants (van Doorn 2005). The sequencing of the *Arabidopsis* genome and improvements in bioinformatics tools has aided the identification of plant programmed cell death regulators. A recent study identified a new family of regulators in *Arabidopsis*, IAP- (inhibition of apoptosis) -like proteins (Higashi et al. 2005). IAPs were first identified in the Baculovirus genome and the proteins were found to inhibit apoptosis in humans (Deveraux and Reed 1999; Miller 1999). It has also been found that baculovirus IAP can inhibit some plant caspases (Ciacci-Zanella and Jones 1999). These findings suggest that ILPs in plant (IAP-like proteins) fulfil a similar role as IAPs in animals. Recently *Arabidopsis* BAG (*Bcl2* Associated

*athanoGene*) genes were identified and play a role in the regulation of programmed cell death. It was found that *Ca<sup>2+</sup> induced AtBag5 and 6* and may therefore play a role during cell death (Doukhanina et al. 2006).

### **1.3. Plant gene regulation**

#### **1.3.1. General promoter structure**

After reviewing the process of xylogenesis and the genes involved in this process, it is clear that there is still much to be learned about the transcriptional regulation of the genes involved in this process. All genes, including wood forming genes, are under the control of a transcriptional promoter and a suite of transcription factors that bind to the promoter. The general mechanism of transcription has been well documented in eukaryotes (Sawant et al. 2005), but the regulatory mechanisms that govern spatial or temporal gene expression are far from well understood. There are a number of similarities in the promoters of eukaryotic and prokaryotic organisms such as the fact that they all require an initiation start site and specific binding sites for the proteins that bind to form the transcription initiation complex. However, there are several differences between eukaryotes and prokaryotes such as the fact that eukaryote promoters contain more protein binding sites for the initiation of transcription than prokaryote promoters. Also, eukaryotes have a much longer region upstream of the initiation site in which the protein binding sites can be located as compared to prokaryotes (Kanhare and Bansal 2005). Even though there are more similarities between plant and animal promoters there are still a number of differences between these two groups such as the novel motif sequences identified in a genome wide-analysis of *Arabidopsis* promoters that have not been identified in mammals to date (Molina and Grotewold 2005). Even within an organism there are a variety of promoters such as the different promoters to which, different polymerases bind. These

polymerases are: RNA polymerase I which transcribes ribosomal RNA, RNA polymerase II involved in the transcription of all protein-coding genes and nuclear RNAs and RNA polymerase III that mainly transcribes transfer RNA, but also transcribes a few ribosomal and nuclear RNAs. Each of these processes has been widely studied and reviewed elsewhere (Wolffe 1995; Moss 2004). All protein coding genes including those involved in xylogenesis contain a promoter for RNA polymerase II. This type of promoter has been the focus of many studies and there are numerous reviews on the topic (Smale 1997; Hochheimer and Tjian 2003; Best et al. 2004; Svejstrup 2004; Sawant et al. 2005).

The current textbook (Latchman 1999) model for a RNA polymerase II promoter is that the promoter can be divided in to two parts the distal and proximal regions. The proximal region is the region just upstream of the 5'UTR that contains the transcriptional start site and is the region of DNA were the RNA polymerase complex is assembled. The proximal promoter is also referred to as the core promoter and contains all the necessary motifs to confer basal gene expression (Nikolov et al. 1996; Featherstone 2002). This region acts as the “on/off switch” of the gene. The distal region of the promoter contains cis-regulatory elements, enhancers and repressors that fine-tune the expression by regulating the genes spatial and temporal expression (Tjian and Maniatis 1994; Fessele et al. 2002). This region of the promoter has an undefined length and can stretch from a few hundred base pairs upstream to kilo base pairs upstream of the proximal promoter (Rombauts et al. 2003).

The regulatory sequences, themselves and their positioning in the promoter add complexity to the process of transcription in a number of ways. For example the

TATA-box is an AT-rich region which is expected to be approximately 50 bp upstream of the transcriptional start site at which the initiation complex assembles, but this may vary greatly from species to species and even from gene to gene in an individual (Lynch et al. 2005). The TATA-box is bound by TATA-box binding proteins, which then facilitate the binding of the RNA polymerase and initiation of transcription. In some cases, RNA polymerase II promoters do not have a TATA-box at all and rely on an initiator sequence (Inr) for the initiation of transcription (Smale 1997). The Inr sequence (Py-Py-A-N-(T/A)-Py-Py) is one of the only known alternative transcriptional initiation sites and has been found to function in the same way as the TATA-box (Lynch et al. 2005).

Core promoter cis-regulatory elements are not only found upstream of the transcriptional start site. In *Drosophila* a downstream promoter element (DPE), identified downstream of the transcriptional start site was found to play a core role in the transcription of genes containing an Inr and no TATA-box. The distance between the Inr and DPE played a role in the functioning of these promoters and when this distance was altered by even 1 bp the promoter activity would decrease (Kutach and Kadonaga 2000). Cis-regulatory elements such as enhancer and repressor elements have also been found in the 5'UTR (Mingam et al. 2004), 3'UTR and even introns (Loke et al. 2005). The huge variation of distance between cis-elements that play a role in gene regulation indicate that there are at least two basic forms of regulation (i) sequence-specific regulation where proteins bind to specific sequences in the DNA; and (ii) structure-specific regulation where the chromatin is remodeled to bring certain elements in closer contact or where regions of DNA are methylated to stop transcription factors from binding to the DNA (Rombauts et al. 2003).



### *Sequence dependent regulation*

From the discussion above it is clear that the initiation and regulation of transcription is reliant on the presence of transcription factor binding sites. Transcription factors are proteins that bind to a specific motif sequence within the distal and proximal regions of the promoter of the gene, which they regulate. These binding sites have a short (6-8 bp) core sequence that is very specific and the presence or absence of the binding sites can confer tissue or signal response specificity. The binding sites can form part of a larger consensus sequence, which is more variable (Pabo 1992). These protein-DNA interactions can be simple where a single protein binds to the DNA and affects the transcription of the gene. The protein could also compete with other proteins for the binding site and in this way alter the expression of the gene. In some cases transcription factors can play a dual role, actively repressing transcription by binding to a specific sequence and passively compete for the same binding site as activating transcription factors (Vom Endt et al. 2002).

In some cases two or more transcription factors must be present at the same time in order to produce a particular expression pattern. For example, MYB transcription factors form a heterodimer with a bHLH (Basic-helix-loop-helix) transcription factor (Murre et al. 1989; Grima-Pettenati and Goffner 1999) in order to initiate transcription in the flavonoid biosynthetic pathway of a number of plants (Vom Endt et al. 2002). An even more complex system was described in *Arabidopsis* root epidermis development where a cascade of different transcription factors was required for this development. CAPRICE is a R3 MYB that regulates root hair cell differentiation in *Arabidopsis* and is regulated by a second transcription factor WEREWOLF (WER). WER is a R2R3 MYB and regulates an HD-Zip gene

(Koshino-Kimura et al. 2005), which causes differentiation to hairless cells in the root when part of a heterodimer with a bHLH transcription factor (Lee and Schiefelbein 2002). For this reason, a number of transcription factor binding sites can often be located in a promoter as it is the combination of proteins binding that confers the specific expression pattern.

#### *Structure dependent regulation*

It is reasonable to think that proteins binding to DNA could regulate gene expression, but how does an element kilobases away affect expression? In order to understand this it is important to remember that DNA interacts with proteins in a three-dimensional space and that regions that are hundreds of kilobases apart “in cis” can be in close contact due to the folding of the DNA. This folding can also cause some regions of DNA to be more accessible to proteins than other regions.

Methylation of intergenic DNA also has effects on gene expression. Many promoters have been found to be associated with regions of DNA known as CpG and CpNpG islands (Zhang 1998). These are regions of DNA that have a high GC content and their methylation may indicate functionality. Because the unmethylated CpG and CpNpG regions have been found to be closely associated with active gene promoters, computer programs such as McPromoter (Ohler et al. 1999; Ohler and Niemann 2001) and Eponine (Down and Hubbard 2002) have been produced to scan for the CpG and CpNpG islands in the genome and predict promoters in these regions. These programs and method were well described in the review by Rombauts et al. (2003).

The genes involved in transcriptional gene silencing by altering DNA methylation can be divided into two main groups. The first group is comprised of genes that affect DNA methylation at the whole genome level. The second group of genes affects DNA methylation only in specific regions of the genome. The methylation of CpG and CpNpG islands can repress the genes by the binding of methyl – CpG – binding proteins, that bind to the DNA and hinder the binding of transcription factors and so transcription is repressed (Kass et al. 1997a; Kass et al. 1997b). Recently, 12 such proteins have been identified in *Arabidopsis* and three of these were found to specifically bind to CpG methylated regions and mediate CpG methylation (Zemach and Grafi 2003). In other studies, these proteins have been found to act as structural proteins which recruit a variety of histone deacetylase complexes and chromatin remodeling factors, leading to chromatin compaction and finally to transcriptional repression (Ballestar and Wolffe 2001; Wade 2001a; Wade 2001b).

It has recently been shown that regions of repeat DNA (e.g. microsatellites) in promoters can play a role in gene regulation (Martin et al. 2005). There is very little information on this regulatory mechanism and it is unclear whether these repeat regions bind transcription factors or play some spatial regulatory role. Microsatellites have been found throughout the genomes of many organisms including plants. Most of these organisms have been shown to contain microsatellites in their promoter regions. It is not clear what role these motifs play in gene regulation. The plantCARE database (<http://bioinformatics.psb.ugent.be/webtools/plantcare/html>) lists a CT repeat sequence as a general enhancer of gene expression found in the 5'UTR of some tomato genes (Daraselia et al. 1996). Microsatellites have been found to be present at a higher proportion than expected in introns and in noncoding sequences upstream

and downstream of the genes (Casacuberta et al. 2000). The upstream noncoding DNA regions also contain a high proportion of regulatory motifs and perhaps this indicates a regulatory function for microsatellites in the same intergenic regions. A study in the bacteria *Neisseria meningitidis* showed that the instability of microsatellites could alter promoter activity (Martin et al. 2005). This could be an interesting avenue to explore in biotechnology as it could offer a way of finely regulating beneficial genes.

### **1.3.2. Xylem cis-regulatory motifs and transcription factors**

Gene regulation involves an intricate cascade of mechanisms and in order to better understand the expression of a gene one needs to understand the underlying mechanisms of gene regulation. To date only a few transcription factors and the binding sites (Table 1.1) have been identified as regulators of key genes involved in wood formation. There is still a great deal of work that must be done in order to gain a clearer understanding of the gene regulation during xylem formation.

The MYB proteins constitute a large family of plant specific transcription factors with 125 members identified in *Arabidopsis* (Stracke et al. 2001). Most plant MYB proteins have two helix-turn-helix motifs and it is these R2 and R3 repeats that are responsible for binding to the DNA. MYBs have a diverse range of functions that include seed development, germination, stress response, anther development, photomorphogenesis and lignin deposition. A pine MYB (PtMYB4) was found to play a role in lignification and the same study revealed that it activated transcription in an AC-element dependent fashion (Campbell and Sederoff 1996; Patzlaff et al. 2003). AC-elements (Table 1.1) are sets of consensus sequences that are found in

most of the genes involved in lignin biosynthesis (Table 1.1). They are found in genes such as *PAL* (Hatton et al. 1995), *4CL* and *C3H* and appear to play a role in the production and deposition of lignin in the vascular tissues (Raes et al. 2003). When an *Arabidopsis* MYB (*AtMYB61*) homologous to the pine MYB (*PtMYB4*) was mis-expressed in *Arabidopsis* it also produced ectopic lignification (Newman et al. 2004). This indicates that *PtMYB4* must play a key regulatory role in the deposition of lignin. Recently, MYBs isolated from *Eucalyptus* have been shown to play a role in the lignin biosynthetic pathway and secondary cell wall formation (Lauvergeat et al. 2002; Goicoechea et al. 2005).

The HD-ZIP III family of transcription factors all contain Homeodomain-leucine zipper (HD-ZIP III) domains and have been found to play a role in a number of plant developmental processes. In *Arabidopsis*, five members of this family were implicated in the regulation of vascular development (Ohashi-Ito and Fukuda 2003). The five members of the family are *IFL1/REV* (INTERFASCICULAR FIBERLESS1/REVOLUTA), *ATHB-8*, *ATHB-9*, *ATHB-14* and *ATHB 15*. An in depth study of the different family members revealed that *REV* loss of function mutants, showed defects in the shoot apical meristem, lateral organ patterning, vascular development and plant stature. The other four genes in the family only produced a mutant phenotype when they were knocked out in pairs to produce double mutants (Prigge et al. 2005). These results indicate that these proteins may work together in complexes were different combinations of proteins confer different expression patterns.

**Table 1.1 Some plant regulatory sequences previously identified in promoters of genes that play a role in vascular development.**

These regulatory element sequences were obtained from the PLACE (<http://www.dna.affrc.go.jp/htdocs/PLACE/>) database.

Element Name	Sequence	Species	Binding site Description	Transcription Factor Family	References
ACIII element	GTTAGGTTCC	<i>Phaseolus vulgaris</i>	Vascular-specific	MYB and G-box	Hatton et al. (1995)
ACII element	CCACCAACCCCC	<i>Phaseolus vulgaris</i>	Vascular-specific	MYB and G-box	Hatton et al. (1995)
ACI element	CCCACCTACC	<i>Phaseolus vulgaris</i>	Vascular-specific	MYB and G-box	Hatton et al. (1995)
RNFG box I	GATCATCGATC	<i>Oryza sativa</i>	Phloem-specific	bZIP	Yin and Beachy (1995)
RNFG box II	CCAGTGTGCCCTGG	<i>Oryza sativa</i>	Phloem-specific	bZIP	Yin and Beachy (1995)
GATA motif	CAGAAGATA	<i>Arabidopsis thaliana</i>	Phloem-specific	GATA motif binding factor	Yin et al. (1997)
GATA box	GATA	<i>Arabidopsis thaliana</i>	Light regulated tissue-specific	GATA motif binding factor	Lam and Chua (1989)
Dof binding site	ACTTTA	<i>Nicotiana tobacum</i>	Auxin induced tissue-specific	Dof protein	Baumann et al. (1999)
Oct type I element	CCACGTCANCGATCCG C	<i>Nicotiana tobacum</i>	S-Phase meristem-specific	-	Taoka et al. (1999)
Oct type II element	TCACGCGGATC	<i>Nicotiana tobacum</i>	S-Phase meristem-specific	-	Taoka et al. (1999)
Oct type III element	GATCCGCGNNNNNNNN NNNNNNNACCAATCS	<i>Nicotiana tobacum</i>	S-Phase meristem-specific	-	Taoka et al. (1999)
Stem element I	ATAATGGCCCACTGT GGGGGCAT	<i>Phaseolus vulgaris</i>	Stem-specific enhancer	-	Keller and Heierli (1994)
Stem element II	TTNNNGTAGCTAGTGTA TTTGTAT	<i>Phaseolus vulgaris</i>	Stem-specific	-	Elmayan and Tepfer (1995)
Negative regulatory region	TAGTGGAT	<i>Brassica napus</i>	Represses extensin in root	-	Elliott and Shirsat (1998)
Sugar responsive element	TTATCC	<i>Arabidopsis thaliana</i>	Auxiliary bud gene regulation	-	Tatematsu et al. (2005)
CCR binding site I	AGCGGG	<i>Eucalyptus grandis</i>	Vascular-specific	-	Lacombe et al. (2000)
Auxin response elements	TGTCTC	<i>Glycine max</i>	Auxin response	Auxin response factor	Ulmasov et al. (1999)
HDZIP binding site	GTAATSATTAC	<i>Arabidopsis thaliana</i>	Xylem cell differentiation	ATHB9	Sessa et al. (1998)
VSF-1 binding site	GCTCCGTTG	<i>Lycopersicon esculentum</i>	Xylem-specific	bZIP	Torres-Schumann et al. (1996)
ASL box	GCATCTTTACTTTAGCAT C	<i>Oryza sativa</i>	Phloem-specific	As-1 box binding factor	Yin et al. (1997)
AGAMOUS	TTAATGG	<i>Arabidopsis thaliana</i>	Quiescent centre-specific	WUSCHEL	Lohmann et al. (2001)
AGAMOUS	CCAATGT	<i>Arabidopsis thaliana</i>	Quiescent centre-specific	LEAFY	Lohmann et al. (2001)

Another family of transcription factors that also have a leucine zipper domain is the b-ZIP (basic leucine zipper motifs) family. The b-ZIP proteins bind to a G-box that was found in the promoters of *PAL* (Heinekamp et al. 2002) and F5H (ferulate 5-hydroxylase). The promoter-binding site (G-box) core contains the sequence CACGTG. This sequence is highly conserved and has been found in the promoter regions of genes regulated by environmental and physiological signals (Weisshaar and Jenkins 1998). Although the core sequence is highly conserved, it has been noted that if the sequence differs by one basepair a different b-ZIP family member can bind to the sequence (Williams and Neale 1992), which indicates that these proteins require a highly specific binding site in order to perform their function. B-ZIP proteins have been found to play a role in the formation of hetero- and homodimers, which form protein – DNA complexes (reviewed in Vinson et al. 2006) and for this reason the binding sites of b-ZIP proteins are often found in pairs.

The MADS-box family of transcription factors is also well conserved and is intriguing because, in spite of their conservation, they play roles in many different processes in the plant. The diverse functions of MADS-box proteins are due to their ability to form large protein complexes, where each complex can perform a different function (Theissen 2001). A review by Cseke et al. (2003) listed all the MADS-Box genes isolated thus far and out of 81 genes only seven are expressed in stems. The first to be identified was *AtAGL3* (AGAMOUS-Like), which was expressed in a variety of tissues in *Arabidopsis* (Ma et al. 1991). Cseke et al. (2003) discovered a MADS-box gene in poplar (*Populus tremuloides* MADS-box 5, *PTM5*) that is first expressed in primary vascular tissue, but later is predominantly expressed in mature cambium during the development of secondary vascular tissues as well as leaves and flowers.

This transcription factor appears to be expressed during spring wood formation and is co-expressed with a number of wood specific genes.

NACs are a large family of transcription factors with 110 identified in *Arabidopsis* (Mitsuda et al. 2005). NAC proteins all contain a NAC domain and are named after the first genes to be identified, *NO APICAL MERISTEM*, *ATAF-1*, *ATAF-2* and *CUP-SHAPED COTYLEDONS 2* (Aida et al. 1997). An interesting characteristic of these transcription factors is that they appear to be redundant so the knockdown of one of these genes does not produce the expected mutant phenotype. Mitsuda et al. (2005) identified two NAC genes; *NST1* (*NAC SECONDARY WALL THICKENING PROMOTING FACTOR1*) and *NST2* involved in pollen cell wall thickening that when over-expressed caused ectopic secondary cell wall formation. When *NST1* and *NST2* were knocked out they were found to be redundant and in a dominant repressor system as no mutant stem phenotype was identified. Mitsuda et al. (2005) speculated that since no stem abnormalities were observed in the *NST1* and *NST2* double mutants, there may be a third gene involved in the regulation of secondary cell wall formation. Two other genes from the same family (*VASCULAR-RELATED NAC-DOMAIN*, *VND6* and *7*) have been shown to play a role in the development of primary vascular tissue in *Arabidopsis* (Kubo et al. 2005).

## **1.4. Promoter analysis**

### **1.4.1. Promoter isolation**

At present there is a lack of knowledge regarding the transcription factors involved in wood formation and their interaction with the genes that they regulate. This may be due to the fact that in the past promoter isolation and analysis was a slow and laborious process where promoters were cloned, sequenced and functionally tested.



Now the genomes of a number of organisms have been sequenced and this has led to speedy discovery of putative promoter regions via computational methods. High-throughput transcript profiling techniques such as cDNA-AFLP and microarray analysis (Kuhn 2001) in combination with the sequenced genomes has allowed the comparative analysis of promoters of co-expressed genes. The comparative analysis is based on the assumption that, the promoters of co-expressed genes contain similar cis-acting regulatory elements, and this has greatly increased the accuracy of computational methods.

Genome walking has always been the first step when isolating and characterizing promoters. This technique is still used today for organisms with little available gene sequence information. Once a full-length gene has been isolated, the upstream region can be obtained via genome walking. In this method, genomic DNA is digested with specific restriction enzymes, adaptors are ligated and the upstream regions amplified using a two-step PCR (Iwahana et al. 1994; Siebert et al. 1995). A number of genome walks may be required before the promoter can be obtained, because genome-walking products and promoters can vary in size greatly. Although this technique may appear to be quite laborious the burden on researchers has been lifted by a number of commercially available kits. These kits have streamlined the process and made genome walking more accessible to the scientific community. With the development of genome walking the number of known promoter sequences isolated and cloned has increased with reports appearing on a regular basis on a variety of different species (Connors et al. 2002; Trindade et al. 2003; Trindade et al. 2004; Kim et al. 2005; Wu et al. 2006). Using this method a number of important promoters involved in plant development have been isolated and characterized (Table 1.2).

Plant promoter discovery has been enhanced by the recent sequencing of the *Arabidopsis* (*Arabidopsis* Genome Initiative 2000) and *Populus* (Tuskan et al. 2006) genomes. Now obtaining promoter sequences of these organisms is quite simple and has led to a great increase in the knowledge on the promoters of these organisms using bioinformatics tools. The *Arabidopsis* genome is far better annotated than the poplar genome because first draft of the poplar genome was only recently completed (Tuskan et al 2006). TAIR (The *Arabidopsis* Information Resource) (Huala et al. 2001) is a website that is focused on *Arabidopsis* and houses the entire genome sequence. TAIR has a user-friendly interface through which the user can navigate the genome.

**Table 1.2 Promoters in the PlantpromDB (<http://www.softberry.com/>) that may play a role in wood formation.** All of the promoters in the PlantpromDB have been functionally tested and characterised.

Organism	Gene identity	Accession number
<i>A.thaliana</i>	Calmodulin-binding protein	AF217547
	Serine threonine protein phosphatase pp2A 3	NM129811
<i>E.gunnii</i>	Cinnamyl alcohol dehydrogenase	X75480
	Cinnamoyl-CoA reductase	AJ132750
<i>N.tabacum</i>	Extensin	L38908
	Glucanase	N60402
<i>P.balsamifera</i>	Poplar APATALA3 homolog	AF057708
<i>S.tuberosum</i>	Phenylalanine ammonia-lyase	X63103
	ADP-glucose pyrophosphorylase	L36648
	Multicystatin	L16456
<i>P.vulgaris</i>	Cellulase	U34754

There are a number of useful tools and links at this website (<http://www.arabidopsis.org>) to aid researchers in their studies of the genome).

TAIR allows the retrieval of the upstream region of any gene of interest simply by entering the locus identifier (At number). For this reason a number of studies have

been done on the upstream regions of *Arabidopsis* genes often taking a genome-wide approach. Using such resources, Molina and Grotewold (2005) analyzed 12,749 *Arabidopsis* proximal promoters (500bp upstream of TSS and 5'UTRs). The TATA box occurred at a high frequency in the *Arabidopsis* core promoters and from this information a TATA nucleotide frequency matrix could be produced that accurately identified TATA-boxes in *Arabidopsis* promoters. They showed that only 29% of *Arabidopsis* promoters contained TATA-boxes and that this is comparable to *Drosophila* promoters. In a study by Lynch et al. (2005) the average length of *Arabidopsis* 5'UTR was investigated and, the authors found that the average 5'UTR length in *Arabidopsis* was 200 bp. Tatematsu et al. (2005) isolated genes with similar expression profiles based on microarray expression profiling in *Arabidopsis*, and obtained their promoter sequences from TAIR. Motifs that were over-represented in the promoters of the co-expressed genes were identified and tested with functional studies. They identified three motifs, one of which was found to be involved in sugar-mediated negative gene regulation during flower decapitation. The results produced by these and similar studies will be useful in the production of more accurate plant promoter models.

With the escalation of available promoter sequences, a number of promoter databases have been compiled such as the eukaryotic promoter database (EPD) compiled by Praz et al. (2002), the *Saccharomyces cerevisiae* promoter database (SCPD) compiled by Zhu and Zang (1999) and Plantprom DB (Shahmuradov et al. 2003; Shahmuradov et al. 2005). Plantprom DB (<http://www.softberry.com>) is a compilation of annotated, non-redundant proximal plant promoter sequences (-200: +51; with the TSS fixed at -201). All of these promoters are for RNA polymerase II promoters and have been experimentally verified transcriptional start sites (Table 1.2). There are 305

promoters, 71 from monocots, 220 dicots and 14 from other plant species. Additional information such as, the taxonomic and promoter type classification, the nucleotide frequency matrices for the promoter elements such as the TATA-box and the Inr motif, is provided (Shahmuradov et al. 2003).

#### **1.4.2 *In silico* analysis of proximal promoter regions**

There are a number of software programs based on different statistical methods that can be used to putatively identify core promoter elements (TATA-box, Inr and TSS) and their transcriptional start sites (Table 1.3). This section will focus on only three of these programs that each utilizes a different statistical approach as many of the programs listed are based on the same or similar statistical approaches. The first of these is McPromoter (Ohler 2000; Ohler and Niemann 2001) and the statistical method used here is based on hidden Markov models that identify eukaryotic polymerase II transcriptional start sites. It is one of the older programs that has been updated and is now available in version 2.0 (<http://genes.mit.edu/McPromoter.html>). The new version has a lower false positive rate of one false positive per 3 kb (Ohler and Niemann 2001). The model used in McPromoter is based on *Drosophila* promoters and for this reason it is useful in human and other animals, but is not as accurate for plants as their promoter architecture may be quite different.

TSSP-TCM (Shahmuradov et al. 2005) is a software program that can be used for the prediction of transcriptional start sites and core cis-elements in plant promoters. It was trained on 132 TATA-box and 104 TATA-less plant promoters obtained from the Plantprom database. A negative training dataset (non-promoter) consisted of coding and intron sequences from GenBank. TSSP first classifies each position in a given promoter sequence as a TATA-box or not. After the promoter is identified as a TAT-

box containing or TATA-less promoter, the next step is to look at the surrounding sequence (-200 to +50) for characteristics such as sequence content. The authors of this program (Shahmuradov et al. 2005) tested it using 40 TATA-box plant promoters and 25 TATA-less promoters that were all fully annotated. TSSP-TCM correctly identified the TSS in 87.5% of the TATA-box promoters and 84% of the TATA-less promoters proving that this is a more accurate method for the prediction of TSSs in plant promoters than many of the mammalian-based software programs (Shahmuradov et al. 2005).

Another statistical method that can be used for transcriptional start site prediction is time-delay neural networks (Waibel et al. 1989). Originally, this form of pattern recognition was used to process speech sequence patterns in a time series with local time shifts. There are a couple of promoter specific hurdles that need to be addressed when looking for patterns within a promoter region. First, the network must learn that the sequence is a feature independent of shifts in position and the second hurdle is that the network has to recognize features even if they have mutation in different positions, both of these concerns could be addressed by time-delay neural networks. This culminated in the NNPP (neural network promoter prediction) program, which was trained on the TATA-box and Inr signals of *Drosophila*. This training also allowed for variable lengths of the sequences being recognized and using the information on the positions of the TATA-box and Inr, NNPP predicts the position of the TSS (Reese et al. 2000).

Although *in silico* analysis of TSSs has become increasingly more precise the most accurate methods for confirming a TSS is by molecular methods of identification. Molecular methods such as 5'RACE and primer extension have been successfully

used to identify TSSs in both animals and plants (Schelling and Jones 1995, Xiao et al. 2006). New techniques, such as 5'SAGE, have been developed to search the whole transcriptome for functional TSSs. 5' SAGE (5'-end serial analysis of gene expression) is an adaptation of the SAGE protocol and can be used to globally identify TSSs and the frequency of individual mRNAs (Hashimoto et al. 2004). The region, which contains the TSS and core promoter elements, can also be identified by deletion studies. In this method the upstream region of a gene is systematically deleted and each section is used in an expression study in conjunction with a reporter gene. When no expression of the reporter gene is observed it indicates that the core elements required for transcription have been deleted and the DNA region just upstream of this will contain the TSS and initiation elements (Rastogi et al. 1997; Farfsing et al. 2005).

**Table 1.3:** Software programs available on the web that can be used to identify the transcriptional start sites and core promoter elements in upstream sequences.

Program	Organism	Website	References
Core-Promoter	Human	<a href="http://rulai.cshl.org/tools/genefinder/CPROMOTER/">http://rulai.cshl.org/tools/genefinder/CPROMOTER/</a>	Zhang (1998)
Dragon GC+ promoter finder	Human	<a href="http://sdmc.lit.org.sg/promoter/CGrich1_0/CGRICH.htm">http://sdmc.lit.org.sg/promoter/CGrich1_0/CGRICH.htm</a>	Bajic et al. (2005a)
Dragon promoter finder	Human	<a href="http://research.i2r.a-star.edu.sg/promoter/promoter1_5/DPF.htm">http://research.i2r.a-star.edu.sg/promoter/promoter1_5/DPF.htm</a>	Bajic et al. (2005b)
Eponine	Mammalian	<a href="http://www.sanger.ac.uk/Users/td2/eponine/">http://www.sanger.ac.uk/Users/td2/eponine/</a>	Down et al. (2002)
Fprom	Human	<a href="http://www.softberry.com/">http://www.softberry.com/</a>	
Hctata	Eukaryotic	<a href="http://25.itba.mi.cnr.it/~webgene/wwwHC_tata.html">http://25.itba.mi.cnr.it/~webgene/wwwHC_tata.html</a>	
Hprom	Any	<a href="http://www.softberry.com/">http://www.softberry.com/</a>	Solovyev et al. (2003)
McPromoter	Human	<a href="http://genes.mit.edu/McPromoter.html">http://genes.mit.edu/McPromoter.html</a>	Ohler et al. (1999; 2001)
NNPP	Drosophila	<a href="http://www.fruitfly.org/seq_tools/promoter.html">http://www.fruitfly.org/seq_tools/promoter.html</a>	Waibel et al. (1989)
Promoter 2.0	Vertebrate	<a href="http://www.cbs.dtu.dk/services/Promoter/">http://www.cbs.dtu.dk/services/Promoter/</a>	Knudsen (1999)
PromoterInspector	Mammalian	<a href="http://www.genomatix.de/products/PromoterInspector/">http://www.genomatix.de/products/PromoterInspector/</a>	Scherf et al. (2000)
PromoterScan	Primate	<a href="http://bimas.dcrn.nih.gov/molbio/proscan/">http://bimas.dcrn.nih.gov/molbio/proscan/</a>	Prestridge (1995)
TSSP	Plant	<a href="http://www.softberry.com/">http://www.softberry.com/</a>	Shahmuradov (2005)
TSSW	Human	<a href="http://www.softberry.com/">http://www.softberry.com/</a>	

### 1.4.3 Database-assisted identification of cis-regulatory elements

There are a number of the databases available online that house large collections of cis regulatory element sequences, many of which are from plants. One of the biggest databases is TRANSFAC, which contains sequences of all known eukaryotic transcription factors (Wingender 2000). TRANSFAC also contains information on the genomic binding sites and DNA binding profiles of these TFs. This website (<http://www.gene-regulation.com/pub/databases.html#transfac>) is home to a number of tools that enable promoter sequence analysis and motif identification (Wingender 2000). Although the TRANSFAC database is not plant specific, it is still useful for identifying motifs not yet identified in plants. There is more information available on other organisms such as humans and so transcription factors that have not yet been identified in plants may already have been identified in humans and homology could identify these motifs.

There are two main plant specific databases that can be used to identify already known plant cis-regulatory elements. The first of these is PLACE (Plant Cis-acting Regulatory DNA Elements) (Higo et al. 1999), which contains the nucleotide sequences of motifs found to act as regulatory elements in plants. For each motif it contains information such as the sequence, definition, description, references, PubMed ID numbers and GenBank Accession numbers. PLACE also provides a signal scan software program which is a homology-based search tool that identifies motifs within a sequence that are identical, or similar to previously reported motifs. The second well-known plant-specific database is PlantCARE (Plant Cis-Acting Regulatory Elements) (Lescot et al. 2002), which contains information on plant cis-acting regulatory elements, enhancers and repressors. The search tool at this website

(<http://bioinformatics.psb.ugent.be/webtools/plantcare/>) represents the elements using a positional matrix, consensus sequences and individual sites on the given sequence.

#### **1.4.4 *In silico* analysis of cis-regulatory elements over-represented in the promoters of co-expressed genes**

The high-throughput transcriptomics era exponentially increased the information available on many plant processes such as wood formation. Advancements in technologies such as EST sequencing, cDNA-AFLP, microarray and SAGE (reviewed in Kuhn 2001) have aided in identifying suite of genes that are expressed at the same time during a particular plant process. Gene expression data is now available in large public databases (Argraves et al. 2005; Maurer et al. 2005; Dietzsch et al. 2006; Hermida et al. 2006) that can be used to identify co-expressed genes. Although there are now vast lists of genes involved in a particular process, little is known about how these genes are regulated to produce specific expression patterns. Genome walking and full-genome sequences have aided in obtaining promoters of these co-expressed genes for comparative studies. The comparative promoter studies are based on the assumption that co-expressed genes should have the same regulatory elements in their promoters in order to drive their co-expression. This approach has been useful in identifying a number of novel motifs in plant promoters (Mohanty et al. 2005; De Bodt et al. 2006).

The vast amount of available expression data has been complimented by a number of software programs based on different algorithms and statistical methods that identify over represented motifs in large sets of promoters from co-expressed genes (Table 1.4). Most of these online tools compare sets of promoters and search for over-represented motifs. MotifSampler (Table 1.4) is the most commonly used of these



programs and is based on the Gibbs sampling algorithm that was first used for protein identification (Neuwald et al. 1995) and then adapted for DNA (Thijs et al. 2001). This algorithm determines which sequence and at what position a statistically over-represented motif is located. MotifSampler has been incorporated into a number of websites such as INCLUSive (Thijs et al. 2002) and PlantCARE (Lescot et al. 2002).

A more recently developed program is POCO (promoters of co-expressed genes) (Kankainen and Holm 2005), which can be used to identify over-represented patterns from either one or two sets of co-expressed gene promoters. It is based on the assumption that a functional transcription factor cannot up- and down-regulate differentially expressed gene sets at the same time and so the element to which it binds should be over or under-represented in two contrasting promoter sets. This program tests this hypothesis by analysing the distribution of a pattern differing among three sets of promoters: up-regulated, down-regulated and a background promoter collection. The program also uses well-known statistical approaches such as ANOVA to obtain F-statistics and P-values to add confidence levels to the results. POCO (Table 1.4) is one of the more versatile programs available as it has been trained on seven different models including *Arabidopsis*.

With such a wide choice of tools for the *in silico* identification of motifs within promoter sequences (Table 1.4) it would be prudent to use a couple of programs based on different statistical methods and then to identify the motifs common to both sets of results. These can then be checked against databases such as those mentioned above to confirm if they are known or novel motifs (Tompa et al. 2005). Even though *in silico* prediction of motifs has come along way since the review by Rombauts et al. (2003) it is still in the growing phase.

Using the *in silico* findings as a guide, more targeted functional testing can be performed and may aid in accurately identify functional regions within the functional regions in the promoter. Once the promoter region has been isolated, its functionality can be tested by placing it in front of a reporter gene such as GUS ( $\beta$ -Glucuronidase: Fedoroff and Smith 1993) and assaying the expression pattern in a model system such as tobacco and *Arabidopsis* (e.g. Lauvergeat et al. 2002). Deletion studies, where parts of the promoter sequence are deleted and the promoter function is tested, can be performed to pinpoint functional regions within the promoter (Jost et al. 2005). The proteins, which bind to the promoter region and confer an expression pattern, can be assayed using techniques such as gel shift assay (Ozyhar and Kiltz 1991), DNase footprinting (Brenowitz M 1986) and ChIP on chip analysis (Liu et al. 2002). Lacombe et al. (2000) identified functional regions of interest in the CCR gene promoter using gelshift assay and DNase footprinting. ChIP on chip assay is a method, which combines Chromatin immuno precipitation and microarray technology to identify transcription factor binding sites in the chromatin (Odom et al. 2004). Methods such as these will aid us in further understanding the transcriptional regulation of plant processes such as xylogenesis.

**Table 1.4** software packages that are available for the identification of cis-regulatory elements.

Program Known Motifs	Organism	Web site	Reference
AliBaba2	Eukaryotic	<a href="http://www.gene-regulation.com/pub/programs/alibaba2/">http://www.gene-regulation.com/pub/programs/alibaba2/</a>	Grabe (2002)
AlignACE	Any	<a href="http://atlas.med.harvard.edu/cgi-bin/alignace.pl">http://atlas.med.harvard.edu/cgi-bin/alignace.pl</a>	Hughes et al. (2006)
Bindgene	Human	<a href="http://www.bioinf.man.ac.uk/~lockwood/bindgene.html">http://www.bioinf.man.ac.uk/~lockwood/bindgene.html</a>	Lockwood and Fraying (2003)
BioProspector	Eukaryotic	<a href="http://ai.stanford.edu/~xslu/BioProspector/">http://ai.stanford.edu/~xslu/BioProspector/</a>	Liu et al. (2001)
Cister	Eukaryotic	<a href="http://zlab.bu.edu/%7Emfrith/cister.shtml">http://zlab.bu.edu/%7Emfrith/cister.shtml</a>	Frith et al. (2001)
CompareProspector	Mammal	<a href="http://ai.stanford.edu/~iliu/CompareProspector/">http://ai.stanford.edu/~iliu/CompareProspector/</a>	Lui et al. (2004)
Gemoda	Any	<a href="http://web.mit.edu/bamel/gemoda/">http://web.mit.edu/bamel/gemoda/</a>	Jensen et al. (2006)
GLAM	Any	<a href="http://zlab.bu.edu/glam/">http://zlab.bu.edu/glam/</a>	Frith et al. (2005)
Improbizer	<i>C. elegans</i>	<a href="http://www.cse.ucsc.edu/~kent/improbizer/improbizer.html">http://www.cse.ucsc.edu/~kent/improbizer/improbizer.html</a>	
JASPAR	Eukaryotic	<a href="http://jaspar.cgb.ki.se/cgi-bin/jaspar_db.pl">http://jaspar.cgb.ki.se/cgi-bin/jaspar_db.pl</a>	Sandelin (2004)
MATCH	Eukaryotic	<a href="http://compel.bionet.nsc.ru/Match/Match.html">http://compel.bionet.nsc.ru/Match/Match.html</a>	Kel et al. (2003)
MatInspector	Eukaryotic	<a href="http://www.genomatix.de/software_services/software/MatInspector">http://www.genomatix.de/software_services/software/MatInspector</a>	Quandt et al. (1995)
Matrix search	Any	<a href="http://bimas.dcrn.nih.gov/molbio/matrixs/">http://bimas.dcrn.nih.gov/molbio/matrixs/</a>	Chen et al. (1995)
MEME	Any	<a href="http://meme.sdsc.edu/meme/website/intro.html">http://meme.sdsc.edu/meme/website/intro.html</a>	Bailey and Gribskov (1997)
MITRA	Any	<a href="http://fluff.cs.columbia.edu:8080/domain/mitra.html">http://fluff.cs.columbia.edu:8080/domain/mitra.html</a>	Eskin and Pevzner (2005)
Motif Analysis	Plant	<a href="http://www.arabidopsis.org/">http://www.arabidopsis.org/</a>	
Motif analysis workbench	Yeast	<a href="http://bioportal.weizmann.ac.il/~lapidotm/rMotif/html/">http://bioportal.weizmann.ac.il/~lapidotm/rMotif/html/</a>	Lapidot (2003)
MotifSampler	Any	<a href="http://homes.esat.kuleuven.be/~thijs/WWork/MotifSampler.html">http://homes.esat.kuleuven.be/~thijs/WWork/MotifSampler.html</a>	Thijs et al. (2005)
Motifviz	Plant	<a href="http://biowulf.bu.edu/MotifViz/">http://biowulf.bu.edu/MotifViz/</a>	Fu et al. (2004)
NSITE	Mammal	<a href="http://www.softberry.com/">http://www.softberry.com/</a>	
NSITE-PL	Plant	<a href="http://www.softberry.com/">http://www.softberry.com/</a>	
NSITEH	Any	<a href="http://www.softberry.com/">http://www.softberry.com/</a>	
NSITEM	Mammal	<a href="http://www.softberry.com/">http://www.softberry.com/</a>	
NSITEM-PL	Plant	<a href="http://www.softberry.com/">http://www.softberry.com/</a>	
PLACE Signal Scan	Plant	<a href="http://www.dna.affrc.go.jp/htdocs/PLACE/signalup.html">http://www.dna.affrc.go.jp/htdocs/PLACE/signalup.html</a>	Higo et al. (1999)
P-match	Human	<a href="http://www.gene-regulation.com/cgi-bin/pub/programs/pmatch/bin/p-match.cgi">http://www.gene-regulation.com/cgi-bin/pub/programs/pmatch/bin/p-match.cgi</a>	Chekmenov et al. (2005)
POCO	Any	<a href="http://ekhidna.biocenter.helsinki.fi/poco">http://ekhidna.biocenter.helsinki.fi/poco</a>	Kankainen and Holm (2005)
RSA-tools	Any	<a href="http://rsat.scmbb.ulb.ac.be/rsat/">http://rsat.scmbb.ulb.ac.be/rsat/</a>	Van Helden (1998; 2000)
ScanWM-PL	Plant	<a href="http://www.softberry.com/">http://www.softberry.com/</a>	
Search for CARE	Plant	<a href="http://intra.psb.ugent.be:8080/PlantCARE/">http://intra.psb.ugent.be:8080/PlantCARE/</a>	Lescot (2003)
SignalScan	Any	<a href="http://bimas.dcrn.nih.gov/molbio/signal/">http://bimas.dcrn.nih.gov/molbio/signal/</a>	Prestridge (1991)
Tess	Eukaryotic	<a href="http://www.cbil.upenn.edu/tess/">http://www.cbil.upenn.edu/tess/</a>	Wallace et al. (1997)
TFBind	Eukaryotic	<a href="http://tfbind.ims.u-tokyo.ac.jp/">http://tfbind.ims.u-tokyo.ac.jp/</a>	Tsunoda and Takagi (1998)
TFSEARCH	Any	<a href="http://www.cbrc.jp/research/db/TFSEARCH.html">http://www.cbrc.jp/research/db/TFSEARCH.html</a>	Akiyama (1995)
VISTA	Any	<a href="http://genome.lbl.gov/vista/index.shtml">http://genome.lbl.gov/vista/index.shtml</a>	Loots et al. (2002)
Weeder	Eukaryotic	<a href="http://159.149.109.16:8080/weederWeb/">http://159.149.109.16:8080/weederWeb/</a>	Pavesi et al. (2004)

## 1.5. Conclusion

Wood formation is a very important developmental process, but little is known about the transcriptional regulation of this process. The sequencing of two plant genomes

(*Populus* and *Arabidopsis*) has helped in gaining more insight into the genes involved in xylogenesis and how they are regulated. These two genomes have also helped in gaining a better understanding of plant promoters in general and how they are different from animal promoters. The large amount of plant promoter sequence now available has also aided in the production and refinement of motif identification tools specific for plants. Although the promoter and motif identification tools have improved greatly, they are still not very accurate and it is suggested that as there are so many of these tools available it would be best to use a combination of tools to increase the confidence in the results obtained. Any information obtained from the *in silico* identification of promoters or motifs will have to be verified by functional testing, but at least these tools can aid in targeted functional testing. These new tools and resources will prove to be indispensable while attempting to unravel the genetic regulation of xylogenesis.

The discussion above highlights the fact that although, there are many studies on the process of wood formation and the genes involved, the regulatory mechanisms which underlie the process are still poorly described. Regulatory gene networks need to be constructed to gain a better understanding of xylogenesis. While some aspects of wood formation such as lignin biosynthesis have been well characterised other pathways such as cellulose biosynthesis still require much investigation. The poplar genome sequence and much anticipated *Eucalyptus* genome sequence will aid in the construction of the regulatory networks involved in wood formation.

## **1.6 Aim of the study**

The first report of a cellulose synthase being isolated from *Eucalyptus* was from a cDNA-AFLP study by Ranik et al. (2006) and this led to the isolation of six other

*Eucalyptus* cellulose synthase genes Ranik and Myburg (2006). The Aim of this study was to investigate the transcriptional regulation of different members of the *CesA* gene family in *Eucalyptus*. The *EgCesA* promoter regions were isolated and comparatively analysed with the orthologous regions in *Arabidopsis* and *Populus*. Bioinformatics tools were used to identify putative regulatory motifs that play a role in the *CesA* genes expression patterns. The promoters of the genes identified in this study will be useful in further understanding the regulation of wood formation and could be used to express transgenes in a tissue specific manner in order to enhance wood characteristics.

## 1.7 References

- Aida, M., T. Ishida, H. Fukaki, H. Fujisawa and M. Tasaka (1997) Genes involved in organ separation in *Arabidopsis*: an analysis of the cup-shaped cotyledon mutant. *Plant Cell* 9:841-57
- Akiyama, Y. (1995) TFSEARCH: searching transcription factor binding sites.
- Argraves, G. L., S. Jani, J. L. Barth and W. S. Argraves (2005) ArrayQuest: a web resource for the analysis of DNA microarray data. *BMC Bioinformatics* 6:287
- Arioli, T., L. C. Peng, A. S. Betzner, J. Burn, W. Wittke, W. Herth, C. Camilleri, H. Hofte, J. Plazinski, R. Birch, et al. (1998) Molecular analysis of cellulose biosynthesis in *Arabidopsis*. *Science* 279:717-720
- Bailey, T. L. and M. Gribskov (1997) Score distributions for simultaneous matching to multiple motifs. *J. Comput. Biol.* 4:45-59
- <sup>a</sup>Bajic, V. B., Chong, A., et al. (2002). An intelligent system for vertebrate promoter recognition. *Ieee Intelligent Systems* 17:64-70.
- <sup>b</sup>Bajic, V. B., Seah, S. H., et al. (2002). Dragon Promoter Finder: recognition of vertebrate RNA polymerase II promoters. *Bioinformatics* 18:198-199.
- Ballestar, E. and A. P. Wolffe (2001) Methyl-CpG-binding proteins: Targeting specific gene repression. *Eur. J. Biochem.* 268:1-6
- Bannan, M. W. (1957) The relative frequency of the different types of anticlinal divisions in conifer cambia. *Can. J. Bot.* 35:875-884
- Baumann, K., A. De Paolis, P. Costantino and G. Gualberti (1999) The DNA binding site of the Dof protein NtBBF1 is essential for tissue-specific and auxin-regulated expression of the rolB oncogene in plants. *Plant Cell* 1:323-34
- Best, A. A., H. G. Morrison, A. G. McArthur, M. L. Sogin and G. J. Olsen (2004) Evolution of eukaryotic transcription: insights from the genome of *Giardia lamblia*. *Genome Res.* 14:1537-1547
- Bowman, J. L., Eshed, Y. and S. F. Baum (2002) Establishment of polarity in angiosperm lateral organs. *Trends Genet.* 18:134-141
- Brenowitz, M. S. D., Shea, M. A., Ackers, G.K. (1986) Quantitative DNase footprint titration: a method for studying protein-DNA interactions. *Methods Enzymol.* 130:132-81
- Brown, R. M. and Saxena, I. M. (2000) Cellulose biosynthesis: A model for understanding the assembly of biopolymers. *Plant Physiol. Biochem.* 38:57-67
- Burn, J. E., Hocart, C. H., Birch, R. J., Cork, A. C. and Williamson, R. E. (2002) Functional analysis of the cellulose synthase genes *CesA1*, *CesA2*, and *CesA3* in *Arabidopsis*. *Plant Physiol.* 129:797-807

- Burton, R. A., Shirley N. J., King B. J., Harvey A. J. and Fincher G. B. (2004). The *CesA* gene family of barley: Quantitative analysis of transcripts reveals two groups of co-expressed genes. *Plant Physiol.* 134:224-236
- Campbell, M. M. and Sederoff, R. (1996) Variation in lignin content and composition. *Plant Physiol.* 110:3-13
- Casacuberta, E., Puigdomenech, P. and Monfort, A. (2000) Distribution of microsatellites in relation to coding sequences within the *Arabidopsis thaliana* genome. *Plant Sci.* 157:97-104
- Chekmenov, D. S., Haid, C. and Kel, A. E. (2005) P-Match: transcription factor binding site search by combining patterns and weight matrices. *Nucleic Acids Res.* 33:432-437
- Chen, Q. K., Hertz, G. Z., et al. (1995). MATRIX SEARCH 1.0: a computer program that scans DNA sequences for transcriptional elements using a database of weight matrices. *Comput Appl Biosci* 11:563-6.
- Ciacci-Zanella, J. R. and Jones, C. (1999) Fumonisin B1, a mycotoxin contaminant of cereal grains, and inducers of apoptosis via the tumour necrosis factor pathway and caspase activation. *Food Chem. Toxicol.* 37:703-12
- Connors, B. J., Miller, M., Maynard, C. A. and Powell, W. A. (2002) Cloning and characterization of promoters from American chestnut capable of directing reporter gene expression in transgenic *Arabidopsis* plants. *Plant Sci.* 163:771-781
- Cosgrove, D. J. (1998) Cell wall loosening by expansins. *Plant Physiol.* 118:333-339
- Cosgrove, D. J. (2005). Growth of the plant cell wall. *Nature Rev. Mol. Cell Biol.* 6:850-861
- Cosgrove, D. J., Li, L. C., Cho, H. T., Hoffmann-Benning, S., Moore, R. C. and Blecker, D. (2002) The growing world of expansins. *Plant and Cell Physiol* 43:1436-1444
- Cseke, L. J., Zheng, J. and Podila, G. K. (2003) Characterization of *PTM5* in aspen trees: a *MADS-box* gene expressed during woody vascular development. *Gene* 318:55-67
- Daraselia, N. D., Tarchevskaya, S. and Narita, J. O. (1996) The promoter for tomato 3-hydroxy-3-methylglutaryl coenzyme A reductase gene 2 has unusual regulatory elements that direct high-level expression. *Plant Physiol.* 112:727-733
- De Bodt, S., Theissen, G. and Van de Peer, Y. (2006) Promoter analysis of MADS-Box genes in eudicots through phylogenetic footprinting. *Mol. Biol. Evol.* 23:1293-1303
- Desprez, T., Vernhettes, S., Fagard, M., Refregier, G., Desnos, T., Aletti, E., Py, N., Pelletier, S. and Hofte, H. (2002) Resistance against herbicide isoxaben and cellulose deficiency caused by distinct mutations in same cellulose synthase isoform CESA6. *Plant Physiol.* 128:482-490
- Deveraux, Q. L. and Reed, J. C. (1999) IAP family proteins--suppressors of apoptosis. *Genes Dev.* 13:239-52
- Dietzsch, J., Gehlenborg, N. and Nieselt, K. (2006) Mayday: a microarray data analysis workbench. *Bioinformatics* 22:1010-1012

- Djerbi, S., Aspeborg, H., Nilsson, P., Sundberg, B., Mellerowicz, E., Blomqvist, K. and Teeri, T. T. (2004) Identification and expression analysis of genes encoding putative cellulose synthases (*CesA*) in the hybrid aspen, *Populus tremula* (L.) x *P. tremuloides* (Michx.). Cellulose 11:301-312
- Doukhanina, E. V., S. Chen, E. van der Zalm, A. Godzik, J. Reed and M. B. Dickman (2006). "Identification and functional characterization of the BAG protein family in *Arabidopsis thaliana*." J. Biol. Chem. 281:18793-18801
- Down, T. A. and Hubbard, T. J. P. (2002) Computational detection and location of transcription start sites in mammalian genomic DNA. Genome Res. 12:458-461
- Elliott, K. A. and Shirsat, A. H. (1998) Promoter regions of the *EXTA* extensin gene from *Brassica napus* control activation in response to wounding and tensile stress. Plant Mol. Biol. 37:675-87
- Elmayan, T. and Tepfer, M. (1995) Evaluation in tobacco of the organ specificity and strength of the *rolD* promoter, domain A of the 35S promoter and the 35S2 promoter. Transgenic Res. 4:388-96
- Emery, J. F., Floyd, S. K., Alvarez, J., Eshed, Y., Hawker, N. P., Izhaki, A., Baum, S. F. and Bowman, J. L. (2003) Radial patterning of *Arabidopsis* shoots by class III HD-ZIP and *KANADI* genes. Curr. Biol. 13:1768-74
- Emons, A. M. C. and Mulder, B. M. (2000) How the deposition of cellulose microfibrils builds cell wall architecture. Trends Plant Sci. 5:35-40
- Eskin, E. and Pevzner, P. A. (2002). Finding composite regulatory patterns in DNA sequences. Bioinformatics 18 Suppl 1:S354-63.
- Fagard, M., Hofte, H. and Vernhettes, S. (2000) Cell wall mutants. Plant Physiol. Biochem. 38:15-25
- Farfaring, J. W., Auffarth, K. and Basse, C. W. (2005) Identification of cis-active elements in *Ustilago maydis mig2* promoters conferring high-level activity during pathogenic growth in maize. Mol. Plant Microbe Interact. 18:75-87
- Featherstone, M. (2002) Co-activators in transcription initiation: here are your orders. Curr. Opin. Genet. Dev. 12:149-155
- Fedoroff, N. V. and Smith, D. L. (1993). A versatile system for detecting transposition in *Arabidopsis*. Plant J. 3:273-89
- Fessele, S., Maier, H., Zischek, C., Nelson, P. J. and Werner, T. (2002) Regulatory context is a crucial part of gene function. Trends Genet. 18:60-63
- Frith, M. C., Hansen, U. and Weng, Z. P. (2001) Detection of cis-element clusters in higher eukaryotic DNA. Bioinformatics 17:878-889
- Fukuda, H. (2000) Programmed cell death of tracheary elements as a paradigm in plants. Plant Mol. Biol. 44:245-253



- Fukuda, H. and Komamine, A. (1980). Establishment of an experimental system for the tracheary element differentiation from single cells isolated from the mesophyll of *Zinnia elegans*. *Plant Physiology* 65:57-60.
- Goicoechea, M., Lacombe, E., Legay, S., Mihaljevic, S., Rech, P., Jauneau, A., Lapiere, C., Pollet, B., Verhaegen, D., Chaubet-Gigot, N. et al. (2005) *EgMYB2*, a new transcriptional activator from *Eucalyptus* xylem, regulates secondary cell wall formation and lignin biosynthesis. *Plant J.* 43:553-67
- Grabe, N. (2002) AliBaba2: context specific identification of transcription factor binding sites. *In Silico Biol.* 2:S1-15
- Grima-Pettenati, J. and Goffner, D. (1999) Lignin genetic engineering revisited. *Plant Sci.* 145:51-65
- Groover, A. and Jones, A. M. (1999) Tracheary element differentiation uses a novel mechanism coordinating programmed cell death and secondary cell wall synthesis. *Plant Physiol.* 119:375-384
- Groover, A. T. (2005) What genes make a tree a tree? *Trends Plant Sci.* 10:210-214
- Gunnerås, S. A. (2005) Wood formation and transcript analysis with focus on tension wood and ethylene biology. Department of Forest Genetics and Plant Physiology. Umeå, Swedish University of Agricultural Sciences
- Gupta, P. K., Balyan, I. S., Sharma, P. C. and Ramesh, B. (1996) Microsatellites in plants: a new class of molecular markers. *Curr. Sci.* 70:45-54
- Harding, S. A., J. Leshkevich, V. L. Chiang and C. J. Tsai (2002). "Differential substrate inhibition couples kinetically distinct 4-coumarate: coenzyme A ligases with spatially distinct metabolic roles in quaking aspen." *Plant Physiol* 128:428-438.
- Hashimoto, S., Suzuki, Y., Kasai, Y., Morohoshi, K., Yamada, T., Sese, J., Morishita, S., Sugano, S. and Matsushima, K. (2004) 5'-end SAGE for the analysis of transcriptional start sites. *Nat Biotechnol* 22:1146-1149
- Hatton, D., Sablowski, R., Yung, M. H., Smith, C., Schuch, W. and Bevan, M. (1995) Two classes of cis sequences contribute to tissue-specific expression of a *PAL2* promoter in transgenic tobacco. *Plant J.* 7:859-76
- Hauffe, K. D., Paszkowski, U., Schulzelefert, P., Hahlbrock, K., Dangl, J. L. and Douglas, C. J. (1991) A parsley *4CL-1* promoter fragment specifies complex expression patterns in transgenic tobacco. *Plant Cell* 3:435-443
- Heinekamp, T., Kuhlmann, M., Lenk, A., Strathmann, A. and Droge-Laser, W. (2002) The tobacco bZIP transcription factor BZI-1 binds to G-box elements in the promoters of phenylpropanoid pathway genes in vitro, but it is not involved in their regulation in vivo. *Mol. Genet. Genomics* 267:16-26

- Hermida, L., Schaad, O., Demougin, P., Descombes, P. and Primig, M. (2006) MIMAS: an innovative tool for network-based high-density oligonucleotide microarray data management and annotation. *BMC Bioinformatics* 7:190
- Higashi, K., Takasawa, R., Yoshimori, A., Goh, T., Tanuma, S. and Kuchitsu, K. (2005) Identification of a novel gene family, paralogs of inhibitor of apoptosis proteins present in plants, fungi, and animals. *Apoptosis* 10:471-80
- Higo, K., Ugawa, Y., Iwamoto, M. and Korenaga, T. (1999) Plant cis-acting regulatory DNA elements (PLACE) database. *Nucl. Acids Res.* 27:297-300
- Hochheimer, A. and Tjian, R. (2003) Diversified transcription initiation complexes expand promoter selectivity and tissue-specific gene expression. *Genes Dev.* 17:1309-1320
- Hu, W. J., Popko, J. L., Lung, J. R., Kawaoka, A., Kao, Y. Y., Hideki, S., Stokke, S. S., Tsai, C. J. and Chiang, V. L. (1998) Transgenic aspen trees with reduced lignin quantity and increased cellulose content. *Papers of the American Chemical Society* 215
- Huala, E., Dickerman, A. W., Garcia-Hernandez, M., Weems, D., Reiser, L., Lafond, F., Hanley, D., Kiphart, D., Zhuang, M. Z., Huang, W. et al. (2001) The *Arabidopsis* Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucl. Acids Res.* 29:102-105
- Hughes, T. A. (2006) Regulation of gene expression by alternative untranslated regions. *Trends Genet.* 22:119-22
- Iwahana, H., Mizusawa, N., Ii, S., Yoshimoto, K. and Itakura, M. (1994) An end-trimming method to amplify adjacent cDNA fragments by PCR. *Biotechniques* 16:94-8
- Jensen, K. L., Styczynski, M. P., Rigoutsos, I. and Stephanopoulos, G. N. (2006) A generic motif discovery algorithm for sequential data. *Bioinformatics* 22:21-28
- Joshi, C., Bhandari, S., Ranjan, P., Kalluri, U. C., Liang, X., Fujino, T. and Samuga, A. (2004) Genomics of cellulose biosynthesis in poplars. *New Phytol.* 164:53-61
- Jost, W., Link, S., Horstmann, V., Decker, E. L., Reski, R. and Gorr, G. (2005) Isolation and characterisation of three moss-derived beta-tubulin promoters suitable for recombinant expression. *Curr. Genet.* 47:111-20
- Kanhere, A. and Bansal, M. (2005) Structural properties of promoters: similarities and differences between prokaryotes and eukaryotes. *Nucl. Acids Res.* 33:3165-3175
- Kankainen, M. and Holm, L. (2005) POCO: discovery of regulatory patterns from promoters of oppositely expressed gene sets. *Nucleic Acids Res.* 33:427-431
- <sup>a</sup>Kass, S. U., Landsberger, N. and Wolffe, A. P. (1997) DNA methylation directs a time-dependent repression of transcription initiation. *Curr. Biol.* 7:157-65
- <sup>b</sup>Kass, S. U., Pruss, D. and Wolffe, A. P. (1997) How does DNA methylation repress transcription? *Trends Genet* 13:444-9

- Kawaoka, A. and Ebinuma, H. (2001) Transcriptional control of lignin biosynthesis by tobacco LIM protein. *Phytochemistry* 57:1149-57
- Kawaoka, A., Kaothien, P., Yoshida, K., Endo, S., Yamada, K. and Ebinuma, H. (2000) Functional analysis of tobacco LIM protein *NtLIM1* involved in lignin biosynthesis. *Plant J.* 22:289-301
- Kel, A. E., Gossling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O. V. and Wingender, E. (2003) MATCH (TM): a tool for searching transcription factor binding sites in DNA sequences. *Nucl. Acids Res.* 31:3576-3579
- Keller, B. and Heierli, D. (1994) Vascular expression of the *grp1.8* promoter is controlled by three specific regulatory elements and one unspecific activating sequence. *Plant Mol. Biol.* 26:747-56
- Kim, M. J., Shin, J. S., Kim, J. K. and Suh, M. C. (2005) Genomic structures and characterization of the 5'-flanking regions of acyl carrier protein and Delta4-palmitoyl-ACP desaturase genes from *Coriandrum sativum*. *Biochim. Biophys. Acta.* 1730:235-44
- Knudsen, S. (1999) Promoter 2.0: for the recognition of PolII promoter sequences. *Bioinformatics* 15:356-361
- Ko, J. H., Han, K. H., Park, S. and Yang, J. M. (2004) Plant body weight-induced secondary growth in *Arabidopsis* and its transcription phenotype revealed by whole-transcriptome profiling. *Plant Physiol.* 135:1069-1083
- Koshino-Kimura, Y., Wada, T., Tachibana, T., Tsugeki, R., Ishiguro, S. and Okada, K. (2005) Regulation of CAPRICE transcription by MYB proteins for root epidermis differentiation in *Arabidopsis*. *Plant Cell Physiol.* 46:817-826
- Kubo, M., Udagawa, M., Nishikubo, N., Horiguchi, G., Yamaguchi, M., Ito, J., Mimura, T., Fukuda, H. and Demura, T. (2005) Transcription switches for protoxylem and metaxylem vessel formation. *Genes Dev.* 19:1855-1860
- Kuhn, E. (2001) From library screening to microarray technology: strategies to determine gene expression profiles and to identify differentially regulated genes in plants. *Ann. Botany* 87:139-155
- Kutach, A. K. and Kadonaga, J. T. (2000) The downstream promoter element (DPE) appears to be as widely used as the TATA-box in *Drosophila* core promoters. *Mol. Cell. Biol.* 20:4754-4764
- Lacombe, E., Van Doorselaere, J., Boerjan, W., Boudet, A. M. and Grima-Pettenati, J. (2000) Characterization of cis-elements required for vascular expression of the cinnamoyl CoA-reductase gene and for protein-DNA complex formation. *Plant J.* 23:663-676
- Lam, E. and Chua, N. H. (1989) ASF-2: a factor that binds to the cauliflower mosaic virus 35S promoter and a conserved GATA motif in Cab promoters. *Plant Cell* 1:1147-56
- Lapidot, M. and Pilpel, Y. (2003) Comprehensive quantitative analyses of the effects of promoter sequence elements on mRNA transcription. *Nucl. Acids Res.* 31:3824-3828


- Latchman, D. S. (1999) Transcription factors. New York, Oxford University Press. pp 1-319
- Lauvergeat, V., Rech, P., Jauneau, A., Guez, C., Coutos-Thevenot, P. and Grima-Pettenati, J. (2002) The vascular expression pattern directed by the *Eucalyptus gunnii* cinnamyl alcohol dehydrogenase *EgCAD2* promoter is conserved among woody and herbaceous plant species. *Plant Mol. Biol.* 50:497-509
- Lee, M. M. and Schiefelbein, J. (2002) Cell pattern in the *Arabidopsis* root epidermis determined by lateral inhibition with feedback. *Plant Cell* 14:611-618
- Lescot, M., Dehais, P., Thijs, G., Marchal, K., Moreau, Y., Van de Peer, Y., Rouze, P. and Rombauts, S. (2002) PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for *in silico* analysis of promoter sequences. *Nucl. Acids Res.* 30:325-327
- Leyva, A., Liang, X., Pintor-Toro, J. A., Dixon, R. A. and Lamb, C. J. (1992) Cis-element combinations determine phenylalanine ammonia-lyase gene tissue-specific expression patterns. *Plant Cell* 4:263-271
- Li, L., Zhou, Y., Cheng, X., Sun, J., Marita, J. M., Ralph, J. and Chiang, V. L. (2003) Combinatorial modification of multiple lignin traits in trees through multigene cotransformation. *PNAS* 100:4939-4944
- Liu, X., Brutlag, D. L. and Liu, J. S. (2001) Bioprospector: Discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac. Symp. Biocomput.* pp 127-138
- Liu, X. S., Brutlag, D. L. and Liu, J. S. (2002) An algorithm for finding protein-DNA binding sites with applications to chromatin-immuno-precipitation microarray experiments. *Nat. Biotechnol.* 20:835-839
- Liu, Y., Liu, X. S., Wei, L., Altman, R. B. and Batzoglou, S. (2004) Eukaryotic regulatory element conservation analysis and identification using comparative genomics. *Genome Res.* 14:451-458
- Lockwood, C. R. and Frayling, T. M. (2003) Combining genome and mouse knockout expression data to highlight binding sites for the transcription factor HNF1 alpha. *In Silico Biol.* 3:57-70
- Lohmann, J. U., Hong, R. L., Hobe, M., Busch, M. A., Parcy, F., Simon, R. and Weigel, D. (2001) A molecular link between stem cell regulation and floral patterning in *Arabidopsis*. *Cell* 105:793-803
- Loke, J. C., Stahlberg, E. A., Strenski, D. G., Haas, B. J., Wood, P. C. and Li, Q. Q. (2005) Compilation of mRNA polyadenylation signals in *Arabidopsis* revealed a new Signal element and potential secondary structures. *Plant Physiol.* 138:1457-1468
- Long, J. A., Moan, E. I., Medford, J. I. and Barton, M. K. (1996) A member of the KNOTTED class of homeodomain proteins encoded by the *STM* gene of *Arabidopsis*. *Nature* 379:66-9

- Loots, G. G., Ovcharenko, I., Pachter, L., Dubchak, I. and Rubin, E. M. (2002) rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Res.* 12:832-839
- Lynch, M., Scofield, D. G. and Hong, X. (2005) The evolution of transcription-initiation sites. *Mol. Biol. Evol.* 22:1137-1146
- Ma, H., Yanofsky, M. F. and Meyerowitz, E. M. (1991) *AGL1-AGL6*, an *Arabidopsis* gene family with similarity to floral homeotic and transcription factor genes. *Genes Dev.* 5:484-95
- Martin, P., Makepeace, K., Hill, S. A., Hood, D. W. and Moxon, E. R. (2005) Microsatellite instability regulates transcription factor binding and gene expression. *PNAS* 102:3800-3804
- Maurer, M., Molidor, R., Sturn, A., Hartler, J., Hackl, H., Stocker, G., Prokesch, A., Scheideler, M. and Trajanoski, Z. (2005) MARS: microarray analysis, retrieval, and storage system. *BMC Bioinformatics* 6:101
- McQueen-Mason, S. J. and Cosgrove, D. J. (1995) Expansin mode of action on cell walls: Analysis of wall hydrolysis, stress relaxation, and binding. *Plant Physiol.* 107:87-100
- Mele, G., Ori, N., Sato, Y., and Hake, S. (2003) The knotted1-like homeobox gene *BREVIPEDICELLUS* regulates cell differentiation by modulating metabolic pathways. *Genes Dev.* 17:2088-2093
- Mellerowicz, E. J., Baucher, M., Sundberg, B. and Boerjan, W. (2001) Unravelling cell wall formation in the woody dicot stem. *Plant Mol. Biol.* 47:239-274
- Miller, L. K. (1999). An exegesis of IAPs: salvation and surprises from BIR motifs. *Trends Cell Biol.* 9:323-328
- Mingam, A., Toffano-Nioche, C., Brunaud, V., Boudet, N., Kreis, M. and Lecharny, A. (2004) DEAD-box RNA helicases in *Arabidopsis thaliana*: establishing a link between quantitative expression, gene structure and evolution of a family of genes. *Plant Biotech. J.* 2:401-415
- Mitsuda, N., Seki, M., Shinozaki, K. and Ohme-Takagi, M. (2005) The NAC transcription factors NST1 and NST2 of *Arabidopsis* regulate secondary wall thickenings and are required for anther dehiscence. *Plant Cell* 17:2993-3006
- Mohanty, B., Krishnan, S. P. T., Swarup, S. and Bajic, V. B. (2005) Detection and preliminary analysis of motifs in promoters of anaerobically induced genes of different plant species. *Ann. Botany* 96:669-681
- Molina, C. and Grotewold, E. (2005) Genome wide analysis of *Arabidopsis* core promoters. *BMC Genomics.* 6:25-35
- Moss, T. (2004) At the crossroads of growth control: making ribosomal RNA. *Curr. Opin. Genet. Dev.* 14:210-217

- Moyle, R., Moody, J., Phillips, L., Walter, C. and Wagner, A. (2002) Isolation and characterization of a *Pinus radiata* lignin biosynthesis-related O-methyltransferase promoter. *Plant Cell Rep.* 20:1052-1060
- Murre, C., McCaw, P. S., Vaessin, H., Caudy, M., Jan, L. Y., Jan, Y. N., Cabrera, C. V., Buskin, J. N., Hauschka, S. D, Lassar, A. B. et al. (1989) Interactions between heterologous helix-loop-helix proteins generate complexes that bind specifically to a common DNA sequence. *Cell* 58:537-44
- Neuwald, A. F., Liu, J. S. and Lawrence, C. E. (1995) Gibbs motif sampling: detection of bacterial outer membrane protein repeats. *Protein Sci.* 4:1618-32
- Newman, L. J., Perazza, D. E., Juda, L. and Campbell, M. M. (2004) Involvement of the R2R3-MYB, AtMYB61, in the ectopic lignification and dark-photomorphogenic components of the det3 mutant phenotype. *Plant J.* 37:239-250
- Nikolov, D. B., Chen, H., Halay, E. D., Hoffman, A., Roeder, R. G. and Burley, S. K. (1996) Crystal structure of a human TATA box-binding protein/TATA element complex. *PNAS* 93:4862-4867
- Odom, D. T., Zizlsperger, N., Gordon, D. B., Bell, G. W., Rinaldi, N. J., Murray, H. L., Volkert, T. L., Schreiber, J., Rolfe, P. A. and Gifford, D. K. (2004) Control of pancreas and liver gene expression by HNF transcription factors. *Science* 303:1378-1381
- Ohashi-Ito, K. and Fukuda, H. (2003) HD-Zip III homeobox genes that include a novel member, *ZeHB-13* (*Zinnia*)/*ATHB-15* (*Arabidopsis*), are involved in procambium and xylem cell differentiation. *Plant Cell Physiol.* 44:1350-1358
- Ohler, U. (2000) Promoter prediction on a genomic scale: the ADH experience. *Genome Res.* 10:539-42
- Ohler, U., Harbeck, S., Niemann, H., Noth, E. and Reese, M. G. (1999) Interpolated markov chains for eukaryotic promoter recognition. *Bioinformatics* 15:362-369
- Ohler, U. and Niemann, H. (2001) Identification and analysis of eukaryotic promoters: recent computational approaches. *Trend Genet.* 17:56-60
- Ozyhar, A. and Kiltz, H. H. (1991) High-resolution gel filtration of the ecdysteroid receptor-DNA complex: an alternative to the electrophoretic mobility shift assay. *J. Chromatogr.* 587:11-7
- Pabo, C. O. (1992) Transcription factors: Structural families and principles of DNA recognition *Annu. Rev. Biochem.* 61:1053-1095
- Patzlaff, A., Newman, L. J., Dubos, C., Whetten, R., Smith, C., McInnis, S., Bevan, M. W., Sederoff, R. R. and Campbell, M. M. (2003) Characterisation of *PtMYB1*, an R2R3-MYB from pine xylem. *Plant Mol. Biol.* 53:597-608
- Pavesi, G., Mereghetti, P., Mauri, G. and Pesole, G. (2004) Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucl. Acids Res.* 32:199-203



- Plomion, C., Leprovost, G. and Stokes, A. (2001) Wood formation in trees. *Plant Physiol.* 127:1513-1523
- Praz, V., Perier, R., Bonnard, C. and Bucher, P. (2002) The eukaryotic promoter database, EPD: new entry types and links to gene expression data. *Nucl. Acids Res.* 30:322-324
- Prestridge, D. S. (1991) SIGNAL SCAN: a computer program that scans DNA sequences for eukaryotic transcriptional elements. *Comput. Appl. Biosci.* 7:203-206
- Prestridge, D. S. (1995) Predicting pol II promoter sequences using transcription factor binding sites. *J. Mol. Biol.* 249:923-932
- Prigge, M. J., Otsuga, D., Alonso, J. M., Ecker, J. R., Drews, G. N. and Clark, S. E. (2005) Class III homeodomain-leucine zipper gene family members have overlapping, antagonistic, and distinct roles in *Arabidopsis* development. *Plant Cell* 17:61-76
- Quandt, K., Frech, K., Karas, H., Wingender, E. and Werner, T. (1995) MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucl. Acids Res.* 23:4878-4884
- Raes, J., Rohde, A., Christensen, J. H., Van de Peer, Y. and Boerjan, W. (2003) Genome-wide characterization of the lignification toolbox in *Arabidopsis*. *Plant Physiol.* 133:1051-1071
- Ranik, M., Creux, N. M. and Myburg, A. A. (2006) Within-tree transcriptome profiling in wood-forming tissues of a fast-growing *Eucalyptus* tree. *Tree Physiol.* 26:365-375
- Ranik, M. and Myburg, A. A. (2006) Six new cellulose synthase genes from *Eucalyptus* are associated with primary and secondary cell wall biosynthesis. *Tree Physiol.* 26:545-56
- Rastogi, R., Bate, N. J., Sivasankar, S. and Rothstein, S. J. (1997) Footprinting of the spinach nitrite reductase gene promoter reveals the preservation of nitrate regulatory elements between fungi and higher plants. *Plant Mol. Biol.* 34:465-476
- Reese, M. G., Hartzell, G., Harris, N. L., Ohler, U., Abril, J. F. and Lewis, S. E. (2000) Genome annotation assessment in *Drosophila melanogaster*. *Genome Res.* 10:483-501
- Richmond, T. A. and Somerville, C. R. (2000) The cellulose synthase superfamily. *Plant Physiol.* 124:495-498
- Roberts, K. and McCann, M. C. (2000) Xylogenesis: the birth of a corpse. *Curr. Opin. Plant Biol.* 3:517-522
- Rombauts, S., Florquin, K., Lescot, M., Marchal, K., Rouze, P. and Van de Peer, Y. (2003) Computational approaches to identify promoters and cis-regulatory elements in plant genomes. *Plant Physiol.* 132:1162-1176
- Rose, J. K. C., Lee, H. H. and Bennett, A. B. (1997) Expression of a divergent expansin gene is fruit-specific and ripening-regulated. *PNAS* 94:5955-5960
- Samuels, A. L., Kaneda, M. and Rensing, K. H. (2006) The cell biology of wood formation: from cambial divisions to mature secondary xylem *Can. J. Bot.* 84:631-639

- 
- Samuga, A. and Joshi, C. P. (2004) The *PtrCesA2* gene from aspen xylem is orthologous to *Arabidopsis AtCESA7* (IRX5) gene associated with secondary cell wall synthesis. *Gene* 296:37-44
- Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W. W. and Lenhard, B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucl. Acids Res.* 32:91-94
- Sawant, S. V., Kiran, K., Mehrotra, R., Chaturvedi, C. P., Ansari, S. A., Singh, P., Lodhi, N. and Tuli, R. (2005) A variety of synergistic and antagonistic interactions mediated by cis-acting DNA motifs regulate gene expression in plant cells and modulate stability of the transcription complex formed on a basal promoter. *J. Exp. Bot.* 56:2345-2353
- Scarpella, E., Boot, K. J. M., Rueb, S. and Meijer, A. H. (2002) The procambium specification gene *OSHOX1* promotes polar auxin transport capacity and reduces its sensitivity toward inhibition. *Plant Physiol.* 130:1349-1360
- Schelling, D. and Jones, G. (1995) Functional identification of the transcription start site and the core promoter of the juvenile hormone esterase gene in *Trichoplusia*. *Biochem. Biophys. Res. Commun.* 214:286-294
- Scherf, M., Klingenhoff, A. and Werner, T. (2000) Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: A novel context analysis approach. *J. Mol. Biol.* 297:599-606
- Schrader, J., Moyle, R., Bhalerao, R., Hertzberg, M., Lundeberg, J., Nilsson, P. and Bhalerao, R. P. (2004) Cambial meristem dormancy in trees involves extensive remodelling of the transcriptome. *Plant J.* 40:173-187
- Sessa, G., Steindler, C., Morelli, G. and Ruberti, I. (1998) The *Arabidopsis Athb-8, -9* and *-14* genes are members of a small gene family coding for highly related HD-ZIP proteins. *Plant Mol. Biol.* 38:609-22
- Shahmuradov, I. A., Gammerman, A. J., Hancock, J. M., Bramley, P. M. and Solovyev, V. V. (2003) PlantProm: a database of plant promoter sequences. *Nucl. Acids Res.* 31:114-117
- Shahmuradov, I. A., Solovyev, V. V. and Gammerman, A. J. (2005) Plant promoter prediction with confidence estimation. *Nucl. Acids Res.* 33:1069-1076
- Siebert, P. D., Chenchik, A., Kellogg, D. E., Lukyanov, K. A. and Lukyanov, S. A. (1995) An improved PCR method for walking in uncloned genomic DNA. *Nucl. Acids Res.* 23:1087-1088
- Smale, S. T. (1997) Transcription initiation from TATA-less promoters within eukaryotic protein-coding genes. *Biochim Biophys Acta* 1351:73-88
- Solovyev, V. V. and Shahmuradov, I. A. (2003) PromH: Promoters identification using orthologous genomic sequences. *Nucl. Acids Res.* 31:3540-3545
- Sterky, F., Regan, S., Karlsson, J., Hertzberg, M., Rohde, A., Holmberg, A., Amini, B., Bhalerao, R., Larsson, M., Villarroel, R., et al. (1998) Gene discovery in the wood-



- forming tissues of poplar: Analysis of 5,692 expressed sequence tags. PNAS 95:13330-13335
- Stracke, R., Werber, M. and Weisshaar, B. (2001) The R2R3-MYB gene family in *Arabidopsis thaliana*. Curr. Opin. Plant Biol. 4:447-456
- Svejstrup, J. Q. (2004) The RNA polymerase II transcription cycle: cycling through chromatin. Biochim Biophys Acta 1677:64-73
- Szyjanowicz, P. M. J., McKinnon, I., Taylor, N. G., Gardiner, J., Jarvis, M. C. and Turner, S. R. (2004) The irregular xylem 2 mutant is an allele of KORRIGAN that affects the secondary cell wall of *Arabidopsis thaliana*. Plant J. 37:730-740
- Taoka, K., Kaya, H., Nakayama, T., Araki, T., Meshi, T. and Iwabuchi, M. (1999) Identification of three kinds of mutually related composite elements conferring S phase-specific transcriptional activation. Plant J. 18:611-23
- Tatematsu, K., Ward, S., Leyser, O., Kamiya, Y. and Nambara, E. (2005) Identification of cis-elements that regulate gene expression during initiation of axillary bud outgrowth in *Arabidopsis*. Plant Physiol. 138:757-766
- Taylor, N. G., Howells, R. M., Huttly, A. K., Vickers, K. and Turner, S. R. (2003) Interactions among three distinct CESA proteins essential for cellulose synthesis. PNAS 100:1450-1455
- Taylor, N. G., Laurie, S. and Turner, S. R. (2000) Multiple cellulose synthase catalytic subunits are required for cellulose synthesis in *Arabidopsis*. Plant Cell 12:2529-2539
- Taylor, N. G., Scheible, W. R., Cutler, S., Somerville, C. R. and Turner, S. R. (1999) The irregular xylem3 locus of *Arabidopsis* encodes a cellulose synthase required for secondary cell wall synthesis. Plant Cell 11:769-779
- Theissen, G. (2001) Development of floral organ identity: stories from the MADS house. Curr. Opin. Plant Biol. 4:75-85
- Thijs, G., Lescot, M., Marchal, K., Rombauts, S., De Moor, B., Rouze, P. and Moreau, Y. (2001) A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. Bioinformatics 17:1113-1122
- Thijs, G., Moreau, Y., De Smet, F., Mathys, J., Lescot, M., Rombauts, S., Rouze, P., De Moor, B. and Marchal, K. (2002) INCLUSive: INtegrated clustering, upstream of sequence retrieval and motif sampling. Bioinformatics 18:331-332
- Tjian, R. and Maniatis, T. (1994) Transcriptional activation: a complex puzzle with a few easy pieces. Cell 77:5-8
- Tompa, M., Li, N., Bailey, T. L., Church, G. M., De Moor, B., Eskin, E., Favorov, A. V., Frith, M. C., Fu, Y. T., Kent, W. J. et al. (2005) Assessing computational tools for the discovery of transcription factor binding sites. Nature Biotechnol. 23:137-144

- Torres-Schumann, S., Ringli, C., Heierli, D., Amrhein, N. and Keller, B. (1996) In vitro binding of the tomato bZIP transcriptional activator VSF-1 to a regulatory element that controls xylem-specific gene expression. *Plant J.* 9:283-96
- Trindade, L. M., Horvath, B., Bachem, C., Jacobsen, E. and Visser, R. G. F. (2003) Isolation and functional characterization of a stolon specific promoter from potato (*Solanum tuberosum L.*). *Gene* 303:77-87
- Trindade, L. M., Horvath, B. M., van Berloo, R. and Visser, R. G. F. (2004) Analysis of genes differentially expressed during potato tuber life cycle and isolation of their promoter regions. *Plant Sci.* 166:423-433
- Tsunoda, T. and Takagi, T. (1998) Automatic extraction of position specific co-occurrence of transcription factor bindings on promoters. *Pac. Symp. Biocomput.* pp 252-263
- Turner, S. R. and Somerville, C. R. (1997) Collapsed xylem phenotype of *Arabidopsis* identifies mutants deficient in cellulose deposition in the secondary cell wall. *Plant Cell* 9:689-701
- Tuskan, G. A., DiFazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A. et al. (2006) The genome of black cottonwood, *populus trichocarpa* (torr. & gray). *Science* 313:1596-1604
- Ulmasov, T., Hagen, G. and Guilfoyle, T. J. (1999) Activation and repression of transcription by auxin-response factors. *PNAS* 96:5844-5849
- van Doorn, W. G. (2005) Plant programmed cell death and the point of no return. *Trend Plant Sci.* 10:478-483
- van Helden, J., Andre, B. and Collado-Vides, J. (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies *J. Mol. Biol.* 281:827-842
- van Helden, J., Rios, A. F. and Collado-Vides, J. (2000) Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. *Nucl. Acids Res.* 28:1808-1818
- Vinson, C., Acharya, A. and Taparowsky, E. J. (2006) Deciphering B-ZIP transcription factor interactions in vitro and in vivo. *Biochim Biophys Acta* 1759:4-12
- Vom Endt, D., Kijne, J. W. and Memelink, J. (2002) Transcription factors controlling plant secondary metabolism: what regulates the regulators? *Phytochemistry* 61:107-114
- <sup>a</sup>Wade, P. A. (2001) Methyl CpG-binding proteins and transcriptional repression. *Bioessays* 23:1131-1137
- <sup>b</sup>Wade, P. A. (2001) Methyl CpG binding proteins: coupling chromatin architecture to gene regulation. *Oncogene* 20:3166-3173
- Waibel, A. H., Hanazawa, T., Hinton, G. E., Shikano, K. and Lang, K. J. (1989) Phoneme recognition using time-delay neural networks. *IEEE Trans Acoustic Speech Signal Process* 37:328-339

- Wallace, A. C., Borkakoti, N. and Thornton, J. M. (1997) TESS: A geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites *Protein Sci.* 6:2308-2323
- Weisshaar, B. and Jenkins, G. I. (1998) Phenylpropanoid biosynthesis and its regulation. *Curr. Opin. Plant Biol.* 1:251-257
- Williams, C. G. and Neale, D. B. (1992) Conifer wood quality and marker-aided selection: a case-study. *Can. J. of Forest Res* 22:1009-1017
- Wingender, E. (2000) The TRANSFAC system on gene regulation. *Trends Glycosci. Glycotechnol.* 12:255-264
- Wolffe, A. P. (1995) RNA Polymerase III transcription. *Trends Genet.* 11:32
- Wu, A. M., Ling, C. and Liu, J. Y. (2006) Isolation of a cotton reversibly glycosylated polypeptide (*GhRGPI*) promoter and its expression activity in transgenic tobacco. *J Plant Physiol.* 163:426-35
- Yamamoto, R., Demura, T. and Fukuda, H. (1997) Brassinosteroids induce entry into the final stage of tracheary element differentiation in cultured *Zinnia* cells. *Plant Cell Physiol.* 38:980-983
- Yamamoto, R., Fujioka, S., Demura, T., Takatsuto, S., Yoshida, S. and Fukuda, H. (2001) Brassinosteroid levels increase drastically prior to morphogenesis of tracheary elements. *Plant Physiol.* 125:556-563
- Yang, X. H., Xu, Z. H. and Xue, H. W. (2005) *Arabidopsis* membrane steroid binding protein 1 is involved in inhibition of cell elongation. *Plant Cell* 17:116-31
- Yin, Y. and Beachy, R. N. (1995) The regulatory regions of the rice tungro bacilliform virus promoter and interacting nuclear factors in rice (*Oryza sativa L.*). *Plant J.* 7:969-80
- Yin, Y., Chen L., and Beachy, R. (1997) Promoter elements required for phloem-specific gene expression from the RTBV promoter in rice. *Plant J.* 12:1179-88
- Zemach, A. and Grafi, G. (2003) Characterization of *Arabidopsis thaliana* methyl-CpG-binding domain (MBD) proteins. *Plant J.* 34:565-572
- Zhang, M. Q. (1998) Identification of human gene core promoters *in silico*. *Genome Res.* 8:319-326
- Zhao, C. S., Johnson, B. J., Kositsup, B. and Beers, E. P. (2000) Exploiting secondary growth in *Arabidopsis*: construction of xylem and bark cDNA libraries and cloning of three xylem endopeptidases. *Plant Physiol.* 123:1185-1196
- Zhong, R. Q., Taylor, J. J. and Ye, Z. H. (1997) Disruption of interfascicular fiber differentiation in an *Arabidopsis* mutant. *Plant Cell* 9:2159-2170
- Zhu, J. and Zhang, M. (1999) SCPD: a promoter database of the yeast *Saccharomyces cerevisiae*. *Bioinformatics* 15:607-611

## Chapter 2

### Isolation and Sequence Characterization of Promoter

### Regions of Six Cellulose Synthase Genes in

### *Eucalyptus grandis*

N.M. Creux<sup>1</sup>, M. Ranik<sup>1</sup>, D.K. Berger<sup>2</sup>, A.A. Myburg<sup>1</sup>

<sup>1</sup>*Department of Genetics, Faculty of Natural and Agricultural Sciences, Forestry Agricultural Biotechnology Institute (FABI), University of Pretoria, Pretoria, 2000, South Africa*

<sup>2</sup>*Department of Botany, Faculty of Natural and Agricultural Sciences, Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Pretoria, 2000, South Africa*

This chapter has been prepared in the format of a manuscript for a research journal (e.g. Plant Molecular Biology). I performed the majority of the genome walking, cloning and sequencing. I performed all of the data analyses and prepared this manuscript. M. Ranik provided the *Eucalyptus* mRNA *CesA* sequences, aided in the genome walking of *EgCesA2* and *EgCesA7* and guided me with the planning of the work. Prof. A.A. Myburg and Prof. D. Berger provided advice, direction and supervision in the planning of the project. They also provided direction in the interpretation of the results and provided critical revision of the manuscript. All other technical assistance is acknowledged at the end of the chapter.

## 2.1 Abstract

Wood formation is a developmentally regulated process comprised of a number of phases, one of which is secondary cell wall formation. During secondary cell wall formation, large amounts of cellulose are deposited on the outside of the plasma membrane. Cellulose is produced and deposited by a membrane bound enzyme complex. This complex comprises multiple cellulose synthase (CESA) proteins, encoded by a family of cellulose synthase (*CesA*) genes. *CesA* genes have been isolated from a number of plants including members of the genera *Eucalyptus*, *Populus* and *Arabidopsis*. Expression studies and mutant phenotypes of the *CesA* genes suggest that at least six of these genes are required for cell wall formation and that they can be divided into two distinct groups based on their expression patterns. A specific set of three *CesA* genes are expressed during primary cell wall formation and three other *CesA* genes are expressed during secondary cell wall formation. The regulatory mechanisms that underlie the unique expression patterns of the two groups of *CesA* genes are poorly understood. In this study six *Eucalyptus CesA* gene promoters (*EgCesA1-5* and 7) were isolated using genome walking. The *Eucalyptus* promoter regions and the orthologous promoter regions from *Populus* and *Arabidopsis* were analysed using TSSP (Transcriptional start site plant promoter prediction) and NNPP (Neural network promoter prediction) software packages. The *in silico* results were compared among species and it was found that the predicted transcriptional start sites and the core elements of the *CesA* gene promoters showed some structural conservation. The isolation and basic characterization of the promoters of the *CesA* gene family in *Eucalyptus* paves the way for detailed *in silico* and molecular characterization of promoter elements in these important vascular-specific proteins (Chapter 3).

## 2.2 Introduction

Xylogenesis is the process of wood formation and is characterised by four main steps namely, cell division, cell differentiation and elongation, cell wall formation and finally programmed cell death (Roberts and McCann 2000). Each of these processes involves complex transcriptional regulation that requires spatial and temporal specificity. Many of the key genes involved in xylogenesis have been identified by gene expression studies using techniques such as cDNA-AFLP (amplified fragment length polymorphism) analysis (Miloni et al. 2001; Prassinis et al. 2005; Ranik et al. 2006), microarray analysis (Hertzberg et al. 2001; Demura et al. 2002; Yang et al. 2003) and EST (Expressed sequence tag) sequencing (Allona et al. 1998; Sterky et al. 1998; Andersson et al. 2004). Although a large amount of data on the expression patterns of genes involved in wood formation is now available, little is known about the regulatory networks and gene regulatory regions that result in the observed expression patterns

Cellulose is the most abundant biopolymer on earth and therefore cellulose biosynthesis is an important aspect of wood formation. A complex of membrane bound enzymes known as cellulose synthases (CESA) facilitate cellulose biosynthesis. This complex is comprised of six catalytic subunits arranged in a rosette-like conformation and each catalytic subunit contains six CESA proteins (Brown and Saxena 2000). The individual CESA proteins add activated glucose molecules to the ends of the growing cellulose chains to form cellulose polymers. The polymers are associated through hydrogen bonding to produce cellulose microfibrils (Saxena and Brown 2005), which are laid down on the outside of the plasma membrane (Kerstens and Verbelen 2003) to strengthen the cell walls.

Cellulose synthase (*CesA*) genes have been isolated from a number of different plants including *Arabidopsis* (Richmond and Somerville 2000), barley (Burton et al. 2004), rice (Tanaka et al. 2003), *Populus* (Joshi et al. 2004) and *Eucalyptus* (Ranik and Myburg 2006). The *Arabidopsis thaliana* genome contains 10 cellulose synthase genes and six (*AtCesA1-4*, *AtCesA7* and *AtCesA8*) of these genes play major roles in plant cell wall formation (Taylor et al. 2000; Taylor et al. 2003), which also appears to be true for the *CesA* genes isolated from other plants (Tanaka et al. 2003; Appenzeller et al. 2004; Burton et al. 2004; Djerbi et al. 2004; Ranik and Myburg 2006). When the *CesA* genes are clustered by expression pattern they fall into two distinct groups, those associated with primary cell wall formation and those associated with secondary cell wall formation. Numerous expression studies have confirmed the association of separate *CesA* gene products in primary and secondary cell wall formation (Turner and Somerville 1997; Taylor et al. 2000; Burn et al. 2002; Ranik and Myburg 2006). The phylogenetic structure of the *CesA* gene tree in higher plants also confirms the early differentiation of the gene family into at least six clades (Ranik and Myburg 2006).

The 5'UTR (five prime untranslated region) is the transcribed region of DNA immediately upstream of the translational start codon and has been shown to play an integral role in gene expression. In general, the 5'UTR regulates the gene at a translational level, bound by translational proteins, but it also plays a role in the transcriptional regulation of some genes. Genes may have more than one transcriptional start site and this could cause varying lengths of 5'UTR (Hughes 2006). Shorter 5'UTRs are often found in conjunction with higher gene expression (Molina and Grotewold 2005), and multiple TSSs (transcriptional start site) can thus produce different levels of expression in different tissues. The 5'UTR may also

contain important motifs for transcriptional gene regulation (Mingam et al. 2004). A number of 5'UTRs contain introns that house regulatory motifs. Mutation of these motifs may alter the expression of the genes they regulate (Chen et al. 2002). Alternate splicing of the introns in the 5'UTR or first introns in the gene may also play a role in producing alternate expression patterns (Loke et al. 2005; Jeong et al. 2006).

Despite the growing number of *CesA* expression studies, the regulatory mechanisms that direct their expression patterns are poorly understood. The TSS is the point at which transcription of a gene begins. In general, transcription commences when the RNA polymerase II binds to the motifs just upstream of the TSS (Svejstrup 2004; Sawant et al. 2005). Important initiation motifs located in the proximal promoter (core promoter) region of a gene such as the TATA-box or Initiator sequence (Inr) are vital for gene function. The TATA-box is an AT-rich region usually found approximately 25-50 bp upstream of the TSS and plays an important role in the initiation of transcription (Burley and Roeder 1996). The TATA-box is not present in all of the promoters of the *Arabidopsis* genome, only 29% of the promoters analysed by Molina and Grotewold (2005) contained a TATA-box.

The TATA-less promoters rely on elements such as the Inr to initiate transcription (Smale 1997). It has been suggested that the promoters of genes that require highly specific expression patterns are more likely to contain a TATA-box than those with a more constitutive expression pattern (Molina and Grotewold 2005). It has also been shown in *Arabidopsis* that TATA-less promoters often have longer 5'UTRs, which may contain other important regulatory features (Molina and Grotewold 2005). A DPE (Downstream promoter element) motif may be located 50 bp downstream of the



TSS and functions in conjunction with the Inr to initiate transcription, often in the absence of the TATA-box (Kutach and Kadonaga 2000). These variations in core initiator elements make it difficult to create a single accurate plant promoter model.

With the large amounts of DNA sequence now available, it has become important to find fast and effective ways of analysing the promoter regions of genes and identifying the proteins, which bind to them. Current functional techniques that can be used to identify the TSS and core promoter elements (Farfsing et al. 2005) are very accurate, but may take months to identify a single functional TSS. A less accurate, but faster method is to use an algorithm that compares the promoter region of interest to a dataset it has already been trained on, to predict possible TSS and core promoter elements. There are a number of software tools that can aid in this approach, but most of these are mammal/human orientated and are less accurate when used for TSS prediction in plants (Rombauts et al. 2003).

When using an *in silico* approach for the prediction of TSSs and core promoter elements (TATA-box and Inr), multiple algorithms should be used in order to increase the accuracy of the results. The only plant-specific tool for TSS prediction is the Transcriptional Start Site Plant promoter prediction (TSSP) algorithm (Shahmuradov et al. 2005). This algorithm was trained on 307 experimentally characterised plant promoters located in plantpromDB (Shahmuradov et al. 2003: [www.softberry.com](http://www.softberry.com)). The dragon promoter finder was originally designed on human promoters (Mohanty et al. 2005) but, subsequently, the motif identification module has been used in *Arabidopsis* (Mohanty et al. 2005). Another software tool that can be used in plants is NNPP (neural network promoter prediction), which uses a neural network algorithm to identify the transcriptional start site of eukaryote promoters (Reese et al. 2000;

Reese 2001). NNPP has been trained on *Drosophila* promoters and may not be as accurate as TSSP when used in plants. There are many software packages available for TSS identification but there is a lack in plant specific software packages (for review see Chapter 1).

Even with the ever-increasing data on plant regulatory mechanisms (Chapter 1), there is still no information on the core regulatory mechanisms behind the unique expression patterns of the *CesA* genes. Cellulose biosynthesis emerged early in plant evolution (Nobles et al. 2001) and a number of conserved domains still occur in the *CesA* genes of distantly related plant species. Whether this conservation among distantly related species is carried through to the regulatory mechanisms of the genes is not clear. The aim of this study was to isolate and characterize the core promoters of the primary and secondary cell wall associated *CesA* genes in *Eucalyptus*. This was achieved through genome walking, *in silico* predictions of core promoter features and comparative analysis with the putative orthologous promoters in *Arabidopsis* and *Populus*. This chapter reports the isolation of six *Eucalyptus CesA* promoter regions and the identification of the core promoter elements (TSS and transcriptional initiation sequences). The following chapter (Chapter 3) discusses a detailed comparative bioinformatics analysis of the cis-regulatory elements involved in the regulation of the *CesA* genes.

## 2.3 Materials and Methods

### *Plant material*

Leaf samples used for DNA isolation were obtained from a pure-species *E. grandis* clone (TAG14) provided by Mondi Business Paper South Africa. The leaf samples

were submerged in liquid N<sub>2</sub> in the field, transported on dry ice and stored at -80°C until DNA isolation.

#### *DNA isolation*

Genomic DNA was isolated using a CTAB (Cetyltrimethylammonium Bromide) method (Doyle and Doyle 1987). To verify that high quality DNA was isolated, the DNA samples were resolved on a 1% agarose gel. To test the purity of the DNA, a test digestion was performed as described in the instruction manual of the Universal Genome Walker kit (Clontech, Palo Alto, CA.).

#### *Genome walking library construction*

Two sets of genome walking libraries were used. The first library panel was constructed using the Universal Genome Walker kit (Clontech) and the second library panel was produced using the protocol developed by Seibert et al. (1995). The Universal Genome Walker kit provided four restriction enzymes (PvuII, EcoRV, DraI, StuI) for library construction. The second library panel added three restriction enzymes (HindIII, SmaI, XbaI) to the panel of genome walking libraries. The different libraries offered more variety in genome walking products and increased the probability of obtaining a suitable genome walking product.

#### *Genome walking*

Gene-specific primers (Table 2.1) were designed on the DNA sequences described in Ranik and Myburg (2006), or on sequences obtained from previous genome walking steps. These primers and adaptor-specific primers were used in primary and secondary PCRs as described by the kit instruction manual and the protocol by Siebert et al.

(1995). All of the PCR reactions were performed using 0.8 units of Supertherm Excel polymerase (SR Product, Kent, UK), which processes proofreading capacity. Standard dNTP (MBI Fermentas, Hanover, MD) concentrations (0.2 mM of each base) were used and no magnesium was added since Exsel *Taq* polymerase buffer (1x) contains magnesium at the correct concentrations for the reaction. The thermal cycling conditions for the primary genome walking PCR consisted of seven cycles of 94°C for 2 seconds and 72°C for 3 minutes, followed by 32 cycles of 94°C for 2 seconds and 67°C for 3 minutes. This was followed by a final elongation step of 67°C for 4 minutes. The secondary PCR conditions were similar, but had fewer cycles beginning with five cycles of 94°C for 2 seconds and 72°C for 3 minutes, followed by 20 cycles of 94°C for 2 seconds and 67°C for 3 minutes and ended with a single elongation step of 67°C for 4 minutes.

The products of the secondary PCR were resolved on 1% agarose gels and the largest fragments selected for cloning with the InsT/A clone<sup>TM</sup> PCR Product Cloning Kit (MBI Fermentas, Hanover, MD). Positive bacterial colonies were identified using colony PCR with standard M13 vector primers (Forward: 5'-CACGACGTTGTAAAACGAC-3' and Reverse: 5'-GGAAACAGCTATGACCATG-3'). The colonies were eluted in 5 µl of dH<sub>2</sub>O and heated to 95°C for five minutes to burst the bacterial cells. Following a brief centrifugation 5 µl of the supernatant was used as the DNA template for the PRC reaction with the dNTPs at a final concentration of 0.20 mM each, 1x Exsel buffer containing magnesium, 0.8 U Exsel *Taq* and 0.4 µM of each primer. The thermal cycling conditions began with a denaturation step of 94°C for 20 seconds followed by an annealing step of 53°C for 30 seconds and an elongation step of 72°C for two

minutes. These three steps were repeated in 30 cycles and a final extension step of 72°C for 10 minutes was performed. Plasmid DNA was extracted from positive colonies using the QIAGEN miniprep kit (QIAGEN GmbH, In Germany). The inserts were sequenced using BigDye terminator chemistry (Applied Biosystems, Foster City, CA) on an ABI3100 automated DNA sequencer (Applied Biosystems) initially using the standard M13 primers. The larger fragments were sequenced using insert-specific primers (Table 2.2) in order to cover the whole cloned region. Sequences were aligned in ContigExpress (Vector NTI, Invitrogen) for DNA sequence analysis.

#### *End-to-end amplification of genome walking contigs*

The contigs obtained from genome walking were used for the design of a forward and a reverse primer spanning the full-length of the contig (Table 2.3). The template of the amplification was undigested *E. grandis* genomic DNA (5 ng/μl). Exsel *Taq* polymerase (SR Product, Kent, UK), 1 x Exsel buffer containing MgCl<sub>2</sub> and dNTPs at a final concentration of 0.20 mM each (MBI Fermentas, Hanover, MD) were used for the amplification. The PCR conditions were as follows: one cycle of 2 min at 94°C; 30 cycles of 30 sec at 94°C, 30 sec at 56°C and 2 min at 72°C and a final elongation step of 10 min at 72°C. The fragments produced by this reaction were resolved on a 1% agarose gel, purified with the Qiaquick PCR purification kit (QIAGEN, Germany) and cloned (InsT/A clone™ PCR Product Cloning Kit). Three positive bacterial colonies for each promoter were selected for plasmid extraction and sequenced for sequence verification of the promoter DNA sequence using the standard M13 primers and internal fragment-specific primers (Table 2.2). The DNA sequences were aligned in Vector NTI's ContigExpress module (Invitrogen) with the original sequence assembled from genome walking products. The four sequences were compared and a

consensus DNA sequence was compiled for each promoter from positions where they were the same in at least 3 of the 4 sequences.

#### *In silico identification of transcriptional start sites*

The *in silico* prediction of the transcriptional start site (TSS) of all of the promoter regions was performed using two online software programs. The first program, TSSP (Shahmuradov et al. 2005) is located at [www.softberry.com](http://www.softberry.com). This is a plant-specific program trained on plant promoters located in plantpromDB (Shahmuradov et al. 2003). The second program located at [http://www.fruitfly.org/seq\\_tools](http://www.fruitfly.org/seq_tools) is Neural Network Promoter Prediction or NNPP (Reese et al. 2000; Reese 2001). NNPP predicts transcriptional start sites for eukaryotes and is trained on *Drosophila* promoters. Both programs were used to analyse the seven *Eucalyptus* promoter sequences (*EgCesA1-4*, *EgCesA5A*, *EgCesA5B* and *EgCesA7*).

NNPP and TSSP were also used to predict the transcriptional start sites in the orthologous promoter sequences for *Arabidopsis* obtained from The *Arabidopsis* Information Resource (TAIR) (Huala et al. 2001), <http://arabidopsis.org/> (AT5G44030.1, AT4G18780.1, AT5G17420.1, AT5G05170.1, AT4G32410.1, AT5G64740.1) and *Populus* obtained from <http://genome.jgi-psf.org/poptr1/poptr1.home.html>. The poplar sequences were obtained by first downloading the Poplar *CesA* cDNA sequences from NCBI (Accession numbers: AF072131.1, AY095297.1, AF527387.1, AY162181.1, AY055724.2 and AY196961.1) and then aligning them to the poplar genome and retrieving the 2 kb sequence upstream of the start codon. The default settings of each program were used for the prediction of the TSSs. In most cases, TSSP was taken to be more accurate

than NNPP as it was trained on plant promoters, but in some cases TSSP could not predict a TSS and only NNPP predicted a TSS.

## 2.4 Results

### *Genome walking*

Primary genome walking PCRs were done using the different genome walking libraries as templates and with the adaptor-specific outer primers with the gene-specific outer primers (Table 2.1). The secondary genome walking PCRs was performed using the inner adaptor primer from the kit and the gene-specific inner primers (Table 2.1). The products from this PCR were resolved on a 1% agarose gel and a number of bands were present for each gene (e.g. Figure 2.2). For each promoter, the largest fragment produced was cloned and sequenced. Genome walking products were successfully isolated for the upstream regions of *EgCesA1-5* and 7 producing fragments 1-2 kb (Table 2.4 and Figure 2.3).

In recent research carried out by our group it was found that *EgCesA6* exhibited very low expression in all of the tissues evaluated by Ranik and Myburg (2006). A seventh *CesA* gene (*EgCesA7*) was subsequently isolated from *Eucalyptus grandis* (Unpublished results M. Ranik, J. Bradfield and A.A. Myburg), which exhibited an expression pattern consistent with being one of the primary cell wall associated *CesA* genes and it was also the putative ortholog of *PtrCesA7* (Figure 2.1). This gene was therefore included in the genome walking rather than *EgCesA6*. The genome walking of *EgCesA7* produced an 800 bp fragment in one genome walk (Figure 2.3 and Table 2.4).

### *End-to-end amplification of the genome walking contig*

The genome walking products of the six *CesA* genes were each approximately 1-2 kb in length upstream from the start codon. Contigs of the genome walking products were built (Figure 2.3 and Table 2.4) and, forward and reverse primers (Table 2.3) were used to amplify the region spanning the contig (Figure 2.4). *EgCesA1-5* produced fragments of the expected sizes and ranged between 1 kb and 2 kb (Figure 2.4: arrow heads indicate the bands). The shortest promoter region obtained was *EgCesA7* promoter (787 bp) and the longest promoter fragment amplified was 2000 bp of *EgCesA1* promoter.

When the *EgCesA5* promoter was amplified from *E. grandis* genomic DNA two fragments were obtained (Figure 2.3). The larger fragment (*EgCesA5A*) was of the expected size (1569 bp), but the second fragment, *EgCesA5B*, was 196 bp smaller (Figure 2.4). The *EgCesA3* and 5 promoters were amplified from *E. grandis* DNA and from hybrid *E. grandis x E. nitens* DNA in order to test whether the promoter regions could be amplified from different *Eucalyptus* species using the same primers. There was no amplification of *EgCesA3* promoter from *E. grandis x E. nitens* but, the amplification of *EgCesA5* from *E. grandis x E. nitens* was successful, suggesting there may well be differences among the different species (Figure 2.4). The double band produced by the amplification of *EgCesA5* from *E. grandis* pure species also suggested that there may be significant differences between different alleles within a species. To minimize PCR errors a DNA polymerase with proofreading capabilities was used in all the amplifications. Only the fragments indicated by the arrows, amplified from *E. grandis* pure species, were cloned and sequenced for further analysis (Figure 2.4).



A consensus sequence (Appendix 1) was constructed by comparing the independent clone sequences by including sequence that was the same in at least three of the four sequences. In most cases the four DNA sequences were highly similar. The two putative allelic fragments identified for the *CesA5* promoter were isolated individually from the gel and then cloned separately and sequenced. After sequence analysis of three clones from each fragment, it was clear that the two *EgCesA5* fragments had a high sequence similarity, but the smaller fragment (*EgCesA5B*) contained a 196 bp deletion (Figure 2.5, Group B).

#### *In silico identification of the transcriptional start sites*

The promoter sequences (Appendix 1) used for the *in silico* identification of the transcriptional start sites (TSSs) were the *EgCesA 1 - 4, 7, 5A, and 5B* promoter regions (*5A* is the larger fragment and *5B* is the promoter region with the 196 bp deletion). The orthologs (Figure 2.1) of the *EgCesA* promoters were also obtained from *Populus* and *Arabidopsis* as described in the previous section. The TSSs of the *Arabidopsis CesA* genes were previously computationally predicted and by actual 5'UTRs from cDNA which, are available on TAIR. These were compared to the outputs of NNPP and TSSP. A number of *Arabidopsis* TSSs have been determined experimentally but the TSSs provided by TAIR for the *CesA* genes were not among these, but were predicted using computer models. The *AtCesA* TSSs available on TAIR may be more accurate than TSSP and NNPP as there is more information available on *Arabidopsis* genes and so more accurate models can be produced (Alexandrov et al. 2006).

The data provided on TAIR and the TSS predictions by TSSP and NNPP did not correlate well. NNPP predicted a similar (within 20 bp) TSS to that on TAIR in only

two of the six promoters (*AtCesA1* and 3). TSSP only predicted a similar TSS position in one of the six cases (*AtCesA8*). In one case, *AtCesA4*, NNPP and TSSP predicted a similar TSS, but this was not the same as the TAIR position (Table 2.6). These results confirm that transcriptional start site prediction is very complex and not enough is understood about plant promoters for accurate models to be produced.

NNPP and TSSP were used to predict the TSS of the *Eucalyptus* and *Populus CesA* promoter regions (Table 2.7). Once again, the two program outputs did not correlate with NNPP and TSSP predicting similar TSSs for only two of the six *Populus* promoters (*PtrCesA3* and 7) and two out of the six *Eucalyptus* promoters (*EgCesA4* and 7). It must also be noted that TSSP did not predict a TSS for 3 of the six *Populus* promoters indicating it may not be well suited to the poplar TSS structure and NNPP did not predict TSSs for two of the six *Eucalyptus* promoters (Table 2.7).

Although the results of the two software packages used for TSS prediction agree with each other directly, when the TSS positions were compared among the orthologs (Figure 2.1) some conservation was apparent even when the TSS was predicted by different software packages (Table 2.7 and Figure 2.5). The most striking example of this is in orthologous Group F where the TSSs of the *Arabidopsis*, *Populus* and *Eucalyptus* orthologs are within ten bp of each other, but each was predicted with a different tool (Figure 2.4). The *Eucalyptus* TSS was predicted by TSSP (Table 2.7), the *Populus* TSS was predicted by NNPP (Table 2.7) and the *Arabidopsis* result was obtained from TAIR (Table 2.6).

Another example of conservation of the TSS is represented in orthologous Group A, which contains *EgCesA4*, *PtrCesA5* and *AtCesA3*. All three upstream regions contained an intron in their 5'UTRs (Figure 2.5, Group A). The length of the 5'UTR

including the intron was of a similar size in *Populus* (643 bp) and in *Eucalyptus* (734 bp) while the *Arabidopsis* 5'UTR was smaller (292 bp). The position of the intron also seemed to be conserved among *Eucalyptus* and *Populus*. The intron occurs 158 bp upstream of the start codon in *Eucalyptus* and 156 bp in *Populus*, but this was not the case in *Arabidopsis* where the intron started 56 bp from the start codon. This is perhaps a reflection of the fact that poplar and (presumably) *Eucalyptus grandis* clones have undergone significant expansion (Tuskan et al. 2006).

In orthologous group D (Figure 2.5) the *Populus* and *Eucalyptus* 5'UTRs showed some conservation having only an 8 bp difference in size (218 bp and 210 bp respectively) and again the *Arabidopsis* 5'UTR was somewhat shorter (75 bp). Orthologous Group F (*EgCesA3*, *PtrCesA2* and *AtCesA7*) also showed a high level of conservation in 5'UTR length (Figure 2.5, Group F). The *Arabidopsis* and *Eucalyptus* upstream regions had the same predicted 5'UTR length (45 bp) while *Populus* had a slightly smaller 5'UTR of 32 bp (Table 2.7). The other orthologous Groups B, C and E did not show such clear 5'UTR length conservation (Figure 2.5). This may be due to the fact that the *in silico* analysis is still not highly accurate and that there may be errors in the predictions, which will have to be functionally tested using methods such as primer extension (Kainz and Roberts 1992).

#### *In silico prediction of the TATA-Box and initiator sequence*

The TATA-box is an AT-rich region located 25-50 bp upstream of the transcriptional start site in many eukaryotic genes. Of the 19 promoters analysed here, 17 promoters contained a TATA-box just upstream of the TSS and the distance from the TSS appeared to be conserved within the different orthologous promoter groups (Figure 2.5). This conservation may be an indication of the regulatory mechanisms, but could

also be an artefact of the TSS prediction software as this will be one of the main factors taken into account when identifying the TSS.

In orthologous Group A, the *Populus* and *Eucalyptus* TATA-boxes were found at 27 and 23 bp upstream of the predicted TSS and were more conserved than in the *Arabidopsis* ortholog where the TATA-box was 50 bp from the predicted TSS (Figure 2.5, Group A). Orthologous Group B showed little or no conservation in the TATA-box position not even between the two different *EgCesA5* alleles. In *EgCesA5A* the TATA-box was 26 bp upstream of the predicted TSS, while in *EgCesA5B* the only AT-rich region was 97 bp upstream of the TSS. This was also the case for orthologous Group B where the *Arabidopsis* TATA-box was predicted to be only 15 bp from the TSS, while the TATA-box of the *Eucalyptus* ortholog was further away (24 bp) and the *Populus* orthologs TATA-box was even further upstream (47 bp) of the TSS (Figure 2.5, Groups B and C). The orthologous Groups D and E (Figure 2.5) showed more conservation in TATA-box position with all of the TATA-boxes predicted within 50 bp of the TSS.

Interestingly, the Group F orthologs (Figure 2.5) showed a large amount of conservation in the TSS, but there was little or no conservation in the positioning of the TATA-box with the *Eucalyptus* ortholog not possessing a TATA-box, but only an initiator sequence (Inr). The *Populus* and *Arabidopsis* orthologs both had a predicted TATA-box, but the position was not conserved. The *Populus* orthologs TATA-box was 26 bp from the TSS, while the only AT-rich region up stream of the TSS in the *Arabidopsis* promoter was 113 bp upstream of the TSS. Only two *Eucalyptus* sequences (*EgCesA1* and *EgCesA3*) appeared to have no TATA-box upstream of the

predicted TSS, but did appear to have Inr sequences (Py-Py-A-N-(T/A)-Py-Py)(Lo and Smale 1996).

#### *Identification of microsatellites in promoter regions*

During the sequence analysis of the *CesA* promoter regions it was observed that a number of microsatellites were present in the promoter regions of these genes (Figure 2.5). Six promoters contained these repeat regions (Table 2.5). The microsatellites were not conserved among the different species, but perhaps are a feature of *Eucalyptus* promoters, as three of the seven *Eucalyptus* promoters contained a microsatellite. The CT repeats in the *Eucalyptus* promoters were found just downstream of the predicted TSSs (Figure 2.5) and could play a role in gene regulation. One *Arabidopsis* sequence contained a GGT repeat less than 50 bp upstream of the TATA-box (Figure 2.5A) and a *Populus* sequence had an AT repeat element, which may form part of the TATA-box (Figure 2.5, Group C).

## **2.5 Discussion**

During cell wall formation cellulose is deposited on the outside of the plasma membrane to strengthen the cell wall. The regulatory mechanisms underlying cellulose biosynthesis are still poorly understood. In previous studies it has been suggested that six cellulose synthase genes play a major role in the production and deposition of cellulose within the plant wall (Burn et al. 2002; Tanaka et al. 2003; Appenzeller et al. 2004; Burton et al. 2004). Recently, six cellulose synthases (*CesA*) genes were isolated from *Eucalyptus*, two genes were expressed during primary cell wall formation and the second set of three were expressed during secondary cell wall formation (Ranik and Myburg 2006). In this study the promoter regions of the six

*Eucalyptus* cellulose synthase genes were isolated and analysed. *In silico* prediction methods were used to predict the transcriptional start sites (TSSs) and TATA-box positions in the *CesA* promoters. The orthologous promoter regions in *Arabidopsis* and *Populus* were compared with the *Eucalyptus* promoter regions and conservation among the orthologs investigated.

The promoter regions of the *Eucalyptus CesA* genes were obtained via genome walking due to the lack of available genomic sequence for *Eucalyptus*. The target length for genome walking was 1.5 – 2 kb of DNA sequence upstream of the translational start site to ensure adequate upstream sequence for further analysis. After genome walking the upstream regions obtained for *EgCesA1*, 2, 3, 4 and 5, were between 1.1 – 2 kb (Figure 2.3). The *EgCesA7* promoter region obtained from genome walking was 787 bp in length and, although shorter than the target, was deemed sufficient for further analysis.

The genome walking method entailed the isolation of DNA sequences, which were aligned to build a contig. A problem with this approach is that during the genome walking process fragments could be amplified from different regions of the genome such as from different alleles or different gene family members. In order to ensure that the contigs built from the genome walking fragments represent a single locus, the full contig region was amplified with end-to-end PCR from genomic DNA. If the contig is built from non-continuous fragments it will not be amplified as a whole fragment from genomic DNA. Another problem with genome walking the *Eucalyptus CesA* genes was the lack of available genome sequence on which to design genome walking primers, thus primers were designed on cDNA sequences. The *CesA* genes have very short first introns and to avoid designing a primer over an intron/exon

boundary some of the primers had to be in the second exon of the genes. This hindered the genome walking in some cases (e.g *EgCesA2*), as the first intron may be very long and sometimes contained stretches of repetitive DNA, which can cause the amplifications to fail.

#### *Indel and SSR variation in the Cesa promoter regions*

The amplification of the *EgCesA5* promoter region produced two DNA fragments that differed in length (*EgCesA5A* and *EgCesA5B*, Figure 2.4) but had high sequence similarity compared to the original sequence obtained by genome walking. Although no segregation analysis has been performed yet, it is possible that the two length variants represent the promoters of two alleles of *EgCesA5* in *E. grandis*. The *EgCesA5B* promoter fragment contained a 196 bp deletion approximately 570 base pairs from the transcriptional start site (Figure 2.5B). Within the 196 bp region there may be important *cis*-elements that could affect *Cesa* gene expression and have an effect on cellulose production. It has previously been shown that allele-specific mutations within chalcone synthase promoter may lead to differential light responsive expression patterns (de Meaux et al. 2005). The two *EgCesA5* length variants may represent allelic variation in *Cesa* promoters and functional analysis may reveal allelic differences in gene expression and, ultimately, in the cellulose biosynthetic pathway.

Sequence analysis of the promoter regions revealed the presence of microsatellites in a number of the promoter regions. Microsatellites are commonly found dispersed throughout plant genomes. A recent *E. grandis* genome sample sequencing study found that microsatellites occur on average every 12 kb (Lourenço 2004). Here, only

about 10 kb of *E. grandis* genomic DNA was sequenced and five microsatellites were identified (Table 2.5). This may indicate that microsatellites are present at a higher frequency in promoter regions than in the rest of the genome. This was also suggested when a global analysis of microsatellite distribution in plant genomes was performed (Morgante et al. 2002; Morgante 2006). Martin et al. (2005) showed that changes in the length of a microsatellite in a bacterial promoter region resulted in altered gene expression. Another study performed on mice also showed that, as the length of a microsatellite upstream of a gene varied, so the expression levels of the gene changed (Hammock and Young 2005). None of the observed microsatellites were conserved among the *CesA* promoters of *Eucalyptus*, poplar and *Arabidopsis*, but they could play a role in the expression of the individual genes. Microsatellites are often sites of mutation where a high frequency of mutations may occur. These mutations may lead to changes in the regulatory elements or in the distance between regulatory elements, which can affect gene expression (Tompa et al. 2005).

#### *In silico* identification of TSS positions and core promoter elements

It was important to locate the position of the TSS in each of the promoter regions in order to anchor the promoters for positioning of transcription factor binding sites that were identified in Chapter 3. Accurate definition of each TSS will also aid further functional testing of the promoters. Due to the high amount of sequence divergence among promoter sequences of distantly related plants such as *Arabidopsis*, *Eucalyptus* and poplar it was not possible to obtain accurate sequence alignments of orthologous promoter sequences and to analyse patterns of nucleotide evolution. However, the preliminary *in silico* identification of the TSS and TATA-box allowed comparison of the structural components in the core promoter regions of the orthologous promoters of different species.



In order to compare the accuracy of the TSSP and NNPP software, the TSSs of the *Arabidopsis Cesa* genes predicted by the software's were compared to each other and to the TSS data available in the public database TAIR (Table 2.6). The three sets of results differed significantly with only three of the predicted TSSs matching (within 20 bp) what was provided by TAIR (*AtCesA1*, *AtCesA3*, and *AtCesA8*). These results indicate that *in silico* TSS prediction is very difficult and that none of the software packages available to date are highly accurate. It was subsequently decided that the TSSP results would be used for further inferences in this study, because TSSP is based on functionally tested plant promoters and should be more accurate for plants than NNPP. The TSSP results were also selected above the data on TAIR as TSSP is newer software (Shahmuradov et al. 2005) than the software used to predict some of the TSSs on the TAIR database and may incorporate a more detailed promoter model.

TSSP and NNPP were used to predict the TSS and TATA-box positions of 19 promoters, six *Arabidopsis*, six *Populus* and seven *Eucalyptus* promoter regions (*EgCesA5A* and *B* were analysed separately). TSSP and NNPP both predicted a TSS position for 12 of the 19 promoters and of these, five were within 50 bp of each other (Table 2.6 and 2.7). NNPP was unable to predict TSS positions for two promoter regions (*EgCesA2* and 3), while TSSP produced no result for five promoter regions (Table 2.6 and 2.7). Surprisingly, of the five promoter regions not predicted by TSSP, three were poplar promoters, indicating that TSSP may not be an accurate tool for use on poplar promoters. The use of the two types of prediction programs ensured that a TSS was predicted for all of the promoters in the data set. Interestingly, Alexandrov et al. (2006) predicted that in the *Arabidopsis* genome the transcriptional start site is usually an adenine base. This was true for most of the *Arabidopsis* and *Eucalyptus*

TSS predictions, but was not the case for the *Populus Cesa* promoters predicted. This may explain the difficulties encountered by TSSP in the prediction of poplar TSSs. Another reason for the difficulties encountered by TSSP could be the presence of multiple TSSs in a single promoter region (for review see Hughes 2006) and this may disrupt the algorithm.

A TATA-box is an AT-rich region found 25-50 bp upstream of the transcriptional start site. A TATA-box was identified in 17 of the 19 promoters and this is not surprising because it has been suggested that TATA-boxes are present in genes that require highly specific expression patterns (Molina and Grotewold 2005). The TATA-box identified in the *AtCesa* promoters showed high similarity to the model produced by Molina and Grotewold (2005). The TATA-boxes identified in the *Eucalyptus* and *Populus* promoters were more variable but this is to be expected since the model was based on *Arabidopsis* promoters and it is known that the TATA-box consensus sequences vary substantially among different species (Smale and Kadonaga 2003). In two of the promoters analysed, no AT-rich region was identified upstream of the predicted transcriptional start site (Figure 2.5, Group D and Group F). This result could indicate that the predicted TSS is incorrect or it could be that these promoters are in fact TATA-less promoters, but this will have to be verified using molecular techniques.

With the increased availability of promoter sequences, the need for fast accurate TSS predictions has increased. Although not yet accurate the current tools are a step in the right direction. Through the testing of these tools on new sequence data from diverse species, their strengths and weakness will be identified and the algorithms can be refined. The complete sequencing of other plant genomes may help in producing more

accurate plant models. In particular the *Populus* and *Eucalyptus* genomes will aid in the development of gene and promoter models specific to woody plant species. Other problems facing *in silico* identification of the TSS are that, in some cases, a gene may have a number of TSSs and each is used in a different tissue in an organism to give that gene differential expression patterns (Hughes et al. 2006). It has been shown in mammals that multiple TSSs in a gene occurs in at least 13% of all mammalian genes (Carninci et al. 2005). *In silico* prediction of TSS for genes, which have more than one TSS may lead to inaccurate results.

Although *in silico* analysis of TSSs has become increasingly more precise the most accurate methods for confirming a TSS is by molecular methods of identification. Molecular methods such as 5'RACE and primer extension have been successfully used to identify TSSs (Schelling and Jones 1995). A new technology, 5' SAGE (5'-end serial analysis of gene expression) is an adaptation of the SAGE protocol and can be used to globally identify TSSs and the frequency of individual mRNAs (Hashimoto et al. 2004). The region, which contains the TSS and core promoter elements, can also be identified by deletion studies. In this method the upstream region of a gene is systematically deleted and each section is used in an expression study in conjunction with a reporter gene. When no expression of the reporter gene is observed it indicates that the core elements required for transcription have been deleted and the DNA region just upstream of this will contain the TSS and initiation elements (Rastogi et al. 1997; Farfsing et al. 2005).

Even with the above-mentioned inaccuracies, the predicted TSS positions were comparable among the different orthologs. This was even the case for sets of orthologs where different programs predicted the TSS. In Figure 2.5 Group A, the

*Eucalyptus* TSS was predicted by TSSP to be 734 bp from the start codon while the *Populus* ortholog was predicted by NNPP to be 643 bp from the start codon (Figure 2.5). Both programs predicted a long 5'UTR, which included an intron in the 5' UTR, suggesting that these programs may be useful when used together on orthologous promoters in different plant species.

#### *CesA promoters exhibit structural conservation in 5'UTR length and structure*

When comparing the *Eucalyptus* promoter regions with those of their orthologs in *Populus* and *Arabidopsis* (Figure 2.5, Group A) it was clear that there is some structural conservation between the upstream regions of the different orthologs. This was interesting, as it mirrored the conservation of structure in the coding regions of the *CesA* genes (Samuga and Joshi 2002; Ranik and Myburg 2006). For example, within the *CesA* genes, the intron-exon patterns are conserved between species and this is the same for the introns located in the 5'UTR of *EgCesA4* promoter and its orthologs (*PtrCesA5* and *AtCesA3*). The conservation of this intron is of great interest as it has previously been shown that introns in 5'UTRs can play a role in gene regulation (Chen et al. 2002). The *EgCesA1* promoter and its *PtrCesA1* ortholog also exhibited highly conserved 5'UTR lengths only differing by a few base pairs.

It has been shown by Molina and Grotewold (2005) that the length of the 5'UTR can play a role in gene regulation. It has also been hypothesized that genes with short 5'UTRs are more highly expressed than genes with longer 5'UTRs (Rogozin et al. 2001; Hughes 2006). Accordingly, *EgCesA3* and its two orthologs (*AtCesA7* and *PtrCesA2*) had very short 5'UTRs (Figure 2.5, Group F), and *EgCesA3* (and its orthologs *AtCesA7* and *PtrCesA2*) was the most highly expressed *CesA* genes in all

three plants species (Ha et al. 2002; Samuga and Joshi 2002; Persson et al. 2005; Ranik et al. 2006; Ranik and Myburg 2006).

The isolation and detailed analysis of *CesA* promoters is important as there is very little known about the regulatory mechanisms behind cellulose production. Also, these are the first *Eucalyptus* cellulose synthase gene promoters to be reported and investigated as a group with the orthologous promoters of other woody and non-woody plants. The promoters isolated in this study together with their detailed analyses, are highly valuable as they play a specific role in cell wall formation and modulate the gene expression at very specific levels, which differs from gene to gene and in different cell types. These promoters could be used to express transgenes in a cell type-specific manner at a desired level thus giving more control over the expression of a transgene than is normally possible. The next step in this study was to use *in silico* tools to identify putative cis-regulatory elements within these sequences (see Chapter 3). The promoter regions will be tested in expression studies to ensure that they do indeed confer the expected tissue-specific expression patterns. Deletion studies will be used to identify the core functional promoter and to functionally confirm putative cis-regulatory elements.

## 2.6 Acknowledgments

The authors would like to thank J. Bradfield for her assistance with the isolation of the *EgCesA2* and *EgCesA7* promoter regions. This work was supported with funding provided by Mondi Business Paper South Africa, through the Wood and Fibre Molecular Genetics Programme, the Technology and Human Resources for Industry Programme (THRIP) and the National Research Foundation of South Africa (NRF).

## 2.7 Tables

**Table 2.1 Gene-specific primers used for genome walking of the six *CesA* promoters.** The primers were designed using the corresponding *CesA* cDNA sequence or the previous genome walking DNA sequence.

Gene	Gene-specific primer name <sup>a</sup>	Primer sequence (5'-3')
<i>EgCesA1</i>	1A - EgCesA1_GW_PR_297 (outer)	GTTGCACTCTTGACAAGCCACGAAGAC
	1B - EgCesA1_GW_PR_238 (inner)	CAGCCTCTCCGCAAGTGTGCACAG
	1C - EgCesAExon1GWUp1 (outer)	CAACCTCAATTCCTCCCACGAAATCA
	1D - EgCesAExon1GWUp2 (inner)	TTCCCTCCTTCAGCCCGAGAGAAGT
<i>EgCesA2</i>	2A - EgCesA2_KEPP_GW_97 (outer)	GCAGCATCGAAGCGCCATCATCAG
	2B - EgCesA2_KEPP_GW_40 (inner)	CAACGGCCAGGATTGAGAGGACAG
	2C - EgCesA2_GW_86 (outer)	CGAACTTCATTCTCGCCGCGGTTCT
	2D - EgCesA2_GW_30 (inner)	TTCGAAGTGGTCGTCTCGTCGTCTC
	2E - EgCesA2_GW_274 (outer)	CCATGAACACGTGCGTAATGCGTATAA
	2F - EgCesA2_GW_240 (inner)	TACAATGTCTCGCGTAGCTCGAGTATC
<i>EgCesA3</i>	3A - EgCesA3_GW_149 (outer)	TTGTGCTTGTCTGCTCATCTTCA
	3B - EgCesA3_GW_115 (inner)	CGTGCTCGAGATCATCAATGTCTT
<i>EgCesA4</i>	4A - EgCesA4_GW_171 (outer)	GATGCGCTTTCAGTCTTCCGGTT
	4B - EgCesA4_GW_118 (inner)	CACTCACACTATCATCAGCATCAG
	4C - EgCesA4_GW_224 (outer)	CTCTGGTTCTCTCGCTCGCGCTTTGT
	4D - EgCesA4_GW_183 (inner)	GGGATGTACCCACTAGCGGGCAATGTGT
	4E - EgCesA4_GW_264 (outer)	CTCCTTCCACAAGCCCAAGATCGCTCCA
	4F - EgCesA4_GW_180 (inner)	GCATAATAGGCAACGATTCTTCAGCTTAG
	4G - EgCesA4_GW_467(outer)	CGCTCGCGCTTTGTAGATGCGATGTG
	4H - EgCesA4_GW_313 (inner)	TCCGAGTGTCGGAGTCCAGCTGTAGT
<i>EgCesA5</i>	5A - EgCesA5_GW_453 (outer)	CGGACCAGCTCGTTCCTCTTGTAAGAT
	5B - EgCesA5_GW_405 (inner)	GCCTCCATCGTCGTCTCTTCTTCTCT
<i>EgCesA7</i>	7A - EgCesA7_5'UTR_Rev1 (outer)	CAGAACAGACCCGGATCTCCGCATTGCC
	7B - EgCes7_5'UTR_Rev2 (inner)	CTACGACGAGGAATGCAGCGGCCGATCT

<sup>a</sup> -The first two letters of the primer name (e.g. A1) refers to the primer codes in Figure 2. This is followed by the gene name of the sequence on which the primers were designed. GW indicates that these are genome walking primers and the number thereafter indicates the position of the primer in the sequence. The brackets at the end of the name denote whether they were the inner or outer primer of the nested primer set. KEPP indicates a conserved protein region in the CESAs around which the primers were designed.

**Table 2.2 Insert-specific primers used to close sequencing gaps in the larger genome walking products.**

<b>Gene name</b>	<b>Primer name <sup>a</sup></b>	<b>Sequence (5'-3')</b>
<i>EgCesA1</i>	EgCesA1_SW_222	GTTAACCCACCAACTACC
	EgCesA1_SW_184	GAGAGGGAGGGTAAGTTC
	EgCesA1_SW_101	TCAGCAGGTGGTTCATTAGC
<i>EgCesA2</i>	EgCesA2_SW_193	CAGAGATTTATGGCGATCC
<i>EgCesA3</i>	EgCesA3_SW_267	CAACGCACTCGCACGCACAT
	EgCesA3_SW_473	CTCGATCCGCTCAAGAGTAA
<i>EgCesA5</i>	EgCesA5_SW_218	GCGGATGTAGCTGGACTGAA

<sup>a</sup> The naming convention begins with the species followed by the gene name. SW refers to the function of the primer, which was used for sequencing of large cloned genome walking fragments, and the number indicates a unique primer number.

**Table 2.3 Forward and reverse primers used for the end-to-end amplification of the six *EgCesA* genome walking contig regions from *E. grandis* genomic DNA.**

Gene name	Primer name <sup>a</sup>	Sequence (5'-3')
<i>EgCesA1</i>	EgCesA1_39_F	CCTTGCACATCCAATTGC
<i>EgCesA1</i>	EgCesA1_2286_R	CAATCTTCTCGCGACCCAAT
<i>EgCesA2</i>	EgCesA2_Prom_F	TGGAGCATCGAGCTTCAAGG
<i>EgCesA2</i>	EgCesA2_Prom_R	GGCCACGGGCCGAGCGGGAA
<i>EgCesA3</i>	EgCesA3_MS1Prom_F	GTTCCCAACTCACTCACCTA
<i>EgCesA3</i>	EgCesA3_GW_149	TTGTGCTTGTCTGCTCATCTTCA
<i>EgCesA4</i>	EgCesA4_181_F	CGCCACAAATTGCCTCAAATG
<i>EgCesA4</i>	EgCesA4_1152_R	TCCTGGCTCGGATGCTAAGA
<i>EgCesA5</i>	EgCesA5_25_F	GCTGGTCTGCTTGACGAACT
<i>EgCesA5</i>	EgCesA5_GW_405	GCCTCCATCGTCGTCCTTCTCCTCCT
<i>EgCesA7</i>	EgCesA7_Prom_F	AAAGGAAAGACGCGACAGCCAGAA
<i>EgCesA7</i>	EgCesA7_5'UTR_R	ACCAGAACGAGAGGACCCGACTCA

<sup>a</sup> The first part of the primer name provides the gene name for which the primer is designed. The numbering indicates where in the sequence it is positioned. 5'UTR and prom indicate the region in which the primer was designed. F and R indicate the direction of the primers. The GW indicates that in some cases the original genome walking primer was used. MS1 indicates this primer was originally designed to amplify a microsatellite region.



**Table 2.4 Length of promoter regions isolated by genome walking for each *EgCesA* gene, the number of walks required and the libraries from which the fragments were amplified.**

Gene name	Promoter length <sup>a</sup>	Number of genome walks	Libraries and enzymes <sup>b</sup>
<i>EgCesA1</i>	2000 bp	2	Kit: <i>PvuII</i> and <i>EcoRV</i>
<i>EgCesA2</i>	1142 bp	3	Kit: <i>HindIII</i> , <i>DraI</i> and <i>PvuII</i>
<i>EgCesA3</i>	1312 bp	1	Kit: <i>PvuII</i>
<i>EgCesA4</i>	1537 bp	4	Siebert: <i>PvuII</i> and <i>XbaI</i> Kit: <i>PvuII</i> and <i>DraI</i>
<i>EgCesA5</i>	1363 bp	1	Kit: <i>PvuII</i>
<i>EgCesA7</i>	787 bp	1	Kit: <i>DraI</i>

<sup>a</sup> Total Length of the upstream region obtained from the genome walking of each gene

<sup>b</sup> The genome walking library the fragments were isolated from (The universal genome walker kit or the libraries produced according to the protocol of Siebert et al. (1995)). Therefore 'Kit *PvuII* and *DraI*' indicates that there were two walks, that produced fragments from the universal genome walker kit library panel and that the first fragment was isolated from the *PvuII* library and the second fragments was isolated from the *DraI* library.

**Table 2.5 Microsatellites identified in *CesA* promoter regions. The number of repeats, type of repeat and distance from the TSS are indicated.**

Gene Promoter	Repeat Unit	Number of repeats	Distance from TSS <sup>a</sup>
<i>EgCesA1</i>	CT	11	225 bp
<i>EgCesA2</i>	GCA	6	65 bp
<i>EgCesA3</i>	CT	12	1069 bp
<i>EgCesA3</i>	TCC	7	1 bp
<i>EgCesA4</i>	CT	14	636 bp
<i>PtrCesA7</i>	AT	11	15 bp
<i>AtCesA3</i>	GGT	5	90 bp

<sup>a</sup> Distance from the microsatellite to the TSS (translational start site) in base pairs.

**Table 2.6 Comparison of the transcriptional start sites (TSS) for the *Arabidopsis Cesa* promoter regions as predicted by TSSP, NNPP and that listed in TAIR.**

Gene names of <i>Arabidopsis</i> promoter regions	NNPP <sup>a</sup>	TSSP <sup>b</sup>	TAIR <sup>c</sup>
<i>AtCesA1</i>	267 bp	332 bp	277 bp
<i>AtCesA2</i>	73 bp	474 bp	173 bp
<i>AtCesA3</i>	311 bp	-	292 bp
<i>AtCesA4</i>	225 bp	228 bp	68 bp
<i>AtCesA7</i>	268 bp	-	45 bp
<i>AtCesA8</i>	194 bp	73 bp	67 bp

<sup>a</sup> TSS position predicted by NNPP (number of bp upstream of the ATG).

<sup>b</sup> TSS position predicted by TSSP (number of bp upstream of the ATG). The dashes indicate where the program was unable to predict a promoter.

<sup>c</sup> TSS position as listed on TAIR (number of bp upstream of the ATG).

**Table 2.7 Comparison of the predicted TSSs of the *Eucalyptus CesA* genes and *Populus* orthologs as predicted by TSSP and NNPP.**

Orthologous Groups <sup>a</sup>	<i>Eucalyptus</i>			<i>Populus</i>		
	Gene <sup>b</sup>	NNPP <sup>c</sup>	TSSP <sup>d</sup>	Gene <sup>b</sup>	NNPP <sup>c</sup>	TSSP <sup>d</sup>
Group D	<i>EgCesA1</i>	979 bp	210 bp	<i>PtrCesA1</i>	77 bp	221 bp
Group E	<i>EgCesA2</i>	-	154 bp	<i>PtrCesA3</i>	801 bp	800 bp
Group F	<i>EgCesA3</i>	-	45 bp	<i>PtrCesA2</i>	32 bp	-
Group A	<i>EgCesA4</i>	693 bp	734 bp	<i>PtrCesA5</i>	169 bp	-
Group B	<i>EgCesA5A</i>	270 bp	99 bp	<i>PtrCesA4</i>	180 bp	-
	<i>EgCesA5B</i>	260 bp	82 bp			
Group C	<i>EgCesA7</i>	256 bp	256 bp	<i>PtrCesA7</i>	176 bp	151 bp

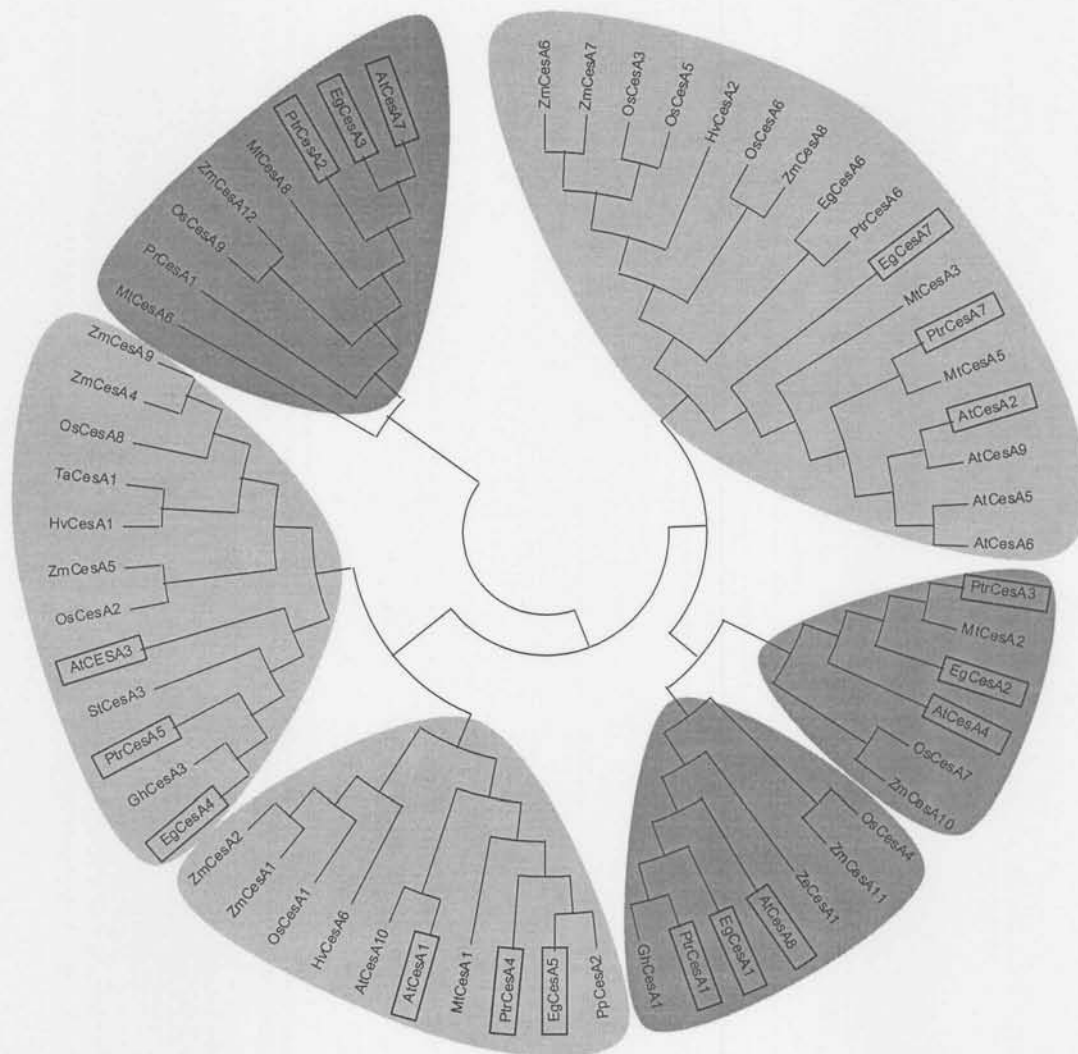
<sup>a</sup> Orthologous promoters are grouped together in groups listed A-F as in Figure 2.5.

<sup>b</sup> Names of the *CesA* genes from *Populus* and *Eucalyptus* whose promoter regions were analyzed using TSSP and NNPP.

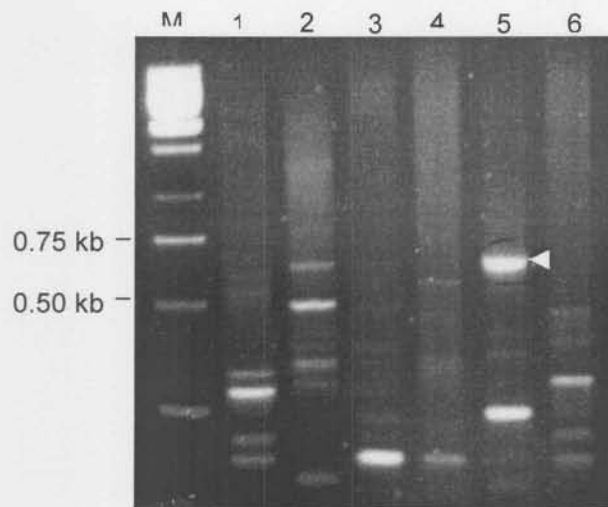
<sup>c</sup> Transcriptional start sites predicted by NNPP represented as the number of bp upstream of the start codon. The dashes indicate were the program was unable to predict a promoter.

<sup>d</sup> Transcriptional start sites predicted by TSSP, represented as the number of bp upstream of the start codon. The dashes indicate were the program was unable to predict a promoter.

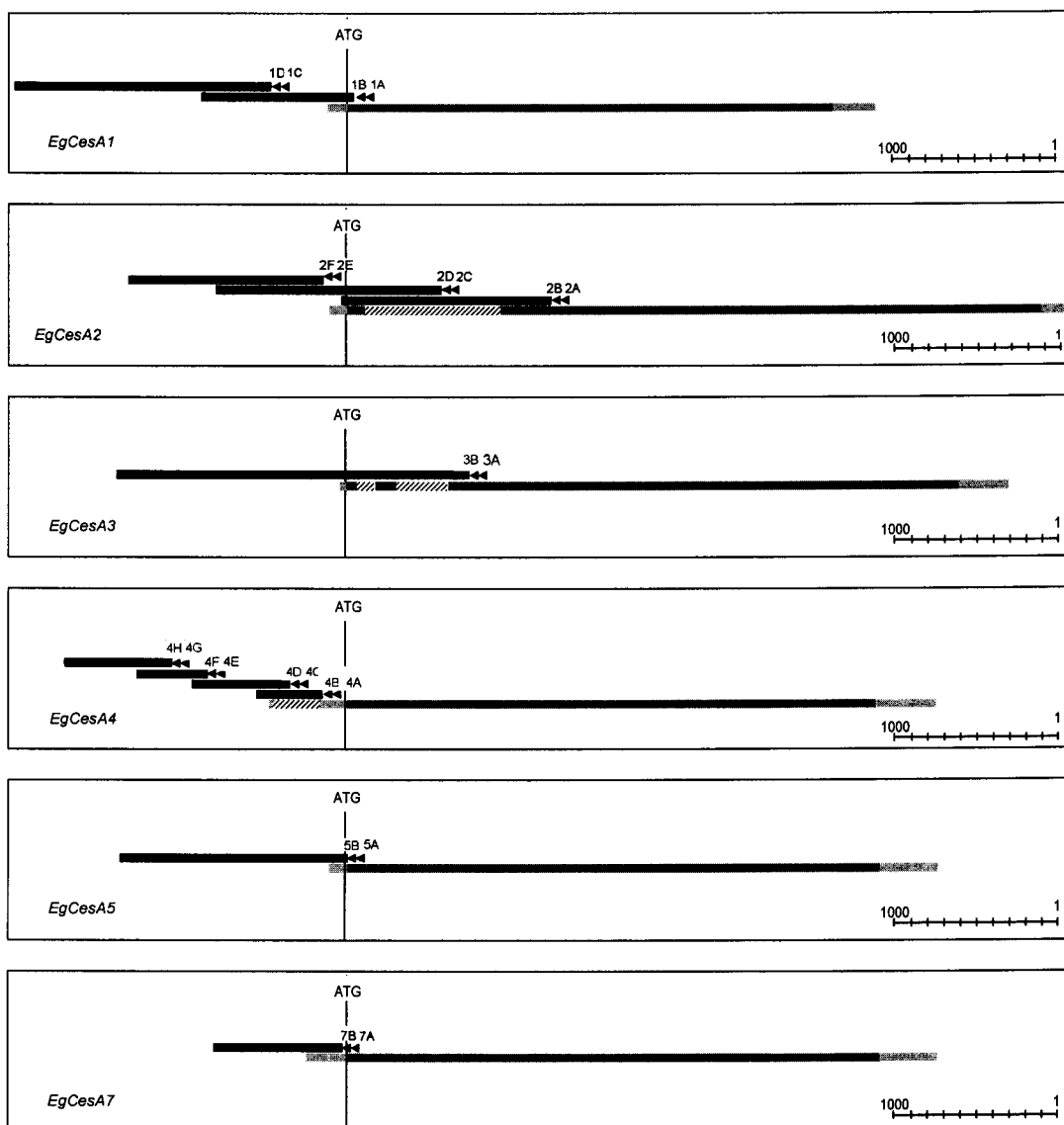
## 2.8 Figures



**Figure 2.1** Unrooted neighbour-joining tree derived from the alignment of the deduced amino acid sequences encoded by 60 full-length *CesaA* gene sequences from different plant species. 10,000 bootstrap replicates were conducted and only branches with support of 80% or greater were considered for the development of the tree. Clades containing *CesaAs* associated with primary cell wall synthesis are indicated by a light grey shading while; the dark grey shading shows those linked to secondary cell wall synthesis. Genes for which, promoters were analysed are indicated by boxes. Species names were abbreviated – At: *Arabidopsis thaliana*, Eg: *Eucalyptus grandis*, Gh: *Gossypium hirsutum*, Hv: *Hordeum vulgare*, Mt: *Medicago truncatula*, Os: *Oryza sativa*, Pr: *Pinus radiata*, Ptr: *Populus tremuloides*, St: *Solanum tuberosum*, Ta: *Triticum aestivum*, Ze: *Zinnia elegans*, Zm: *Zea mays*.

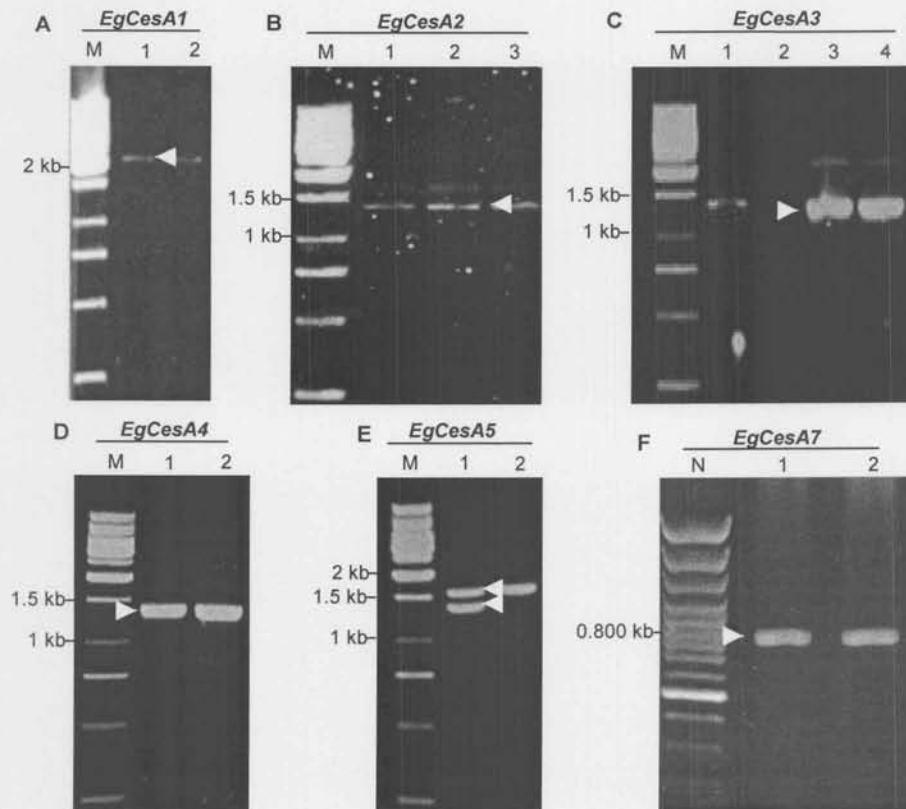


**Figure 2.2** Representative genome walking result depicting the last genome walk performed for *EgCesA4*. The first lane on the left marked M contains a 1 kb size standard. The amplified fragments were produced using genome walking primers, 4G-EgnCesA4\_GW\_467 and 4H-EgnCesA4\_GW\_313 (Table 2.1). Genome walking was performed using a panel of genome walking libraries constructed with *E. grandis* DNA. Lane 1 = *Hind*III, lane 2 = *Xba*I, lane 3 = *Sma*I, lane 4 = *EcoRV*, lane 6 = *Stu*I and Lane 5 = *Pvu*II. White arrowhead represents the fragment isolated for cloning and sequencing.



**Figure 2.3. Schematic representation of genome walking products for the six *EgCesA* promoters.** The bottom left corner of each block contains the name of the gene (*EgCesA1-5 & 7*). The vertical line indicates the position of the start codon and the bottom right of the block provides a scale in base pairs. The dark grey bars indicate the coding region of the gene of which the sequence was used to design the first set of genome walking primers. The light grey bars at each end of the coding region represent the UTRs of the gene. The black bars indicate the DNA sequence obtained from genome walking and the diagonal striped bars indicate introns that were crossed during the genome walking. The black arrows indicate the position of the primers used for the genome walking. The lettering above the arrows (e.g. 1A) indicates the name and sequence of the primers as listed in Table 1. The total distance genome walked in each case is represented in Table 2.4.





**Figure 2.4.** Agarose gel electrophoresis of the end-to-end amplification of the *EgCesA* genome walking contigs amplified from *Eucalyptus* genomic DNA. Gene name is indicated at the top of each gel image and below this indicates the lane numbers. M represents the 1 kb molecular marker (Fermentas) in each gel. Gel A (*EgCesA1*), Lane 1 and lane 2 contain a 2 kb fragment amplified from *E. grandis* genomic DNA. Gel B contains the *EgCesA2* amplification and lanes 1-3 show fragments amplified from *E. grandis*. For C (*EgCesA3*) the end-to-end amplification was performed on *E. grandis* x *E. nitens* genomic DNA (lane 1 & 2) and from *E. grandis* genomic DNA (lanes 3 & 4). In D (*EgCesA4*), lane 1 and 2 show the end-to-end amplification from *E. grandis*. In E (*EgCesA5*), lane 1 indicates the end-to-end amplification from *E. grandis*. The two bands indicate a possible difference in the two alleles. Lane 2 represents the full-length amplification of the *EgCesA5* promoter from *E. grandis* x *E. nitens* genomic DNA. F indicates the end-to end amplification of *EgCesA7* genome walking contig from *E. grandis* (Lanes 1 and 2). N indicates the 100 bp molecular marker (Fermentas) in this panel.



---

**Figure 2.5. Comparison of the predicted *Eucalyptus*, *Arabidopsis* and *Populus Cesa* upstream regions, TSS positions and core promoter elements.** Each *EgCesa* promoter is shown with its orthologous *Arabidopsis* and *Populus* promoter region in separate groups (*EgCesa4* Group A, *EgCesa5A* and *B* Group B, *EgCesa7* Group C, *EgCesa1* Group D, *EgCesa2* Group E and *EgCesa3* Group F). In each group the black bar represents the promoter sequence excluding the start codon and 5'UTR. The line labeled ATG represents the start codon. The dark grey bar indicates the 5' UTR. The transcriptional start site (TSS) is indicated by the arrow and indicates the direction of transcription. A hashed line indicates an intron in 5' UTR Group A. The 196 bp deletion in *EgCesa5B* is indicated by the black lines indicating the deletion of the light grey region and in *EgCesa5A* the light grey bar indicates the corresponding region (Group B). The column to the left of the image gives the gene name of each promoter in the set and the far left column indicates whether the genes are expressed during primary or secondary cell wall formation. The positions of the microsatellite repeats are indicated in the relevant blocks (Groups A, C, D and F).



Group A	<i>Eucalyptus grandis</i> ( <i>EgCesA4</i> )	
	<i>Populus trichocarpa</i> ( <i>PtrCesA5</i> )	
	<i>Arabidopsis thaliana</i> ( <i>AtCesA3</i> )	

Group B	<i>Eucalyptus grandis</i> ( <i>EgCesA5A</i> )	
	<i>Eucalyptus grandis</i> ( <i>EgCesA5B</i> )	
	<i>Populus trichocarpa</i> ( <i>PtrCesA4</i> )	
	<i>Arabidopsis thaliana</i> ( <i>AtCesA1</i> )	



Group C	<i>Eucalyptus grandis</i> ( <i>EgCesA7</i> )	
	<i>Populus trichocarpa</i> ( <i>PtrCesA7</i> )	
	<i>Arabidopsis thaliana</i> ( <i>AtCesA2</i> )	
Group D	<i>Eucalyptus grandis</i> ( <i>EgCesA1</i> )	
	<i>Populus trichocarpa</i> ( <i>PtrCesA1</i> )	
	<i>Arabidopsis thaliana</i> ( <i>AtCesA8</i> )	



Group E	<i>Eucalyptus grandis</i> ( <i>EgCesA2</i> )	
	<i>Populus trichocarpa</i> ( <i>PtrCesA3</i> )	
	<i>Arabidopsis thaliana</i> ( <i>AtCesA4</i> )	
Group F	<i>Eucalyptus grandis</i> ( <i>EgCesA3</i> )	
	<i>Arabidopsis thaliana</i> ( <i>AtCesA7</i> )	
	<i>Populus trichocarpa</i> ( <i>PtrCesA2</i> )	

## 2.9 References

- Alexandrov, N. N., Troukhan, M. E., Brover, V. V., Tatarinova, T., Flavell, R. B. and Feldmann, K. A. (2006) Features of *Arabidopsis* genes and genome discovered using full-length cDNAs. *Plant Mol. Biol.* 60:69-85
- Allona, I., Quinn, M., Shoop, E., Swope, K., St. Cyr, S., Carlis, J., Riedl, J., Retzel, E., Campbell, M. M., Sederoff, R. et al. (1998) Analysis of xylem formation in pine by cDNA sequencing. *PNAS* 95:9693-9698
- Andersson, A., Keskitalo, J., Sjodin, A., Bhalerao, R., Sterky, F., Wissel, K., Tandre, K., Aspeborg, H., Moyle, R., Ohmiya, Y. et al. (2004) A transcriptional timetable of autumn senescence. *Genome Biol.* 5:24
- Appenzeller, L., Doblin, M., Barreiro, R., Wang, H. Y., Niu, X. M., Kollipara, K., Carrigan, L., Tomes, D., Chapman, M. and Dhugga, K. S. (2004) Cellulose synthesis in maize: isolation and expression analysis of the cellulose synthase (*CesA*) gene family. *Cellulose* 11:287-299
- Brown, R. M. and Saxena, I. M. (2000) Cellulose biosynthesis: a model for understanding the assembly of biopolymers. *Plant Physiol. Biochem.* 38:57-67
- Burley, S. K. and Roeder, R. G. (1996) Biochemistry and structural biology of transcription factor IID (TFIID). *Annu. Rev. Biochem.* 65:769-799
- Burn, J. E., Hocart, C. H., Birch, R. J., Cork, A. C. and Williamson, R. E. (2002) Functional analysis of the cellulose synthase genes *CesA1*, *CesA2*, and *CesA3* in *Arabidopsis*. *Plant Physiol.* 129:797-807
- Burton, R. A., Shirley, N. J., King, B. J., Harvey, A. J. and Fincher, G. B. (2004) The *CesA* gene family of barley. Quantitative analysis of transcripts reveals two groups of co-expressed genes. *Plant Physiol.* 134:224-236
- Carninci, P., Kasukawa, T., Katayama, S., Gough, J., Frith, M. C., Maeda, N., Oyama, R., Ravasi, T., Lenhard, B., Wells, C. et al. (2005) The transcriptional landscape of the mammalian genome. *Science* 309:1559-1563
- Chen, W. Q., Provart, N. J., Glazebrook, J., Katagiri, F., Chang, H. S., Eulgem, T., Mauch, F., Luan, S., Zou, G. Z., Whitham, S. A., et al. (2002) Expression profile matrix of *Arabidopsis* transcription factor genes suggests their putative functions in response to environmental stresses. *Plant Cell* 14:559-574
- de Meaux, J., Goebel, U., Pop, A. and Mitchell-Olds, T. (2005) Allele-specific assay reveals functional variation in the chalcone synthase promoter of *Arabidopsis thaliana* that is compatible with neutral evolution. *Plant Cell* 17:676-690
- Demura, T., Tashiro, G., Horiguchi, G., Kishimoto, N., Kubo, M., Matsuoka, N., Minami, A., Nagata-Hiwatashi, M., Nakamura, K., Okamura, Y., et al. (2002) Visualization by

- comprehensive microarray analysis of gene expression programs during transdifferentiation of mesophyll cells into xylem cells. PNAS 99:15794-15799
- Djerbi, S., Aspeborg, H., Nilsson, P., Sundberg, B., Mellerowicz, E., Blomqvist, K. and Teeri, T. T. (2004) Identification and expression analysis of genes encoding putative cellulose synthases (*CesA*) in the hybrid aspen, *Populus tremula* (L.) x *P. tremuloides* (Michx.). Cellulose 11:301-312
- Doyle, J. J. and Doyle, J. L. (1987) A rapid DNA isolation procedure from small quantities of fresh leaf tissues. Phytochem Bull. 19:11-15
- Farfaring, J. W., Auffarth, K. and Basse, C. W. (2005) Identification of cis-active elements in *Ustilago maydis mig2* promoters conferring high-level activity during pathogenic growth in maize. Mol. Plant Microbe Interact. 18:75-87
- Fujimori, S., Washio, T., Higo, K., Ohtomo, Y., Murakami, K., Matsubara, K., Kawai, J., Carninci, P., Hayashizaki, Y., Kikuchi, S., et al. (2003) A novel feature of microsatellites in plants: a distribution gradient along the direction of transcription. Febs Letters 554:17-22
- Ha, M. A., MacKinnon, I. M., Sturcova, A., Apperley, D. C., McCann, M. C., Turner, S. R. and Jarvis, M. C. (2002) Structure of cellulose-deficient secondary cell walls from the *irx3* mutant of *Arabidopsis thaliana*. Phytochemistry 61:7-14
- Hammock, E. A. and Young, L. J. (2005) Microsatellite instability generates diversity in brain and sociobehavioral traits. Science 308:1630-1634
- Hashimoto, S., Suzuki, Y., Kasai, Y., Morohoshi, K., Yamada, T., Sese, J., Morishita, S., Sugano, S. and Matsushima, K. (2004) 5'-end SAGE for the analysis of transcriptional start sites. Nat Biotechnol 22:1146-1149
- Hertzberg, M., Aspeborg, H., Schrader, J., Andersson, A., Erlandsson, R., Blomqvist, K., Bhalerao, R., Uhlen, M., Teeri, T. T., Lundeberg, J. et al. (2001) A transcriptional roadmap to wood formation. PNAS 98:14732-14737
- Huala, E., Dickerman, A. W., Garcia-Hernandez, M., Weems, D., Reiser, L., Lafond, F., Hanley, D., Kiphart, D., Zhuang, M. Z., Huang, W., et al. (2001) The *Arabidopsis* Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. Nucl. Acids Res. 29:102-105
- Hughes, T. A. (2006) Regulation of gene expression by alternative untranslated regions. Trends Genet 22:119-122
- Jeong, Y. M., Mun, J. H., Lee, I., Woo, J. C., Hong, C. B. and Kim, S. G. (2006) Distinct roles of the first introns on the expression of *Arabidopsis* profilin gene family members. Plant Physiol. 140:196-209
- Joshi, C. P., Bhandari, S., Ranjan, P., Kalluri, U. C., Liang, X., Fujino, T. and Samuga, A. (2004) Genomics of cellulose biosynthesis in poplars. New Phytol. 164:53-61

- Kainz, M. and Roberts, J. (1992) Structure of transcription elongation complexes in vivo. *Science* 255:838-841
- Kerstens, S. and Verbelen, J. P. (2003) Cellulose orientation at the surface of the *Arabidopsis* seedling. Implications for the biomechanics in plant development. *J. Struct. Biol.* 144:262-270
- Kutach, A. K. and Kadonaga, J. T. (2000) The downstream promoter element (DPE) appears to be as widely used as the TATA-box in *Drosophila* core promoters. *Mol. Cell. Biol.* 20:4754-4764
- Lo, K. and Smale, S. T. (1996) Generality of a functional initiator consensus sequence. *Gene* 182:13-22
- Loke, J. C., Stahlberg, E. A., Strenski, D. G., Haas, B. J., Wood, P. C. and Li, Q. Q. (2005) Compilation of mRNA polyadenylation signals in *Arabidopsis* revealed a new signal element and potential secondary structures. *Plant Physiol.* 138:1457-1468
- Lourenço, R. (2004) Genomic structure of three mega basepairs of genomic shotgun DNA of *Eucalyptus grandis*: nucleotide content, repetitive sequences and genes, Univ of Campinas: pp117
- Martin, P., Makepeace, K., Hill, S. A., Hood, D. W. and Moxon, E. R. (2005) Microsatellite instability regulates transcription factor binding and gene expression. *PNAS* 102:3800-3804
- Milioni, D., Sado, P. E., Stacey, N. J., Domingo, C., Roberts, K. and McCann, M. C. (2001) Differential expression of cell-wall-related genes during the formation of tracheary elements in the *Zinnia* mesophyll cell system. *Plant Mol. Bio.* 47:221-238
- Mingam, A., Toffano-Nioche, C., Brunaud, V., Boudet, N., Kreis, M. and Lecharny, A. (2004) DEAD-box RNA helicases in *Arabidopsis thaliana*: establishing a link between quantitative expression, gene structure and evolution of a family of genes. *Plant Biotechnol. J.* 2:401-415
- Mohanty, B., Krishnan, S. P. T., Swarup, S. and Bajic, V. B. (2005) Detection and preliminary analysis of motifs in promoters of anaerobically induced genes of different plant species. *Ann. Botany* 96:669-681
- Molina, C. and Grotewold, E. (2005) Genome wide analysis of *Arabidopsis* core promoters. *BMC Genomics* 6:25-36
- Morgante, M. (2006) Plant genome organisation and diversity: the year of the junk! *Curr. Opin. Biotechnol.* 17:168-73
- Morgante, M., Hanafey, M. and Powell, W. (2002) Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. *Nat. Genet.* 30:194-200
- Nobles, D. R., Romanovicz, D. K. and Brown, R. M., Jr. (2001) Cellulose in cyanobacteria. Origin of vascular plant cellulose synthase? *Plant Physiol.* 127:529-542

- Persson, S., Wei, H. R., Milne, J., Page, G. P. and Somerville, C. R. (2005) Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets. *PNAS* 102:8633-8638
- Prassinou, C., Ko, J. H. and Han, K. H. (2005) Transcriptome profiling of vertical stem segments provides insights into the genetic regulation of secondary growth in hybrid aspen trees. *Plant Cell Physiol.* 46:1213-1225
- Ranik, M., Creux, N. M. and Myburg, A. A. (2006) Within-tree transcriptome profiling in wood-forming tissues of a fast-growing *Eucalyptus* tree. *Tree Physiol.* 26:365-375
- Ranik, M. and Myburg, A. A. (2006) Six new cellulose synthase genes from *Eucalyptus* are associated with primary and secondary cell wall biosynthesis. *Tree Physiol.* 26:545-556
- Rastogi, R., Bate, N. J., Sivasankar, S. and Rothstein, S. J. (1997) Footprinting of the spinach nitrite reductase gene promoter reveals the preservation of nitrate regulatory elements between fungi and higher plants. *Plant Mol. Biol.* 34:465-476
- Reese, M. G. (2001) Application of a time-delay neural network to promoter annotation in the *Drosophila melanogaster* genome. *Comput. Chem.* 26:51-56
- Reese, M. G., Hartzell, G., Harris, N. L., Ohler, U., Abril, J. F. and Lewis, S. E. (2000) Genome annotation assessment in *Drosophila melanogaster*. *Genome Res.* 10:483-501
- Richmond, T. A. and Somerville, C. R. (2000) The cellulose synthase superfamily. *Plant Physiol.* 124:495-498
- Roberts, K. and McCann, M. C. (2000) Xylogenesis: the birth of a corpse. *Curr. Opin. Plant Biol.* 3:517-522
- Rogozin, I. B., Kochetov, A. V., Kondrashov, F. A., Koonin, E. V. and Milanesi, L. (2001) Presence of ATG triplets in 5' untranslated regions of eukaryotic cDNAs correlates with a 'weak' context of the start codon. *Bioinformatics* 17:890-900
- Rombauts, S., Florquin, K., Lescot, M., Marchal, K., Rouze, P. and Van de Peer, Y. (2003) Computational approaches to identify promoters and cis-regulatory elements in plant genomes. *Plant Physiol.* 132:1162-1176
- Samuga, A. and Joshi, C. P. (2002) A new cellulose synthase gene (*PtrCesA2*) from aspen xylem is orthologous to *Arabidopsis AtCesA7 (irx3)* gene associated with secondary cell wall synthesis. *Gene* 296:37-44
- Sawant, S. V., Kiran, K., Mehrotra, R., Chaturvedi, C. P., Ansari, S. A., Singh, P., Lodhi, N. and Tuli, R. (2005) A variety of synergistic and antagonistic interactions mediated by cis-acting DNA motifs regulate gene expression in plant cells and modulate stability of the transcription complex formed on a basal promoter. *J. Exp. Bot.* 56:2345-2353
- Saxena, I. M. and Brown, R. M. (2005) Cellulose biosynthesis: Current views and evolving concepts." *Ann. Botany* 96:9-21



- Schelling, D. and Jones, G. (1995) Functional identification of the transcription start site and the core promoter of the juvenile hormone esterase gene in *Trichoplusia*. *Biochem. Biophys. Res. Commun.* 214:286-294
- Shahmuradov, I. A., Gammerman, A. J., Hancock, J. M., Bramley, P. M. and Solovyev, V. V. (2003) PlantProm: a database of plant promoter sequences. *Nucl. Acids Res.* 31:114-117
- Shahmuradov, I. A., Solovyev, V. V. and Gammerman, A. J. (2005) Plant promoter prediction with confidence estimation. *Nucl. Acids Res.* 33:1069-1076
- Siebert, P. D., Chenchik, A., Kellogg, D. E., Lukyanov, K. A. and Lukyanov, S. A. (1995) An improved PCR method for walking in uncloned genomic DNA. *Nucl. Acids Res.* 23:1087-1088
- Smale, S. T. (1997) Transcription initiation from TATA-less promoters within eukaryotic protein-coding genes. *Biochim Biophys Acta* 1351:73-88
- Smale, S. T. and Kadonaga, J. T. (2003) The RNA Polymerase II core promoter. *Annu. Rev. Biochem.* 72:449-479
- Sterky, F., Regan, S., Karlsson, J., Hertzberg, M., Rohde, A., Holmberg, A., Amini, B., Bhalerao, R., Larsson, M., Villarroel, R. et al. (1998) Gene discovery in the wood-forming tissues of poplar: Analysis of 5,692 expressed sequence tags. *PNAS* 95:13330-13335
- Svejstrup, J. Q. (2004) The RNA polymerase II transcription cycle: cycling through chromatin. *Biochim Biophys Acta* 1677:64-73
- Tanaka, K., Murata, K., Yamazaki, M., Onosato, K., Miyao, A. and Hirochika, H. (2003) Three distinct rice cellulose synthase catalytic subunit genes required for cellulose synthesis in the secondary wall. *Plant Physiol.* 133:73-83
- Taylor, N. G., Howells, R. M., Huttly, A. K., Vickers, K. and Turner, S. R. (2003) Interactions among three distinct CESA proteins essential for cellulose synthesis. *PNAS* 100:1450-1455
- Taylor, N. G., Laurie, S. and Turner, S. R. (2000) Multiple cellulose synthase catalytic subunits are required for cellulose synthesis in *Arabidopsis*. *Plant Cell* 12:2529-2539
- Tompa, M., Li, N., Bailey, T. L., Church, G. M., De Moor, B., Eskin, E., Favorov, A. V., Frith, M. C., Fu, Y. T., Kent, W. J., et al. (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotechnol.* 23:137-144.
- Turner, S. R. and Somerville, C. R. (1997) Collapsed xylem phenotype of *Arabidopsis* identifies mutants deficient in cellulose deposition in the secondary cell wall. *Plant Cell* 9:689-701
- Tuskan, G. A., DiFazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A., et al. (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science* 313:1596-1604

Yang, J. M., Park, S., Kamdem, D. P., Keathley, D. E., Retzel, E., Paule, C., Kapur, V. and Han, K. H. (2003) Novel gene expression profiles define the metabolic and physiological processes characteristic of wood and its extractive formation in a hardwood tree species, *Robinia pseudoacacia*. *Plant Mol. Biol.* 52:935-956

## Chapter 3

# ***In silico* analysis of Cis-acting elements in the cellulose synthase promoters of *Eucalyptus*, *Populus* and *Arabidopsis***

<sup>1</sup>N.M. Creux, <sup>2</sup>D.K. Berger and <sup>1</sup>A.A. Myburg

<sup>1</sup>*Department of Genetics, Faculty of Natural and Agricultural Sciences, Forestry Agricultural  
Biotechnology Institute (FABI), University of Pretoria, Pretoria, South Africa*

<sup>2</sup>*Department of Botany, Faculty of Natural and Agricultural Sciences, Forestry and  
Agricultural Biotechnology Institute (FABI), University of Pretoria, Pretoria, South Africa*

This chapter has been prepared in the format of a manuscript for a research journal (e.g. Plant Molecular Biology). I identified the software programs to be used for the sequence analysis and performed all of the *in silico* analyses, interpreted the data and prepared the manuscript. Prof. A.A. Myburg and Prof. D.K. Berger provided valuable advice, direction and supervision in the planning and implementation of the project. They also provided important direction and critical revision of the manuscript. All other technical assistance is listed in the acknowledgements.

### 3.1 Abstract

Cellulose is one of the most abundant biopolymers on earth and is produced by a multi-subunit complex of membrane bound proteins known as cellulose synthases (CESAs). *CesA* genes have been isolated from several different plant species including *Arabidopsis*, *Populus* and *Eucalyptus*. Expression analysis of the *CesA* genes in these plants has revealed the presence of two distinct groups of *CesA* genes. A set of three *CesA* genes are co-expressed in tissues undergoing secondary cell wall formation and a different set of *CesA* genes are co-expressed during primary cell wall formation. Although their expression patterns have been extensively investigated, little is known about the regulatory mechanisms that govern their unique expression patterns. Six *Eucalyptus CesA* gene promoters were isolated by genome walking (Chapter 2) and used in a comparative *in silico* analysis with the orthologous promoter regions from *Arabidopsis* and *Populus*. This is the first study in which the promoters of the *CesA* gene family are characterized in *Arabidopsis*, *Populus* and *Eucalyptus*. Three software packages (Weeder, POCO and MotifSampler) were used to analyse the promoter regions and identify over represented motif sequences. A number of key stem-specific and xylem-specific motifs such as the AC-motif and G-box motif were identified as well as a number of novel motifs. This Chapter gives a detailed list of the possible motifs involved in the transcriptional regulation of the cellulose synthase genes. Although all of the predicted motifs identified here will have to be functionally tested, the results of this study provide a good map for directed deletion studies and functional testing of the *CesA* promoters.

### 3.2 Introduction

Cellulose is one of the main components of the cell wall and much is now understood about its biosynthesis and deposition. Cellulose is synthesized and deposited in the cell wall by a complex of membrane bound enzymes. A number of different *CesA* genes produce each of the proteins, which make up the cellulose synthase complex. Interestingly these genes have variable expression patterns where some genes are associated with primary cell wall formation and a different set of *CesA* genes are associated with secondary cell wall formation (Delmer et al. 2000; Burn et al. 2002). The actual regulatory mechanisms behind their unique expression patterns have not been well characterized and there is no information available about the promoter sequences and cis-regulatory elements involved in this process.

Cellulose is a key component of all plant cell walls and the cellulose fibers produced by the plants are used in a number of different products including paper. Cellulose is deposited in the plant cell wall to strengthen the cell wall. In the stem cellulose is deposited in a two-fold process where the less structured cellulose is deposited in the primary cell wall to allow for growth. And then more crystalline and ridged cellulose is deposited in the secondary cell wall to provide strength to the cells (Emons and Mulder 2000). A large complex of membrane bound enzymes synthesizes and deposits the cellulose in the cell walls. This complex is comprised of six catalytic subunits arranged in a rosette structure in the cell membrane. Each catalytic subunit consists of six cellulose synthase (CESA) proteins also in a rosette configuration (Brown and Saxena 2000). These proteins are encoded for by a number of *CesA* genes. The *CesA* gene family is well conserved in higher plant species with at least seven different genes and several paralogs reported by a number of studies in different plant species (Burn et al. 2002; Joshi 2003; Burton et al. 2004; Ranik and Myburg

2006). Ranik and Myburg (2006) recently isolated six cellulose synthase genes from *Eucalyptus* and these sequences were used to isolate the promoter regions by genome walking (Chapter 2).

Expression studies of the *CesA* genes revealed that three *CesA* genes (*AtCesA8*, *AtCesA4* and *AtCesA7*) were expressed during secondary cell wall formation and another set of genes were associated with primary cell wall formation (Turner and Hall 2000; Burn et al. 2002; Hamann et al. 2004). In further studies it was found that this expression pattern was maintained in a number of different species including *Populus* (Joshi et al. 2004), *Eucalyptus* (Ranik and Myburg 2006) and barley (Burton et al. 2004). Ranik and Myburg (2006) performed expression studies on the *CesA* genes isolated from *Eucalyptus* and found that apart from the different expression patterns among the different tissues it appeared that the different *CesA* genes were expressed at different levels within one tissue.

Although the expression patterns of the *CesA* genes have been studied in some detail there is little or no information available on the regulatory mechanisms that determine these expression patterns. Persson et al. (2005) performed a genome wide study to identify all the genes in *Arabidopsis* that were co-expressed with the *CesA* genes, associated with primary cell wall formation and secondary cell wall formation. The up-stream promoter sequence of these genes are available on TAIR but have not yet been investigated. It is expected that the promoters of co-expressed genes will have a number of cis-regulatory motifs in common. Thus it is expected that the promoters of the two sets of genes identified by Persson et al. (2005) will have some motifs in common but between the two sets there could be a number of different cis-regulatory motifs specific to only one set.

Lignin biosynthesis is also an important process in xylogenesis and a number of studies have shown that many of the genes involved in lignin biosynthesis are co-expressed with the secondary cell wall associated *CesA* genes (Hertzberg et al. 2001; Demura et al. 2002; Ranik et al. 2006) and thus may share a number of important cis-regulatory elements. The promoters of some key lignin genes have been isolated and cis-regulatory elements involved in xylem-specific expression have been identified (Lacombe et al. 2000; Andersson et al. 2004). Some of the key cis-regulatory elements involved in xylem-specific expression of the lignin genes are the AC rich elements (AC I, ACII and ACIII) and the G-box. Hatton et al. (1995) found that the G-Box is bound by a b-ZIP protein and this in conjunction with an AC element was involved in the light responsiveness of the genes. This study also showed that ACI and ACII bound MYB (MYeloBlastosis virus) proteins and regulated xylem-specific expression. In *Eucalyptus* the *CAD* (Cinnomyl Alcohol Dehydrogenase) gene promoter contained an AC-element that binds a similar MYB transcription factor (Lacombe et al. 2000; Goicoechea et al. 2005). These previously identified cis-regulatory elements may also play a role in the regulation of the *CesA* genes that are co-expressed with these lignin genes (For a complete review of the transcription factors involved in wood formation see Chapter 1).

The sequencing of a number of plant genomes has created a need for faster methods of cis-regulatory element identification. The newly available promoter sequences and large quantity of expression data from methods such as microarray and cDNA-AFLP (Kuhn 2001) have lead to the development of *in silico* methods for cis-regulatory motif prediction. Tompa et al. (2005) reviewed a number of the software packages available at present and show that if two software packages, which are based on different search methods, were used the accuracy of motif prediction was greatly

increased. MotifSampler (Thijs et al. 2001; Tompa et al. 2005) and Weeder used in conjunction performed better than any of the other software programs tested. This study also showed that of the software packages tested, Weeder (Tompa et al. 2005) was the most accurate. Weeder uses a pattern-based search to identify over-represented motifs in a set of co-expressed gene promoters, while MotifSampler uses a traditional alignment driven search, which is the basis of a number of prediction tools. POCO (promoters of co-expressed genes) was not tested in the Tompa review (2005) but is a novel approach to motif identification as it compares two sets of promoters of oppositely expressed genes (Kankainen and Holm 2005). POCO then searches for motifs that are present in one set and absent in the other set, as well as motifs over-represented in both sets, relative to the background. The background used in POCO is a collection of the upstream regions of nearly all of the genes in the selected genome thus finding motifs that are over represented in a subset of promoters when compared to all the promoters.

The high-through put methods for motif identification, and large quantity of promoter sequence available for functional testing has lead to an increase in the number of motifs being identified and to the development of a number of cis-regulatory element databases. For example, TRANSFAC is a database that houses all eukaryotic cis-regulatory elements and the transcription factors that bind to them (Wingender 2000). Two plant-specific databases, PLACE and PlantCARE, list all the plant cis-regulatory elements that have been identified (Higo et al. 1999; Lescot et al. 2002). These databases offer a central resource where one can search for motifs that have been identified either by functional testing or *in silico* methods and also offer a number of tools for searching promoters for known promoter regions.



The differential expression patterns of the *CesA* genes and the isolation of a number of *CesA* promoters from different plant species have lead to a number of interesting questions. The first is weather there are any cis-regulatory elements shared among the *CesA* genes and their co-expressed genes. Are these motifs conserved among different plant species? And what regulates the *CesA* gene expression patterns, i.e. are there different cis-regulatory elements involved in the expression of the primary and secondary associated *CesA* genes? Here we present a preliminary study of the possible motifs involved in the unique expression patterns of the *CesA* genes. This is the first comparative *in silico* analysis of the *CesA* promoters isolated from *Arabidopsis*, *Populus* and *Eucalyptus* to identify possible motifs conserved among the different species that may play a role in the regulation of these genes. This study lists a number of motifs that may be involved in the differential expression of the *CesA* genes, as well as motifs present in the promoters of the co-expressed *Arabidopsis* genes identified by Persson et al. (2005).

### **3.3 Materials and Methods**

#### *Promoter sequences*

The DNA sequences of the cellulose synthase promoters were obtained from *Eucalyptus*, *Arabidopsis* and *Populus*. The *Eucalyptus* promoter sequences were isolated by genome walking as described in Chapter 2. The *Arabidopsis CesA* promoter sequences were downloaded from TAIR ([www.arabidopsis.org](http://www.arabidopsis.org)) using their locus identifiers (Table 3.1 and 3.2). The *Arabidopsis* promoter sequences of 34 *Arabidopsis* genes (Table 3.2) that were highly co-expressed with the primary and secondary *AtCesA* genes were also downloaded from TAIR (Persson et al. 2005). The *Populus* promoter sequences were obtained as described in Chapter 2.

For *in silico* analysis of promoter sequences, exactly 1kb of DNA sequence upstream of the predicted transcriptional start site was downloaded from TAIR (*Arabidopsis*). The *Arabidopsis* promoter regions are provided on TAIR from the transcriptional start site but most of these transcriptional start sites have not been confirmed, but are predicted by software. The *Populus* and *Eucalyptus* promoter regions were therefore also formatted to be 1 kb in length upstream from the predicted transcriptional start site (Chapter 2).

#### *Cellulose synthase promoter data sets*

Two cellulose synthase promoter data sets were compiled from the cellulose synthase promoters of *Eucalyptus*, *Arabidopsis* and *Populus*. The *CesA* promoters were separated into two groups according to the expression patterns of the genes. The first group contained the promoters of the *CesA* genes predicted to be expressed during primary cell wall formation (Table 3.1) and is referred to as **CesA set 1** in the text. The second group of *CesA* promoters were obtained from the *CesA* genes predicted to be expressed during secondary cell wall formation (Table 3.1) and is referred to as **CesA set 2**.

#### *CesA co-expressed gene promoter data sets*

Persson et al. (2005) listed two sets of *Arabidopsis* genes that were co-expressed with the *Arabidopsis* cellulose synthase genes. The first set of genes listed in the study, were all highly co-expressed with the primary cell wall associated *AtCesA* genes. The second set of genes, were highly co-expressed with the secondary cell wall associated *AtCesA* genes. Two data sets were compiled using the promoter regions of a number of the genes listed by Persson et al. (2005). The first data set referred to in this text as **Co-expressed set 1** contained the promoters of 17 genes found to be co-expressed

with the primary cell wall associated *AtCesA* genes and the second data set referred to in this text as **Co-expressed set 2** was compiled using the promoters of 17 genes that were co-expressed with the secondary cell wall associated *AtCesA* genes (Persson et al. 2005). **Co-expressed set 1** and **Co-expressed set 2** also contained the promoters of the respective *AtCesA* genes, thus each data set contained a total of 20 promoters.

### *Motif analysis*

In order to identify cis-regulatory elements that may play a role in the tissue-specific expression of the cellulose synthase genes, accurate *in silico* tools had to be identified from the large number of tools that are available reviewed in Chapter 1 (Table 1.4 and 1.5). Three software tools were selected for the identification of cis-regulatory elements in the *CesA* promoters. MotifSampler (Thijs et al. 2001; Thijs et al. 2002) and Weeder (Pavesi et al. 2004) were selected based on a review by Tompa et al. (2005) who showed that MotifSampler and Weeder produced more accurate results when used in conjunction. The third software program was selected because of its ability to analyze clusters of promoters from clusters of genes with opposite expression patterns, which was how the data sets in this study were structured. The genes of the promoters in *CesA* set 1 are oppositely expressed to the genes of the promoters in *CesA* set 2. This type of analysis may identify motifs that are over-represented in one data set compared to the background but under-represented in the other data set compared to the background. Another reason for selecting these programs was that they all offered an *Arabidopsis* background for comparison and this would be more accurate when analysing plant promoters than a mammalian background, which many software packages use. The background models used were composed of the promoters from the *Arabidopsis* genome. In all of the analyses performed both six and eight length motifs were searched for because most

transcription factor binding sites are between four to eight base pairs in length. Because the algorithms have built in an allowance for mutated bases when searching for eight length motifs the shorter motifs are also identified. This was true here, where often the six length motifs were contained in the eight length motifs (Appendix 2.8).

MotifSampler (<http://homes.esat.kuleuven.be/~thijs/Work/MotifSampler.html>) is based on the Gibbs sampling method and allows for the identification of statistically over-represented motifs in a set of un-aligned sequences. The algorithm determines at what positions and in which sequences a statistically over-represented motif is present (Thijs et al. 2002). This algorithm also incorporates the use of a higher order background model, which increases its prediction accuracy (Thijs et al. 2001). MotifSampler has a number of user specified parameters, which gives the user more control over the analysis. For this study most of the default settings were used except that length 6 motifs were also searched for in both strands instead of only length 8 motifs. The other setting altered for this analysis was the number of motifs obtained per run. The default is set at six, but the software recommends searching for fewer motifs per search and performing more searches and so this was set at two and three motifs per search and then ten searches were performed for each setting.

Weeder Web (<http://159.149.109.16:8080/weederWeb/>) is the web interface for Weeder, an algorithm that discovers conserved motifs in a set of related regulatory DNA sequences. The algorithm incorporates a significance statistic that is specific for transcription factor binding sites and this measure is used to rank the results in order to identify the best motifs (Pavesi et al. 2001). This measure of significance accounts for the number of sequences the motif appears in, how conserved the motif is and the overall number of occurrences in the input set. Then the ten most significant motifs

for each run are reported to the user (Pavesi et al. 2004). The Weeder algorithm was used in this study with the default settings maintained for most of the parameters. Both strands were searched using the thorough scan option. All of the promoter sets were analysed and the 10 highest scoring motifs per set were presented in the output. These motifs and the reverse complement of the motifs were compared to the motifs identified by MotifSampler in order to identify motifs predicted by both software packages.

POCO (<http://ekhidna.biocenter.helsinki.fi/poco>) was the third software included in the analyses (Kankainen and Holm 2005). The analysis was performed predominately using the default settings. The length of the promoter was automatically set by the software to the length of the shortest promoter in the data sets (1 kb in this study). Both six and eight nucleotide motifs were searched for and the full *Arabidopsis* background provided by POCO was used. The first search was for motifs in the cellulose synthase data sets where CesaA set 1 was entered as the first cluster and the CesaA set 2 was entered as the second cluster. The second search was similar, but in this case the co-expressed data sets were used for the first and second cluster. The search was performed three times for each data set and the motifs not identified in all three sets of results were discarded. The output generated approximately 20 motifs per data set and these were compared to the motifs identified by MotifSampler and Weeder. Motifs that were not predicted by at least two of the three programs and motifs with a P-value (calculated by POCO) greater than 0.05, were discarded.

#### *Motif annotation*

A number of the motifs identified in the study had the same or very similar sequences and were likely the same motif. These motifs were grouped together and a consensus

sequence for each group was generated. A motif was often predicted multiple times by the software packages and these were grouped together, based on sequence similarity, and a consensus motif sequence was generated. The consensus was constructed by observing each position in the different sequences and if the majority of sequences contained a particular base, that base was used in the consensus. In a case where there are only two sequences and they did not agree an ambiguous base is included in the consensus. Some motif sequences were not exactly the same but overlapped with other sequences or, in the case of eight bp, contained other motifs (six bp in length) and these were also included in the motif consensus sequence.

The consensus motifs were used in homology searches of the PLACE database (<http://www.dna.affrc.go.jp/htdocs/PLACE/signalup.html>) for similar known plant cis-regulatory elements (Higo et al. 1999). Results from the PLACE database are listed a measure of the similarity (E-value), which is not provided on PlantCARE. But plantCARE may contain motifs not present in the PLACE database and therefore the motifs that showed very weak, or no similarity to motifs in the PLACE database were also used in homology searches of the PlantCARE database (Lescot et al. 2002). The motifs identified in the four different datasets were compared to identify motifs that are present in two or more of the datasets as this could indicate similar regulatory mechanisms. Two other *Eucalyptus* promoters (*EgCesA2* and 7) were isolated later in the study were not included in the initial datasets. These were later scanned for motifs similar to these identified during the initial data analysis using Vector NTI (Invitrogen) and promoter scan programs on the PLACE and plantCARE websites (Figure 3.1 provides a flow diagram of the method discussed above).

### 3.4 Results

#### *Cis-regulatory motif identification*

The four datasets (Table 3.1 and 3.2) were used as input data for the three different motif identification software packages used in this study (MotifSampler, Weeder and POCO). MotifSampler generated a total of 400 motifs, which was comprised of 100 motifs for each data set (results not shown). Weeder reported a total of 160 motif sequences of which 40 motif sequences were predicted for each dataset with a high significance (results not shown). POCO reported 20 motifs, which were over represented in one dataset when compared to the other datasets. A total of 100 motif sequences were reported by POCO for the CesA datasets, which comprised of 40 sequences predicted to be over-represented in CesA set 1 only, 40 sequences were over represented in CesA set 2 only and the top 20 sequences over-represented in both sets when compared to the background. This was the same for the co-expressed data sets with the top motifs in each category being reported (results not shown). The total number of sequences predicted to be over-represented by the software programs in the four different datasets were 760 motif sequences, which were further sorted and compared to one another.

The MotifSampler results were analyzed and all motifs that were repeated or had similar sequences (sequences that overlapped, were contained in or had similar sequences when ambiguous bases were taken into account) were grouped together for each dataset (results not shown). The POCO motif search was performed three times and the results were compared. Any motif that was not repeated in all three sets of results or was not of length six or eight was removed. The motifs were then compared to the 400 motifs identified by MotifSampler in order to identify motifs that were predicted by both software packages. The 160 motifs predicted by Weeder were also

compared to the MotifSampler and POCO results in order to identify motifs predicted by two or more software programs. Any of the motifs that were not identified by two or more of the software programs were excluded from further analysis. Of the motifs identified by two or more software programs, 65 motifs were identified by MotifSampler and POCO, 56 were identified by MotifSampler and Weeder and 14 were predicted by all three programs, thus a total of 135 motifs were used for further analysis (results not shown). Of the 135 motifs, any that had a P-value predicted by POCO to be more than 0.05 were excluded from the analysis. This resulted in a final set of 81 motifs identified by at least two of the three softwares that had a P-value of less than 0.05 (there is only a 5% chance of the motif being over represented by chance). The number of motifs identified for each dataset was as follows: 24 of the 81 motifs were over-represented in CesA set 1 (Table 3.3), 12 were over-represented in CesA set 2 (Table 3.4), 22 were identified in the Co-expressed set 1 (Table 3.5) and 23 were identified in Co-expressed set 2 (Table 3.6).

#### *Motif annotation*

During the motif analysis approximately 20 consensus sequences representing the motifs identified to have a strong significance were produced for all datasets except for CesA set 2 from which only 12 consensus motifs were identified (Table 3.7, 3.8, 3.9 and 3.10). All of these consensus motifs were used in similarity searches of the PLACE motif database. Some of the motifs had sequence similarity to transcription factor binding sites in the database. A total of 38 motifs in the four different datasets had E-values of less than 1.0 and were likely putative identities for the motifs. The motifs with little (E-value greater than or equal to 1.0) or no similarity to motifs in the PLACE database were used in similarity searches in the PlantCARE database and were also found to have little or no similarity to motifs on this database (results not



shown). These motifs may prove to be of interest as they may be novel motifs specific to primary or secondary cell wall formation.

Some of the consensus motifs identified in Cesa set 1 had similarity to a number of interesting motifs in the PLACE database (Table 3.7). Motifs CEP2, CSP2, CP2, CP5, CP7, CP12 and CP13 (Appendix 2.4, 2.7, 2.9, 2.12, 2.14, 2.19, 2.20) showed similarity to hormonal response elements known to play a role in plant development such as abscisic acid, auxin, ethylene and gibberellic acid. Two well-known stem elements were also identified during the similarity search of motifs CP3 (Hatton et al. 1995) and CP4 (Keller and Baumgartner 1991). Motif CP10 also showed low similarity to a core MYB binding site (Luscher and Eisenman 1990a; Luscher and Eisenman 1990b).

Five of the motifs identified in Co-Expressed set 1 (Table 3.9) showed similarity to a number of key elements identified in Cesa set 1. The elements EP3 and CEP2 (Tables 3.10 and 3.8) also showed similarity to the hormone response elements for gibberellic acid and abscisic acid. This dataset also identified a number of elements involved in general gene regulation such as a general activation element (EP2), an initiator element (EP12) and an element for the modulation of gene activity (EP8). Another important element (EP4), which was identified by the search, is a sugar responsive element to which the WRKY family of transcription factors bind (Sun et al. 2003). Motifs CEP1, CEP2 and CEP3 (Appendix 2.3 -2.5) were identified in both Cesa set 1 and Co-expressed set 1 and showed similarity a negative response element (Ngai et al. 1997), an abscisic acid response element (Kao et al. 1996) and a general activation element (Benfey et al. 1989; Benfey et al. 1990) respectively.

Fewer motifs were identified for *CesA* set 2 than any of the other data sets (Table 3.8). In this set there were fewer hormonal response elements identified than in the *CesA* set 1 and Co-Expressed set 1. Motif CSP2 (Appendix 2.7) showed similarity to the only hormone response element identified in this set, an ethylene response element (Solano et al. 1998). Motif CS1 (Appendix 2.25) highly similar to an element found in a virus that drives phloem specific expression (Yin et al. 1997), but this element is also highly similar to the AC-elements found in lignin gene promoters (Hatton et al. 1995). A number of photo-regulation elements were also identified during the search (motifs CESP1 and CS2), and one element (CS3) was found to play a role in regulation of circadian genes. Motifs CSP1 (Appendix 2.6) and CSP2 were identified in both the *CesA* set 1 and 2 but not in the Co-expressed data sets.

Co-Expressed set 2 (Table 3.10) also had fewer hormone response elements with only motifs ES9 and ES16 showing similarity to auxin response elements. Motif ES1 was interesting as it had similarity to an element identified in the promoters of genes involved in sclareolide-mediated expression and this is a stress response mechanism which can result in programmed cell death, which is one of the main processes in xylogenesis (Grec et al. 2003). A second stem element first identified in bean (Keller and Baumgartner 1991) was also located here (motif ES13). A pollen-specific element (motif ES10) and a number of light responsive elements were also identified in this dataset (motifs ES14, ES20 and CESP1). Some well-known plant transcription factor binding sites such as DOF (ES3) and MADS (ES7 and ES12) sites were also identified in this dataset.

The Co-Expressed data sets were included in this study because they contain the promoters of genes that are co-expressed with the *CesA* genes during primary and

secondary cell wall formation in *Arabidopsis* (Persson et al. 2005). It is speculated that the promoters of the co-expressed genes may share motifs with the promoters of the *CesA* genes with which they are co-expressed. Motif CESP1 (Table 3.7-3.10 and Appendix 2.1) was identified all four datasets and was found to have similarity to an element involved in photo-regulation (Bruce et al. 1991). The CESP2 motif was identified in both *CesA* sets and in Co-expressed set 2. CESP2 (Table 3.7-3.10 and Appendix 2.2) showed high similarity to an element involved in pollen production (Hamilton et al. 1998).

### *Motif abundance*

The *in silico* methods of motif identification used in this study are all based on a motifs number of occurrence in the dataset, where a high number of occurrences indicate a significant motif. The motif with the highest number of occurrences in *CesA* set 1 was CEP2 and in *CesA* set 2, CS1 had the highest number of occurrences. Motif CP4 (Appendix 2.11) was the most abundant motif identified in *CesA* set 1 and the second most abundant of all the motifs identified in the two *CesA* data sets. The third most abundant motif was CS1 identified in *CesA* set 2 with 51 occurrences. Twelve of the motifs identified in *CesA* set 1 and *CesA* set 2 had more than 20 occurrences in their respective datasets, and is an average of two motifs per promoter (Figure 3.2).

### *The motifs spatial distribution in the promoters of the CesA genes*

It is important to note the spatial distribution of motifs as their position in the promoter can often effect the functioning of the gene. For example a motif may only have an effect if it is a specific distance from the transcriptional start site and this type of spatial distribution is often conserved among distantly related species (von

Gromoff et al. 2006). In Figure 3.3, 3.4 and 3.5 the spatial distribution of a number of the motifs identified in the study were mapped to the promoters. The orthologous promoters from *Arabidopsis*, *Populus* and *Eucalyptus* have been grouped together to observe any conservation that may occur in the motif occurrences and positioning.

CesA set 1 contains two sequence versions of the *EgCesA5* (*EgCesA5A* and *EgCesA5B*) promoter (Chapter 2). *EgCesA5B* has a 200 bp deletion between –550 and –750 (Figure 3.3-3.6). When the motifs were mapped to the promoters it became apparent that there were very few motifs predicted in this region and for these two promoters most of the predicted motifs are within the first 550 bp of sequence (Figure 3.3 and 3.4). In Figure 3.3 the orthologous Group B has the occurrences of 4 motifs mapped to the promoters and out of a total 41 occurrences, 37 were within the first 550 bp (+1 to -550). Only one motif was found in the region of the deletion in all of the promoters of group B and only three motifs were located after 750 bp. Similarly of the five motifs identified in CesA set 1 represented here, there were a total of 56 occurrences represented in the orthologous promoter group B and the majority (39) of these motifs were found to be in the region of +1 to –550 bp (Figure 3.4).

Five motifs (CESP1, CESP2, CSP1 and CSP2) were identified in both CesA set 1 and CesA set 2, these motifs were mapped on to the promoters to identify possible positional conservation (Figure 3.3). Firstly the mapping of these motifs (CESP1, CESP2, CSP1 and CSP2) showed that the total number of predicted motifs decreased with distance from the predicted transcriptional start site (Figure 3.3). In all of the promoters analyzed there were only 14 motif occurrences in the region from –800 to –1000 as opposed to the 88 motif occurrences located in the region of +1 to –200 out of a total of 272 motif occurrences. Motif CESP2 appeared to be evenly spread through

all of the promoter regions but there were more occurrences in the promoters of the genes involved in secondary cell wall formation than in the promoters of the genes involved in primary cell wall formation (Figure 3.2 and Figure 3.3).

Motif CSP1 occurrences appeared to decrease sharply from an average of ten occurrences every 100 bp in the region from +1 to – 600 to an average of 3 motifs every 100 bp in the region from –600 to –1000. CSP1 had twice the number of occurrence (2 motifs per promoter) in the promoters involved in primary cell wall formation than in the promoters of the genes involved in secondary cell wall formation (Figure 3.2). In Group A CSP1 displayed a pair of conserved occurrences less than 50 bp apart in the region –100 to –150 in both the *Eucalyptus* and *Populus Cesa* promoters. Group B also displayed conservation in the region +1 to –50 and – 100 to –200 where the motif was located in all four promoters from *Arabidopsis*, *Populus* and *Eucalyptus*. Group C and the promoters in Cesa set 2 do not display conservation in the position of this motif (Figure 3.3).

Motif CESP1 appeared to occur in roughly the same region (-100 to –200) in all of the promoters with 15 occurrences in this region out of 42 occurrences. The remaining 27 occurrences were spread through the rest of the promoters. CESP1 appears to occur in conjunction with CESP2 in a number of the Cesa set 1 promoters, but this is not the case in the Cesa set 2 promoters (Figure 3.3). Unlike the other motifs discussed thus far, CSP2 did not appear to be clustered at a specific region in any of the promoters analyzed. Only in the region –600 to – 700 did there appear to be some slight conservation where four of the *Eucalyptus* promoters, two of the *Populus* promoters and one *Arabidopsis* promoter out of the 19 promoters had an occurrence in approximately the same position. Also it appeared that CSP2 occurred 13 times out of

15 occurrences as a pair with CESP2 in Cesa set2, but this relationship was not observed in Cesa set 1 (Figure 3.3).

*Motifs exclusive to Cesa set 1 and their spatial distribution in the promoters*

A number of the motifs identified during this study were only over-represented in one of the datasets when compared to the background or the other data sets. Figure 3.4 maps the occurrences of six motifs only CEP1, CEP2 and CEP3 were mapped because they were identified in both Cesa set 1 and co-expressed set 1 and so are likely to be involved in primary cell wall formation. CP1 was selected for mapping because it showed the highest similarity to a known motif. CP3 and CP5 (Appendix 2.10 and 2.12) were selected for mapping on the promoters as they both showed similarity to well known stem- or vascular-specific elements. CP1 shows some conservation with the majority of its occurrences occurring in the region +1 to -550. In orthologous Group B promoters in the region -200 to -250 in all four promoters there was one CP1 occurrence and Group C displayed a similar pattern with all three promoters containing CP1 in the region -300 to -350. In Group A, the *Eucalyptus* and *Arabidopsis* promoters contained CP1 in the region -200 to -250 but CP1 was completely absent from the *Populus* promoter of this group (Figure 3.4).

There did not seem to be any positional conservation of CEP3 although there appeared to be a relationship between CEP3 and CP4 where approximately half of the CEP3 occurrences were paired with CP4 occurrences. CP4 also lacked positional conservation, but was highly abundant in the promoters with most promoters having four or more occurrences. CEP1 also showed little conservation among the different orthologous promoters, but an interesting feature to note is that half of the occurrences were in the region -150 to -350 and no motifs occur in any of the promoters in the

region +1 to -200. CEP2 also did not show significant positional conservation, but it appeared to be present in all the *Arabidopsis* and *Eucalyptus* promoters but was not present in two of the three *Populus* promoters (Figure 3.4).

#### *Motifs exclusive to CesA set 2 and their spatial distribution in the promoters*

The five (CS1, CS2, CS3, CS4 and CS5) motifs identified in CesA set 2 with the most similarity to known motifs were mapped to the promoters (Appendix 2.25-2.29). The motifs found in CesA set 2 appeared to form clusters of different motifs in at least eight of the nine promoters analysed. CS1 is a highly abundant motif with each promoter having approximately five occurrences. CS1 showed an interesting distribution, where Group D had the majority of motifs occurring in the region +1 to -500, but in Groups E and F most of the motifs occur in the region -500 to -1000. All of the promoters in Group D have at least one CS1 occurrence in the region -300 to -400. The promoters of Group E all have at least one CS1 occurrence in region -500 to -550 and the promoters of Group F also have at least one occurrence in region -750 to -800 (Figure 3.5).

Motif CS2 was not as abundant as CS1 but there were some interesting correlations that could be seen on the motif map (Figure 3.5). CS2 did not show conservation among the homologous promoters in Group E, but it did form part of a tight cluster consisting of different motifs in the *Populus* promoter (*PtrCesA2*) in the region +1 and -50. Group F displayed this motif as part of a cluster of motifs that appeared to show some conservation among the *Populus* and *Eucalyptus* promoters in the region -150 to -200. The *Populus* and *Arabidopsis* promoters of group D had a similar cluster of motifs in the same region (-150 to -200) and these four clusters both contain CS2 and CS5 in the same order. If one looks at group D separately it is also possible that

the cluster of motifs in the region +1 to -100 in the *Eucalyptus* promoter and region -100 to -200 in the *Populus* promoter were conserved regions in these promoters (Figure 3.5).

CS4 produced two striking features on the motif map (Figure 3.5). The motif was far more abundant in the promoter groups E and F when compared to group D and the second striking feature of this motif was that in the region -500 to -700 in all three *Populus* promoters there appeared to be a conserved region where the motif was repeated two or three times in close succession. CS3 appeared in general to be randomly distributed through the dataset. Although, in the region -200 to -300 six of the nine promoters contain this motif suggesting a possible positional conservation of this motif in the Cesa set 2 promoters. CS5 formed part of the conserved cluster of motifs in the region -150 to -200 in four of the six promoters of group D and F and appeared to have a higher rate of occurrence in these two groups when compared to the occurrences in the promoters of group E. Another interesting conformation produced by this motif occurs in region -700 to -800 where it mirrors the pattern produced by CS1 in this region offset by only approximately 20 bases (Figure 3.5).

A number of motifs had a high similarity to related motifs in the PLACE database. Motifs CS4 and CSP1 showed similarity to IDE1 and IDE2 (Iron Deficiency Element) respectively (Kobayashi et al. 2003). These elements have been found to work in conjunction to confer a specific expression pattern and Figure 3.6 shows a difference in the occurrences of these motifs in Cesa set 1 and 2. CSP1 is present in both datasets while CS4 is only over-represented in Cesa set 2. In Cesa set 2 the CS4 and CSP1 occurrences appear to be conserved at a number of positions in the



orthologous promoters. In the region –250 to –450 of *CesA* set 2 CS4 and CSP1 are approximately 100 bp apart in 5 of the nine promoters (Figure 3.6).

### 3.5 Discussion

Cellulose is the most abundant biopolymer on earth and is used in a number of important industries including the pulp and paper industry. Cellulose is a product of a large complex of enzymes (cellulose synthases), which deposit the cellulose into the cell wall. A number of cellulose synthase genes have been isolated from different plants including *Arabidopsis* (Burn et al. 2002), *Populus* (Samuga and Joshi 2002) and *Eucalyptus* (Ranik and Myburg 2006). Expression studies of these genes revealed that three cellulose synthase genes are associated with the secondary cell wall formation while at least three other cellulose synthase genes are associated with primary cell wall formation (Turner and Somerville 1997; Burton et al. 2004; Persson et al. 2005; Ranik and Myburg 2006).

Although the expression patterns of the *CesA* genes have been well documented there is little information on the regulation of these genes. The cis-regulatory elements involved in the differential expression of these genes are unknown and the promoter regions have not yet been analyzed in different plant species. This Chapter reports the results of a comparative bioinformatics study identifying possible regulatory features in the *CesA* gene promoters of *Eucalyptus*, poplar and *Arabidopsis* plants. In this study a number of key regulatory elements involved in stem-specific expression as well as a number of novel motifs with unknown function were predicted in the *CesA* gene promoters. This is the first comparative study in which the orthologous *CesA* promoter regions were analyzed for motifs conserved among distantly related plant species. A problem that may arise is that *Arabidopsis* is herbaceous where as *Populus*

and *Eucalyptus* have woody stems, thus the three species may not have regulatory elements involved in secondary cell wall formation in common.

Four datasets were used in this study for the identification of possible cis-regulatory motifs. The first data set (CesA set 1) contained the orthologous cellulose synthase promoters associated with primary cell wall formation in *Arabidopsis*, *Eucalyptus* and *Populus*. The second set (CesA set 2) contained all of the *Eucalyptus*, *Populus* and *Arabidopsis CesA* promoters associated with secondary cell wall formation. These two datasets were used to identify motifs that were over-represented in the *CesA* promoters. The motifs over represented in the CesA set 1 and CesA set 2 were compared to differentiate motifs that were shared between the two sets as well as motifs that were only over-represented in one of the two sets. In this way motifs associated with primary or secondary cell wall formation can be predicted. Unfortunately, at the time the datasets were constructed the full-length promoter sequences of *EgCesA2* and *EgCesA7* were not yet available and they were not included in the initial comparison. These promoter sequences were scanned separately for the presence of motifs predicted in the CesA datasets, but this is not optimal as different methods of motif identification were used and some motif occurrences maybe omitted in these promoters.

Two other datasets were also included in this study; they were compiled of promoters from *Arabidopsis* genes that were shown to be co-expressed with the *CesA* genes (Persson et al. 2005). The first set (Co-Expressed set 1) contained the promoters of 17 genes highly co-expressed with the *AtCesA* genes, which are associated with primary cell wall formation. Likewise the second dataset (Co-Expressed set 2) was compiled of promoters of 17 genes highly co-expressed with the *AtCesA* genes thought to play a

role in secondary cell wall formation. These two datasets were included in the study because it is expected that the promoters of co-expressed genes would share a number of cis-regulatory elements. These two datasets were also divided according to genes associated with primary and secondary cell wall formation and so the motifs identified in these datasets could be compared with the motifs identified in Cesa set 1 and Cesa set 2. As these genes share expression patterns it is expected that their promoters will share some key regulatory motifs involved in primary and secondary cell wall formation.

Three different programs (Weeder, POCO and MotifSampler) were used for the motif identification and because they were based on different motif prediction algorithms it increased the accuracy of the predictions (Tompa et al. 2005). The statistical analysis performed by POCO uses bootstrapping with replacement and this lead to slight variations in the results to ensure the accuracy of the results the analysis was performed three times on each dataset and only the motifs identified in all three permutations were used for further analysis. Another problem with *in silico* analyses is that they are only as accurate as the models on which the algorithm is based and in the case of motif analysis the available models still require vast refinements. Thus a number of the motifs identified in this study will prove not to play a role in the tissue-specific expression of the *Cesa* genes. Also building consensus sequences may lead to a sequence being incorporated that could be a separate motif and so will be lost.

One problem, which comes with studying the promoters of genes involved in wood formation in trees such as *Eucalyptus* or poplar, is that most motif prediction tools only incorporate herbaceous plant backgrounds such as *Arabidopsis* or rice. All the software packages used in this study made use of an *Arabidopsis* promoter

background comprised of the promoters in the entire *Arabidopsis* genome. This may skew the results slightly and motifs specific to woody-stemmed plants maybe missed. The prediction of the TSS in Chapter 2 can also lead to errors, in cases where a gene has a long 5'UTR or multiple TSSs are present the TSS prediction software may produce false positives and this could lead to the incorporation of 5'UTR regions in the datasets. Also by not including the region between the TSS and start codon a number of core promoter elements may be missed. But, the focus of this study is on the upstream cis-regulatory elements and not on the core promoter cis-elements.

A final set of motif sequences were generated for each of the datasets and in the two Co-Expressed sets and CesA set 1 over 20 different motifs were predicted per dataset (Tables 3.7, 3.9 and 3.10). The motifs identified in CesA set 1 and Co-Expressed set 1 were compared in order to identify motifs in common and the motifs identified in CesA set 2 and Co-Expressed set 2 were compared to identify motifs shared between the datasets. CesA set 2 only generated 12 motifs (Table 3.8), but there may be a number of different reasons for this. CesA set 2 was the smallest dataset analysed and thus there was less sequence to be analysed and fewer motifs could be identified. There was no overlap in the motifs identified by CesA set2 and Co-expressed set 2. This may be because Co-expressed set 2 is comprised of only *Arabidopsis* promoters, while two thirds of the promoters in CesA set 2 are from *Eucalyptus* and *Populus*. *Eucalyptus* and *Populus* have woody stems and produce great quantities of cellulose, but *Arabidopsis* is herbaceous and only produces wood when induced by stress. Therefore it is likely that *Arabidopsis* has different regulatory elements involved in secondary wood formation compared to that of *Eucalyptus* and *Populus*.

### *Motif location in the Cesa promoters*

Motifs identified in different *Cesa* promoters produced some interesting findings (Figure 3.3, 3.4 and 3.5). The overall impression obtained from the mapping of the motifs is that motif density is highest in the region close to the TSS. This is very apparent in Figure 3.3 and to a lesser extent it is also true for the motifs mapped in Figure 3.4. This is to be expected because even though in theory cis-regulatory elements can be located many kilobases upstream the majority of motifs reported to date are within 200 to 300 base pairs upstream of the transcriptional start site (Turner and Sommerville 1997; Burton et al.2004; Persson et al.2005; Ranik and Myburg 2006).

Two different promoter regions for *EgCesa5* were included in Cesa set 1 because during the isolation of the *EgCesa5* promoter it was found that there were two alleles that differed by a 200 bp indel (Chapter 2). Both versions of the promoter were included in the dataset to determine if this 200 bp region contained any predicted motifs, which may lead to differential expression of the allelic variants. Almost all of the motifs identified in Cesa set 1 mapped to either side of the 200 bp region in both promoters (Figure 3.3 and 3.4). The majority of the motifs could be found in the region +1 to -550 and of all the motifs mapped only CEP2, CSP1, CP5, CP7, CP10 and CP13 (Appendix 2.4, 2.6, 2.12, 2.14, 2.17 and 2.20) had an occurrence in the region of the deletion (-550 to -750).

In *EgCesa5A* CEP1 (Appendix 2.3) occurs in position -800 but in *EgCesa5B* CEP1 occurs at position -650 and has moved 200 bp closer to the transcriptional start site (Figure 3.3). The repositioning of CEP1 may influence the expression of the *EgCesa5B* allele because some motifs only function when in a specific position in the

promoter (de Meaux et al. 2005). Motif CEP2 (Appendix 2.4) also showed variation between the two alleles of *EgCesA5*. CEP2 occurred at two positions in *EgCesA5A* (-600 and -820) but in *EgCesA5B* CEP2 only occurred once in position -650 (Figure 3.4 and Appendix 2.4). The 200 bp indel in this region appears to have deleted one of the CEP2 occurrences and moved the second CEP2 site 200 bp closer to the TSS (transcriptional start site). The deletion and repositioning of CEP2 site could cause differential expression of these alleles but these results will have to be functionally tested (Figure 3.4).

#### *Presence of motifs in multiple datasets*

Motifs common to all the datasets were identified, which is predictable because all the promoters are expressed in vascular tissues and will have some main motifs in common. *CesA* set 1 and Co-expressed set 1, shared 3 motifs, which expected because both sets were comprised of promoters of genes expressed during primary cell wall formation. The two *CesA* promoter datasets had two motifs in common and this may be because they are all promoters of genes which synthesis a similar product and therefore may share some regulatory elements. There was not a large overlap in motifs identified from the different datasets possibly due to the motifs identified in Co-Expressed set 1 and 2 being more specific to *Arabidopsis* promoters. Only key motifs that are highly conserved among species will be identified in both the co-expressed and *CesA* datasets.

Motif CESP1 (Appendix 2.1) was identified in all four datasets where *CesA* set 1 had the highest number of occurrences and *CesA* set 2 had the lowest (Figure 3.2). This difference in CESP1 abundance may indicate a role for CESP1 in the spatial-temporal expression of the cellulose synthase genes. CESP1 showed similarity (Tables 3.3-3.6)

to a motif in the PLACE database known as PE3 (**P**ositive **E**lement **3**). PE3 was identified in the promoters of phytochrome genes and is a light responsive element. This element is 31 bp in length and CESP1 showed similarity to the 5' region of the PE3 element. This is of interest because the first 14 bp of this element contains a motif that is critical for its functioning. PE3 is a positive element, which enhances gene expression during dark periods (Bruce and Quail 1990). This could corroborate the function of CESP1 as it has been shown that cellulose synthase is differentially expressed during different periods of the day with the expression peaking at night (Solomon, Ranik, unpublished results). A number of the CESP1 occurrences appeared to be in the same region (between -100 and -200) of the different *CesA* promoters (Figure 3.3) and PE3 also showed positional conservation in the same region of the phytochrome promoters (Bruce et al. 1991), suggesting similar regulatory mechanisms.

Motif CESP2 (Appendix 2.2) was identified by both *CesA* datasets and Co-expressed set 1, with *CesA* set 2 having twice as many occurrences as *CesA* set 1 (Figure 3.2). CESP2 showed similarity to a pollen-specific element identified in maize pollen gene promoters (Table 3.3 - 3.5). This element was isolated from the promoter of the *ZM13* gene, which is highly expressed in maturing maize pollen (Hamilton et al. 1998). This may be of interest to as pollen cells also undergo secondary cell wall formation during the maturation process and thus it is possible that a similar motif may direct expression in the cellulose gene promoters. A transcription factor known to play a major role in pollen formation in *Arabidopsis* was over-expressed and caused ectopic tracheary element formation indicating that pollen transcription factors or closely related family members may play a role in wood formation (Mitsuda et al. 2005). Many transcription factors are known to work in conjunction with other transcription

factors forming dimers and binding to the DNA in tandem (Zhou 1999; Choi et al. 2005). In Figure 3.3 it is clear that CESP2 occurs as part of a pair with CSP2 in the promoters of Cesa set2 but this is not the case in Cesa set 1 promoters. Suggesting that these two motifs may bind transcription factors in a heterodimer conformation to confer the differential expression patterns.

Motif CSP2 (Appendix 2.7) was identified in both Cesa set 1 and Cesa set 2 with twice as many occurrences in Cesa set 2 than in Cesa set 1 (Figure 3.2). In a similarity search of the PLACE database this motif was found to be similar to an element known as the GCC-box (Table 3.7 and 3.8). This motif has been identified in a number of ethylene responsive genes (Solano et al. 1998). Ethylene is a plant hormone that plays a role in a number of plant development processes such as senescence, cell elongation and the determination of cell fate (Kieber 1997). All of these processes are also involved during wood formation, of which cellulose deposition is an important feature (Eyles et al. 2003). The EIN3 protein, which binds to the GCC-box has been shown to have a similar DNA binding domain to the domain found in the APETALA2 (AP2) protein (Solano et al. 1998). AP2 plays an integral role in the stem cell maintenance in the shoot apical meristem (Wurschum et al. 2006).

Motifs CS4 and CSP1 (Appendix 2.28 and 2.6) were found to have a high similarity to two related elements in the PLACE database, iron response elements one and two (IDE1 and IDE2) respectively (Table 3.7 and 3.8). CSP1 was over-represented in both Cesa set 1 and Cesa set 2 while CS4 was only over-represented in Cesa set 2 (Figure 3.2). IDE1 and IDE2 were found in the promoters of genes that play a role in the iron deficiency response pathway (Kobayashi et al. 2003). One of the key



enzymes in this pathway is S-adenosylmethionine synthase (SAMS). This is interesting because SAMS has also been found to play a major role in the lignin biosynthetic pathway (Shen et al. 2002). SAMS is co-expressed with the *CesA* genes associated with the secondary cell wall formation (Sterky et al. 1998; Hertzberg et al. 2001; Ranik et al. 2006). Also studies show that in the rice iron deficiency gene promoters IDE1 and IDE2 are approximately 100 bp apart (Kobayashi et al. 2005) and in *CesA* set 2 CSP1 and CS4 were also approximately 50 - 100 bp apart (Figure 3.6). This suggests that CS4 and CSP1 may be similar to the IDE elements and could play a role in the regulation cellulose biosynthesis.

#### *Motifs identified in CesA set 1*

Motifs such as CP1 (Appendix 2.8) were identified to be over-represented in *CesA* set 1 and showed high similarity to elements in the PLACE database (Appendix 2.8). This motif was found to have similarity to an element that resembles a R2R3 MYB binding site in the promoter of a maize anthocyanin gene (Hernandez et al. 2004). Anthocyanin is a product of the flavonoid biosynthetic pathway and is a color pigment deposited in the flowers of plants. Lignin is another major product of the flavonoid biosynthetic pathway and a different R2R3 MYB has been found to play an important role lignin deposition (Goicoechea et al. 2005). CP1 may represent the binding site of a MYB involved in primary cell wall cellulose deposition. CP1 occurred frequently in the *CesA* set 1 promoters in the region -100 to -300 (Figure 3.4). This region overlaps with the regions in which, the petunia MYB binds to the promoters of the anthocyanin biosynthetic pathway (Hernandez et al. 2004) and the region in *EgCAD* and *EgCCR* promoters to which the EgMYB binds (Goicoechea et al. 2005).

Motif CP4 (Appendix 2.11) is of interest to this study because it was found to have some similarity to an element identified in bean that is a general enhancer of vascular specific genes (Table 3.7). On its own it does not confer tissue-specific expression, but acts as a general enhancer. When present in conjunction with a second element it causes the gene to be highly expressed in vascular tissue. CP4 was identified in the promoters of the *CesA* genes involved in primary cell wall formation. The expression of these genes is not confined to one specific tissue of the plant and they need to be expressed at high levels explaining the identification of a number of different enhancers in this dataset. The SE2 (Stem element 2) element, which confers vascular tissue-specific expression (Keller and Baumgartner 1991), was identified in Co-Expressed set 2 (Table 3.10). This is interesting as most of these genes are involved in secondary cell wall formation and would need to be highly expressed in the vascular tissue. There could be a number of reasons why this element was not detected in *CesA* set 2, such as there were very few *Arabidopsis* promoters in this dataset and perhaps the element involved in the vascular tissue-specific expression has a very different sequence in poplar and *Eucalyptus* and is not conserved among the species. Another reason could be that these two elements were first isolated from a bean plant, which is also herbaceous and while *Arabidopsis* and beans may share this mechanism of regulation, woody plant species may utilize a different mechanism of regulation.

CEP1 (Appendix 2.3) is of interest to us because it had a high significance (Table 3.3) and was identified in *CesA* set 1 and Co-Expressed set 1 (Tables 3.7 and 3.9). These motifs were similar to an element identified in the promoter of the asparagine synthase gene, which was responsible for the light-repression of the gene (Bruce et al. 1991). This fits well with the identification of the motif similar to PE3 (CESP1), because PE3 is involved the induction of expression in the dark and so these two

elements could work in conjunction to ensure that the genes are only expressed during the dark hours. This is important, because as stated earlier the *CesA* genes are highly expressed during the dark hours (Solomon, Ranik unpublished results). It was also found that promoters with this element confined the GUS expression to the vascular tissue (mainly phloem) making this element of great interest to us because in a previous study it was found that out of all the *CesA* genes, *EgCesA4* and *EgCesA5* (in *CesA* set 1) were the most highly expressed in phloem (Ranik and Myburg 2006).

The motif CEP2 (Appendix 2.4) was identified in *CesA* set 1 and Co-expressed set 1 and were found to be similar to part of a maize element found in gene promoters involved in the anthocyanin biosynthetic pathway (Tables 3.7 and 3.9). This element contains a binding site for the VP1 (Viviparous1) protein and an abscisic acid response motif (Kao et al. 1996). CEP2 overlapped with the abscisic acid response motif (CGTGTC), but also partially overlapped with the VP1 binding site (CGTCCATGCAT). This element forms part of a complex regulatory mechanism in the genes involved in anthocyanin biosynthesis. Anthocyanin biogenesis is part of the flavinoid biosynthetic pathway, which is also responsible for the production of lignin. This is the second element we have identified that showed some similarity to elements involved in the regulation of the anthocyanin pathway. CEP1 and CEP2 have a very similar motif consensus sequence, but hit on 2 different motifs on the PLACE database as discussed above. The two studies (Bruce et al. 1991; Kao et al. 1996) from which the elements were identified may in fact have identified the same motif under different circumstances, because the one element is an abscisic response element and abscisic acid has been shown to play a role in a number of different plant processes.

CEP3 (Appendix 2.5) was identified in *CesA* set 1 and in both of the Co-Expressed sets (Figure 3.2). When used in a homology search of the PLACE database it was found to have similarity (Tables 3.7, 3.9 and 3.10) to AS1 (Activation sequence 1). This sequence was first identified in the 35S promoter of the cauliflower mosaic virus and it does not require binding of any viral proteins to perform its function (Benfey et al. 1989; Benfey et al. 1990). This suggests that the plant contains proteins, which can bind to this motif in the viral DNA and this suggests that the plant proteins must also bind a similar motif in the plant DNA. The region of homology between this element and CEP3 overlapped with one of two TGACG repeats in the element that were necessary for the binding of the plant protein, activation sequence factor-1 (ASF-1). It has also been noted that on its own this factor confers high root specific expression, but when found in conjunction with other upstream regions it conferred high leaf specific expression (Lam et al. 1989). This element has also been found in conjunction with IDE (Iron deficiency response elements) elements (Kobayashi et al. 2005), which we have also identified during this study (CSP1 and CS4). This suggests that CEP3 may be involved in enhancing the expression of the *CesA* genes during primary cell wall formation.

#### *Motifs identified in CesA set 2*

Motifs that were highly abundant in *CesA* set 2 (Table 3.4) and had the highest similarity to elements listed in PLACE (Table 3.8) were mapped to the promoters (Figure 3.5). The first of these motifs is CS1 (Appendix 2.25), which had a high significance and a high number of occurrences (Figure 3.2 and 3.5). This motif was not identified in any of the other datasets and showed high similarity to a phloem-specific element first identified in a promoter of the Rice Tungro Bacilliform virus (Yin and Beachey 1995). This element, known as Box II, or RNFG2 (Rice nuclear

factor) binding site has been found to carry some of the key elements known to be involved in phloem- and xylem-specific expression (Yin and Beachey 1997). RNFG2 (CCAGTGTGCCCTG) contains a number of motifs also found in the well-known AC II element (CCACCACCCC). Motif CS1 overlaps partially with both the CCA and CCCC regions of the RNFG2 element (Appendix 2.25). Table 3.8 only reports the highest similarity score of each motif, but CS1's second highest similarity score was for the AC II element (Hatton et al. 1995), which may add to the suggestion that CS1 is related to the AC-elements. Other phloem gene promoters have also been reported to contain the AC II and AC I elements and it has been shown that these elements and RNFG2 all confer vascular tissue specificity (Yin et al. 1997) but that it is a combination of these elements in the promoters that result in xylem or phloem specific expression suggesting hetero- and homodimer binding.

It has been proposed that perhaps phloem evolved before xylem (Yin and Beachey 1995). This could be why the promoters of xylem-specific genes contain a number of phloem suppressing elements, but to date no xylem-suppressing elements have been identified in the promoters of phloem specific genes. On the other hand there are many transcription factors and their binding sites still need to be identified and so xylem-suppressing elements may still be identified (Yin et al. 1995). Two proteins (RF2a and RF2b) have been identified that form a heterodimer and bind to RNFG2. These proteins belong to the b-ZIP family of transcription factors and other members of this family have been shown to bind to the ACII and ACI elements of lignin genes aiding in their tissue-specific expression (Dai et al. 2003). This is supported by the close proximity of some of the CS1 occurrences (less than 50bp apart in Figure 3.5). Thus it is not surprising that similar AC-type elements (CS1) were identified in the

promoters of the *CesA* genes as they display similar expression patterns to many of the key lignin genes.

CS2 (Appendix 2.26) was over-represented in *CesA* set 2 and showed similarity to an element predicted to be involved in the negative regulation of *PHYA* (**Phytochrome A**) gene promoters (Terzaghi and Cashmore 1995). This motif was identified in one of the first *in silico* motif prediction studies. Bruce et al. (1991) identified a number of motifs known to play a role in *PHYA* regulation such as the G-box motif were also identified suggesting that the *in silico* isolation method was accurate and this adds to the significance of the repression element identified. The *in silico* analysis used here also identified a similar element. This could indicate that the predicted motif is an aberration of the programming in the *in silico* analyses, but since three different software packages were used and this should ensure the motifs validity

Motif CS3 (Appendix 2.27) showed high similarity to an element in the promoter region of the *CAB2* (Chlorophyll a/b binding protein 2) gene (Table 3.8). This gene is also regulated by the phytochrome system and has a number of light responsive elements (Anderson and Carol 2004). This element contained a GATA motif to which CS3 showed similarity and the GATA like motif binds a CGF-1 (CAB GATA Factor 1) protein (Villain et al. 1994). The element in the *CAB2* gene promoter was found to contain 3 GATA elements and the CS3 sequence overlapped with one of the three elements. This may suggest the existence of a similar element in the *CesA* set 2 promoters although CS3 is only 8 bp long and the element from the PLACE database is 32 bp long and so further analysis of the surrounding sequence is required. The GATA motifs are variable and so this may not represent an element with a similar function but a different element also containing a number of GATA repeats because

these motifs confer many different and diverse functions (Zhou et al. 1999). CS3 and CS2 both showed similarity to elements thought to play a role in phytochrome mediated regulation and if one looks at the positioning of the CS2 and CS3 motif occurrences, in six of the nine promoters, the two motifs are found within 100 bp of each other (Figure 3.5). This could indicate a function of these elements in the expression of the *CesA* set 2 genes if not by light then in another capacity and should be investigated further.

CS5 (Appendix 2.29) showed weak similarity (Table 3.8) to an element found to play a role in anaerobic response of the *GapC4* (gluceraldehyde-3-phosphate dehydrogenase 4) gene (Geffers et al. 2000). This motif remained of interest as it had a high significance (Table 3.4) and when mapped to the promoters showed some interesting conservation patterns. CS5 was identified in seven of the nine promoters and was less than 50 bp away from CS1 (Figure 3.5). In the promoter group F at position -750 CS5 was also less than 50 bp from CS1, but even more striking is the fashion in which it mirrors the pattern produced by CS1 (Figure 3.5). It was discovered that the anaerobic element contained a number of GT-motifs. CS1 was shown to have high similarity to a well known xylem-specific element and it is known that a number of different elements must work in conjunction to ensure xylem- or phloem-specific expression and so it is possible that CS5 and CS1 could bind transcription factor dimers in order to confer the xylem-specific expression pattern observed in the *CesA* set 2 genes, but this speculation will have to be followed up with an in-depth investigation of the sequences involved.

### *Putative new and novel motifs identified*

A number of the motifs identified in the *CesA* promoters (CP7, CP8, CP9, CP10, CP11, CP12, CP13, CP14, CP15, CP16, CP17, CS5, CS6, CS7 and CS8) show very little or no similarity to any of the motifs located on the PLACE or PlantCARE databases (Appendix 2.14-2.24 and 2.29-2.32). This is not so surprising as this is the first *CesA* promoter study and to date only a few wood specific elements have been identified (Summary in Chapter 1). The weak similarity shown by some motifs to elements on place may indicate that the motifs bind different proteins in the same family as the proteins that bind to the PLACE motifs. Among these motifs are some interesting candidates, such as CS8 and CS6, which appear to occur in pairs with themselves and since some transcription factors form homodimers when binding with the DNA these may be interesting motifs that could regulate secondary cell wall formation. These motifs will be of interest as there may be some novel motifs, which are important to the regulation of the *CesA* genes that have not yet been studied.

This was a novel study because it was the first time that the *CesA* promoters have been comparatively studied among distantly related species (*Arabidopsis*, *Populus* and *Eucalyptus*). A number of key stem regulatory motifs have been identified and with further testing maybe useful in manipulating the *CesA* gene expression. Also a number of novel or new motifs may have been identified and these may lead to the isolation of novel DNA binding proteins that play a role in the regulation of the *CesA* genes.

### **3.6 Acknowledgements**

The authors would like to thank the designers of the software used in this study in particular Dr G. Pavesi for his prompt response and help with queries pertaining to



Weeder and its output. We would also like to thank M. Ranik for helpful suggestions pertaining to software troubleshooting and guidance with regards to the compilation of the datasets. This work was supported by funding provided by Mondi Business Paper South Africa, through the Wood and Fibre Molecular Genetics Programme, the Technology and Human Resources for Industry Programme (THRIP) and the National Research Foundation of South Africa (NRF).

### 3.7 Tables

**Table 3.1 NCBI accession numbers and gene names of the *Eucalyptus*, *Populus* and *Arabidopsis* cellulose synthase genes of which the promoters were used.**

Data set	Orthologous Groups <sup>a</sup>	Gene Name	Accession number	Reference
CesA set 1	Group A	<i>EgCesA4</i>	DQ014508	(Ranik and Myburg 2006)
		<i>PtrCesA5</i>	AY055724	(Kalluri and Joshi 2003)
		<i>AtCesA3</i>	AF027174	(Arioli et al. 1998)
	Group B	<i>EgCesA5A</i>	DQ014509	(Ranik and Myburg 2006)
		<i>EgCesA5B</i>	DQ014509	(Ranik and Myburg 2006)
		<i>PtrCesA4</i>	AY162181	(Kalluri and Joshi 2003)
		<i>AtCesA1</i>	AF027172	(Arioli et al. 1998)
	Group C	<i>PtrCesA7</i>	AY162180	(Samuga and Joshi 2002)
		<i>AtCesA2</i>	AF027173	(Arioli et al. 1998)
		<i>AtCesA5</i>	NM_121024	(Asamizu et al. 1998)
		<i>AtCesA6</i>	NM_125870	(Desmos et al. 1996)
	CesA set 2	Group D	<i>EgCesA1</i>	DQ014505
<i>PtrCesA1</i>			AF072131	(Wu et al. 2000)
<i>AtCesA8</i>			NM_117994	(Turner and Somerville 1997)
Group E		<i>EgCesA3</i>	DQ104507	(Ranik and Myburg 2006)
		<i>PtrCesA2</i>	AY095297	(Samuga and Joshi 2002)
		<i>AtCesA4</i>	AF458083	(Taylor et al. 2003)
Group F		<i>PtrCesA3</i>	AF527387	(Kalluri and Joshi 2004)
		<i>AtCesA7</i>	AF088917	(Taylor et al. 1999)

<sup>a</sup> The promoters are grouped into orthologous groups based on their nearest neighbours in the phylogenetic tree presented in Figure 2.1.

**Table 3.2 TAIR locus identifiers of the genes whose promoters were used to construct the two Co-expressed data sets used in the motif analysis.**

Data set	Gene Name or putative function <sup>a</sup>	TAIR locus identifier
Co-expressed set 1	<i>AtCesA3</i>	At5G05170
	<i>AtCesA1</i>	At4G32410
	<i>AtCesA6</i>	At5G64740
	<i>COBRA</i>	At5G60920
	Transporter related gene	At1G76670
	Dehydration response-like gene	At1G04430
	<i>CLT1</i>	At1G05850
	Glycerophosphoryl diester phosphodiesterase family	At4G26690
	Dehydration response like	At1G29470
	<i>AtCesA2</i>	At4G39350
	Phosphotranslocator related gene	At1G12500
	Endomembrane protein 70 gene	At5G35160
	Glycosyl transferase family 8 protein	At3G62660
	Expressed protein	At4G39840
	Expressed protein	At2G41770
	Squalene monooxygenase	At1G58440
	Glycosyl transferase family 2	At4G31596
	Leucine-rich repeat protein kinase	At4G18030
	Expressed protein	At1G45688
	Mitogen-activated protein kinase	At2G42880
Co-expressed set 2	<i>AtCesA4</i>	At5G44030
	<i>AtCesA7</i>	At4G18780
	Glycosyl transferase family 8	At5G54690
	<i>AtCesA8</i>	At5G17420
	Putative Laccase	At2G38080
	<i>CTL1-like</i>	At3G16920
	<i>COBL4</i>	At5G15630
	<i>FLA11</i>	At5G03170
	Glycosyl transferase family 43	At2G37090
	Glycogenin glucosyl transferase like	At3G18660
	Expressed protein	At4G27435
	Expressed protein	At5G60720
	<i>GCP10</i>	At3G62020
	<i>NAM</i> family	At4G28500
	Putative Laccase	At5G60020
	<i>FLA12</i>	At5G60490
	Leucine-rich repeat kinase	At1G79620
	Exostosin family protein	At1G72440
	Expressed protein	At1G09610
	Expressed protein	At1G79470

<sup>a</sup> Gene list and putative function as referenced in Persson et al. (2005)

**Table 3.3 Motifs identified in Cesa set 1 by at least two of the three software programs (Weeder, POCO and MotifSampler).** The three major columns represent the outputs of each program. In each case the first column contain the motif sequence identified by the program. The second column indicates the number of times the motifs occurred in the set of promoters and the last column gives a measure of the significance of the motif occurrence provided by the software.

Motif Identity <sup>a</sup>	MotifSampler			POCO			Weeder		
	Sequence <sup>b</sup>	Occurrences <sup>c</sup>	Log likelihood <sup>d</sup>	Sequence <sup>b</sup>	Occurrences <sup>c</sup>	P-Value <sup>e</sup>	Consensus Sequence <sup>f</sup>	Occurrences <sup>c</sup>	Significance <sup>g</sup>
CP1	A-GSKGGYKS	17	60.48				C-GGGGGTGG	22	0.44
	B-GGRKGYKG	18	66.28				I-GGGTGG	59	1.4
	D-GGGTGG	16	72.90						
	E-SGSTGS	18	86.34						
	F-NGCTGG	16	80.73						
	G-SGGTGG	18	77.83						
	H-CCANCC	15	57.48						
CP2	A-CNCNCCTC	12	50.37	B-CNCNNCNC	39	2.81E <sup>-02</sup>			
				C-CNCTNCNC	20	2.63E <sup>-02</sup>			
CP3	A-CCNMCCC	16	63.85				B-CCACCCCC	22	0.44
	C-CCCMCY	14	60.17				D-CCCCCC	22	0.95
CP4	A-GNCASGTN	14	69.29	I-GGTGNNGC	5	3.80E <sup>-03</sup>	B-GACAGTGC	6	0.62
	H-GGKGARGY	15	74.01	E-CNGNCNGT	12	3.94E <sup>-04</sup>	G-GGTGGGGC	22	0.41
	D-NWGTCCKGT	14	64.32	F-CNGTCNNT	17	4.03E <sup>-03</sup>			
	C-CWGKCTGT	22	85.06						
CP5	A-ATTWATTA	20	67.27	B-ATNTNTTA	39	4.81E <sup>-03</sup>			
CP6	A-GCNTGC	19	69.76	G-GCNGGC	40	4.24E <sup>-02</sup>			
	B-GCWWGC	20	76.42	H-GNANGC	44	2.18E <sup>-02</sup>			
	C-GCWNGC	20	77.98	I-GCNTNC	44	2.18E <sup>-02</sup>			
	E-GCAWGC	18	75.46						
	D-GCANGC	20	76.97						
	F-GCWKGC	20	77.17						
CP7	A-NGRCAGTG	14	64.08	B-TGNCNGTG	6	2.42E <sup>-03</sup>			
CP8	A-GYGCTC	15	65.50				B-GCGCTC	19	1.05
CP9	A-GAGCGM	11	52.74				B-GAGCGC	20	1.05
CP10	A-GTCKGT	17	72.27	B-GTCNGT	20	4.13E <sup>-04</sup>			
CP11	A-AGNGAYAG	15	61.08	B-ANTGNNAG	25	2.08E <sup>-02</sup>			
				C-ANNGACNG	17	4.03E <sup>-03</sup>			
				C-ACNGNCNG	12	3.94E <sup>-04</sup>			
CP12	A-ASAGNCTG	19	87.23						
	B-ACAGRCWG	20	92.71						
CP13	A-TTTTTT	19	58.01	B-TTTTTT	202	4.08E <sup>-02</sup>			
				C-TTTTNT	328	4.10E <sup>-02</sup>			
CP14	A-AAAAAA	20	51.56	B-AAAAAA	202	4.08E <sup>-02</sup>			
				C-AAAANA	328	4.16E <sup>-02</sup>			

Motif Identity <sup>a</sup>	MotifSampler			POCO			Weeder		
	Sequence <sup>b</sup>	Occurrences <sup>c</sup>	Log likelihood <sup>d</sup>	Sequence <sup>b</sup>	Occurrences <sup>c</sup>	P-Value <sup>e</sup>	Consensus Sequence <sup>f</sup>	Occurrences <sup>c</sup>	Significance <sup>g</sup>
CP15	A-SYCSWSCC	16	72.07	C-CNCNNCNC	47	2.80E <sup>-02</sup>			
	B-CNCCMCCN	12	64.71						
CP16	A-TNGCTKTC	7	39.36	B-TNNCNTNC	68	2.10E <sup>-02</sup>			
CP17	A-GGSWSGGS	17	83.80	C-GNGNNGNG	47	2.81E <sup>-02</sup>			
	B-GGSWSGRS	15	80.93						
CEP1	A-TGTSGSTN	16	63.43	B-TGNCNGTG	6	2.42E <sup>-03</sup>	C-TGTCGGTG	19	0.46
CEP2	A-TGTCKG	14	65.24				B-TGTCGG	61	1.12
	C-CMGACA	12	45.73				D-CCGACA	61	1.12
CEP3	A-GNCASTGN	15	73.31				B-GACAGTGG	28	0.45
CSP1	A-TCTGTM	14	62.04				B-TCTGTC	28	0.96
	D-GASAGA	16	65.78						
	E-SACAGA	16	71.87						
	C-GACAGA	17	76.22						
	B-CKCCCC	9	45.86				A-CGCCCC	22	1.40
CSP2	B-CKCCCC	9	45.86	B-GCNCACC	5	3.80E <sup>-03</sup>	C-GCTCCACC	22	0.35
CESP1	A-GCTNNMSY	22	88.52				C-AGCCAGC	19	0.72
CESP2	A-AGNSSAGN	20	93.65						
	B-MGCMAGCY	14	62.10						

<sup>a</sup> Motif identity is represented by a series of letters and numbers (C= Cesa set, E= Co-expressed set, P= Primary cell wall associated gene promoters, S= Secondary cell wall associated gene promoters) the numbers were given in the order they were listed. The motif ID refers to the motif consensus sequence (Appendix 2), which was constructed from the different motifs (A, B, C...etc), listed in this table.

<sup>b</sup> Motif sequence predicted by the software (N=A/T/G/C, M=A/C, W=A/T, R=A/G, Y=C/T, S=C/G and K=G/T). The bold lettering at the beginning of each motif sequence represents the motifs position in the motif alignments from which the consensus motif sequence was constructed (Appendix 2).

<sup>c</sup> Number of times the motif was identified in Cesa set 1 consisting of 11 promoters sequences.

<sup>d</sup> Log likelihood score of the motifs predicted by MotifSampler, a measure of the quality of the motif and this depends on the strength of the motif and the total number of instances of the motif (Tompa et al. 2005).

<sup>e</sup> P-value of the motif predicted by POCO, a measure of the significance of each motif by estimating the probability of identifying this motif in a random set of sequences (Kankainen and Holm 2005).

<sup>f</sup> Consensus sequence for each motif as predicted by Weeder.

<sup>g</sup> Measure of significance as predicted by the Weeder algorithm that is specific for transcription factor binding sites and is used to rank the results in order to identify the best motifs. This measure of significance accounts for the number of sequences the motif appears in, how conserved the motif is and the overall number of occurrences in the input set. The higher the statistical score the more significant the predicted motif (Tompa et al. 2005).

**Table 3.4 Motifs identified in Cesa set 2 by at least two of the three software programs (Weeder, POCO and MotifSampler).** The three major columns represent the outputs of each program. In each case the first column contain the motif sequence identified by the program. The second column indicates the number of times the motifs occurred in the set of promoters and the last column gives a measure of the significance of the motif occurrence provided by the software.

Motif Identity <sup>a</sup>	MotifSampler			POCO			Weeder		
	Sequence <sup>b</sup>	Occurrences <sup>c</sup>	Log likelihood <sup>d</sup>	Sequence <sup>b</sup>	Occurrences <sup>c</sup>	P-Value <sup>e</sup>	Consensus Sequence <sup>f</sup>	Occurrences <sup>c</sup>	Significance <sup>g</sup>
CS1	A-RNYSTGCC	14	56.83	N-ANCNTNNC	77	9.35E <sup>-4</sup>	Q-GCTGTGCC	19	0.74
	B-RMCNTGCC	15	57.25	O-CNCNNCNC	40	2.81E <sup>-2</sup>	R-GGGCACAG	24	0.65
	C-RCYSTGCC	14	58.13	P-CNCTNCNC	20	2.63E <sup>-2</sup>	S-GTGCCC	24	1.15
	D-RCYNNGCC	15	61.12						
	E-RCNNGGCC	16	62.98						
	K-CNSWGCCC	14	63.76						
	L-CYSWGCCC	12	60.22						
	M-STGCC	13	54.60						
	F-GGCNNRKN	12	56.70						
	G-GGCASRGY	13	62.91						
	H-RGGCASRG	14	56.86						
	I-GGGCANNAG	14	61.14						
	J-GGGCASNG	13	53.56						
	CS2	B-CARMAGGA	12	48.22	D-CATGNC	21	3.62E <sup>-2</sup>	A-CAGCAGGA	27
C-YNTGCC		15	64.47						
CS3	A-NCNGMAGG	16	61.56				C-GCTGAAGG	25	0.25
	B-SY TSAAGN	12	55.36						
CS4	A-NNGCATGC	10	50.32				D-GAGCATGC	18	0.25
	B-GCATGC	10	46.18				E-GCATGC	28	1.43
	C-GCANGC	15	62.04						
CS5	A-TCCTKYTG	11	51.03				B-TCCTGCTG	25	0.37
CS6	A-NNTTSAAG	9	35.07	B-AANTNAAG	12	1.24E <sup>-2</sup>			
CS7	A-GNNCAGAG	13	57.92	B-GNGNNGNG	39	2.81E <sup>-2</sup>			
				C-GNGNAGNG	20	2.63E <sup>-2</sup>			
				C-ANCNTNNC	77	9.35E <sup>-2</sup>			
CS8	A-AGSTWANC	12	50.53						
	B-RNCYTRCC	13	63.65						
CSP1	F-SMTKCTGT	11	62.64	I-ANGNCATG	12	1.59E <sup>-4</sup>	J-CCTGCTGT	28	0.41
	G-WMRGCAKG	9	41.04				K-ACAGCAGG	28	0.41
	H-WCAGMAKN	9	44.20						
CSP2	C-GGCANGKN	12	55.87	D-GNNANGNT	77	9.35E <sup>-4</sup>			
CESP1	O-GCTGATGK	10	56.06				P-GCTGATGG	19	0.24

Motif Identity <sup>a</sup>	MotifSampler			POCO			Weeder		
	Sequence <sup>b</sup>	Occurrences <sup>c</sup>	Log Likelihood <sup>d</sup>	Sequence <sup>b</sup>	Occurrences <sup>c</sup>	P-Value <sup>e</sup>	Consensus Sequence <sup>f</sup>	Occurrences <sup>c</sup>	Significance <sup>g</sup>
CESP2	<b>F</b> -GGCAGR	15	57.81				<b>H</b> -GGCAGG	28	1.4
	<b>G</b> -GGCWGG	12	49.75				<b>K</b> -GGGCTG	7	1.21
	<b>E</b> -GGCAGG	13	51.37						
	<b>I</b> -SSGCWG	14	55.27						
	<b>J</b> -SNGGWW	15	59.97						
	<b>D</b> -GGGCTG	13	50.08						
	<b>L</b> -YCWGCC	15	64.63						
	<b>N</b> -YCTGCY	15	58.96						
	<b>O</b> -CCTGCC	14	58.11						
	<b>P</b> -SWGCCC	12	50.14						

<sup>a</sup> Motif identity is represented by a series of letters and numbers (C= Cesa set, E= Co-expressed set, P= Primary cell wall associated gene promoters, S= Secondary cell wall associated gene promoters) the numbers were given in the order they were listed. The motif ID refers to the motif consensus sequence (Appendix 2), which was constructed from the different motifs (A, B, C...etc), listed in this table.

<sup>b</sup> Motif sequence predicted by the software (N=A/T/G/C, M=A/C, W=A/T, R=A/G, Y=C/T, S=C/G and K=G/T). The bold lettering at the beginning of each motif sequence represents the motifs position in the motif alignments from which the consensus motif sequence was constructed (Appendix 2).

<sup>c</sup> Number of times the motif was identified in *Cesa* set 2 consisting of 8 promoters sequences.

<sup>d</sup> Log likelihood score of the motifs predicted by MotifSampler, a measure of the quality of the motif and this depends on the strength of the motif and the total number of instances of the motif (Tompa et al. 2005).

<sup>e</sup> P-value of the motif predicted by POCO, a measure of the significance of each motif by estimating the probability of identifying this motif in a random set of sequences (Kankainen and Holm 2005).

<sup>f</sup> Consensus sequence for each motif as predicted by Weeder.

<sup>g</sup> Measure of significance as predicted by the Weeder algorithm that is specific for transcription factor binding sites and is used to rank the results in order to identify the best motifs. This measure of significance accounts for the number of sequences the motif appears in, how conserved the motif is and the overall number of occurrences in the input set. The higher the statistical score the more significant the predicted motif (Tompa et al. 2005).

**Table 3.5 Motifs identified in Co-expressed set 1 by at least two of the three software programs (Weeder, POCO and MotifSampler).** The three major columns represent the outputs of each program. In each case the first column contain the motif sequence identified by the program. The second column indicates the number of times the motifs occurred in the set of promoters and the last column gives a measure of the significance of the motif occurrence provided by the software.

Motif Identity <sup>a</sup>	MotifSampler			POCO			Weeder		
	Sequence <sup>b</sup>	Occurrences <sup>c</sup>	Log likelihood <sup>d</sup>	Sequence <sup>b</sup>	Occurrences <sup>c</sup>	P-Value <sup>e</sup>	Consensus Sequence <sup>f</sup>	Occurrences <sup>c</sup>	Significance <sup>g</sup>
EP1	A-MRGYGG	26	110.86				B-AAGTGG	197	0.63
EP2	A-KGTGCG	21	92.30				B-GTGTCC	142	0.97
EP3	A-KGTGCGKY	20	112.16	C-ANNACNA	320	1.41E <sup>-2</sup>	D-TTGTCCGC	68	0.28
EP4	B-RMCGACAM	25	120.89				E-ACCGACAC	93	0.99
	A-GTCGKT	22	94.09				C-GTCGGT	75	0.95
EP5	B-RMCGAC	24	109.04				D-ACCGAC	92	0.95
	A-TGATTA	21	83.17	C-TNANTA	468	4.67 E <sup>-2</sup>			
EP6	B-TAATCA	24	85.05	D-TANTNA	468	4.67 E <sup>-2</sup>			
	A-CGSGTY	26	121.35				B-CGCGTT	140	0.65
EP7	A-MTCANATC	19	84.51	B-ANNAATC	107	1.73 E <sup>-2</sup>			
EP8	A-GWSAGTGA	22	92.76	B-GNNANTNA	239	4.19 E <sup>-2</sup>			
EP9	A-AAGWARAC	17	96.79	B-AANNNGNC	158	2.03 E <sup>-2</sup>			
EP10	A-GATTSC	30	117.39	B-GATNNC	112	2.81 E <sup>-2</sup>			
EP11	A-TTTATTTW	25	109.99	C-TNTANTNA	195	1.02 E <sup>-2</sup>			
	B-TAWNTTAA	28	104.74	D-TNTATTNA	77	2.49 E <sup>-2</sup>			
EP12				E-TATNNTNA	179	1.15 E <sup>-2</sup>			
	A-CRGTGR	22	100.28	B-CANTNA	245	4.98 E <sup>-2</sup>			
EP13	A-GGTTWA	22	91.36	B-GGNAA	173	3.65 E <sup>-2</sup>			
EP14	A-TAWTTA	26	89.89	B-TNANTA	468	4.67 E <sup>-2</sup>			
EP15	A-YCNCYGTC	30	120.22	D-TNANTNNC	239	4.19 E <sup>-2</sup>			
	B-NCACNGRC	31	130.07	E-TNNCNGNC	70	1.36 E <sup>-2</sup>			
EP16	C-YCACYNWC	27	120.09						
	A-TGSYGGYG	20	81.98	B-TNNTNCG	111	1.94 E <sup>-2</sup>			
EP17	A-NGTSAC	28	104.94						
CEP1	D-NTGTGCKT	17	96.37	G-TNGTNNNT	320	1.41E <sup>-2</sup>	H-CTGTCCGT	197	0.34
	E-KGTGCKT	27	123.34				I-GTGTCCGT	197	0.99
CEP2	F-TGNNGGTG	23	105.28				J-TGTCCGTG	92	0.95
							K-TGACGGTG	97	0.49
CEP3	E-CGACAM	26	113.70				G-CGACAC	92	0.97
	F-TGTGCK	25	102.74	D-TNANTG	245	4.98 E <sup>-2</sup>	H-TGTCCG	197	1.11



Motif Identity <sup>a</sup>	MotifSampler			POCO			Weeder		
	Sequence <sup>b</sup>	Occurrences <sup>c</sup>	Log likelihood <sup>d</sup>	Sequence <sup>b</sup>	Occurrences <sup>c</sup>	P-Value <sup>e</sup>	Consensus Sequence <sup>f</sup>	Occurrences <sup>c</sup>	Significance <sup>g</sup>
CESP1	<b>D</b> -GWCGGTGR	21	94.95	<b>I</b> -GNCNGNNA	70	1.36 E <sup>-2</sup>			
	<b>E</b> -GWCGGTSR	20	103.10						
	<b>F</b> -GNCGGTGR	20	92.97						
	<b>H</b> -GWSRGTTN	23	107.12						
CESP2	<b>Q</b> -RMCGAC	24	109.04				<b>R</b> -GTCGGC	197	0.75

<sup>a</sup> Motif identity is represented by a series of letters and numbers (C= Cesa set, E= Co-expressed set, P= Primary cell wall associated gene promoters, S= Secondary cell wall associated gene promoters) the numbers were given in the order they were listed. The motif ID refers to the motif consensus sequence (Appendix 2), which was constructed from the different motifs (A, B, C,...etc), listed in this table.

<sup>b</sup> Motif sequence predicted by the software (N=A/T/G/C, M=A/C, W=A/T, R=A/G, Y=C/T, S=C/G and K=G/T). The bold lettering at the beginning of each motif sequence represents the motifs position in the motif alignments from which the consensus motif sequence was constructed (Appendix 2).

<sup>c</sup> Number of times the motif was identified in Co-expressed set 1 consisting of 20 promoters sequences.

<sup>d</sup> Log likelihood score of the motifs predicted by MotifSampler, a measure of the quality of the motif and this depends on the strength of the motif and the total number of instances of the motif (Tompa et al. 2005).

<sup>e</sup> P-value of the motif predicted by POCO, a measure of the significance of each motif by estimating the probability of identifying this motif in a random set of sequences (Kankainen and Holm 2005).

<sup>f</sup> Consensus sequence for each motif as predicted by Weeder.

<sup>g</sup> Measure of significance as predicted by the Weeder algorithm that is specific for transcription factor binding sites and is used to rank the results in order to identify the best motifs. This measure of significance accounts for the number of sequences the motif appears in, how conserved the motif is and the overall number of occurrences in the input set. The higher the statistical score the more significant the predicted motif (Tompa et al. 2005).

**Table 3.6 Motifs identified in Co expressed set 2 by at least two of the three software programs (Weeder, POCO and MotifSampler).** The three major columns represent the outputs of each program. In each case the first column contain the motif sequence identified by the program. The second column indicates the number of times the motifs occurred in the set of promoters and the last column gives a measure of the significance of the motif occurrence provided by the software.

Motif Identity <sup>a</sup>	MotifSampler			POCO			Weeder		
	Sequence <sup>b</sup>	Occurrences <sup>c</sup>	Log likelihood <sup>d</sup>	Sequence <sup>b</sup>	Occurrences <sup>c</sup>	P-Value <sup>e</sup>	Consensus Sequence <sup>f</sup>	Occurrences <sup>c</sup>	Significance <sup>g</sup>
ES1	A-GGNAMMGY	22	118.41				D-GGAACAGC	59	0.51
	B-GGWAMNGY	22	110.46						
	C-GCTGWTNN	18	89.33						
ES2	A-TAACTT	30	115.39	B-TAACNT	83	1.26 E <sup>-02</sup>	B-GAGTGT	169	0.59
ES3	A-GAGTGW	17	83.59						
ES4	A-CMAASACA	26	115.84	BCCAANNNA	93	7.17 E <sup>-03</sup>			
ES5	A-CMTTRC	23	103.03				C-CCTTAC	53	0.61
	B-SCTTRC	25	107.93						
ES6	A-TTTAYYTA	27	119.89	B-TNTANTNA	140	1.02 E <sup>-02</sup>	C-TTGCTTGG C-GGGAAC	53	0.25
				C-TNTATTNA	54	2.49 E <sup>-02</sup>			
			D-TNNANCTA	85	1.97 E <sup>-02</sup>				
			B-TNNNTNAC	258	1.29 E <sup>-02</sup>				
			BTNNNTTGG	93	7.17 E <sup>-03</sup>				
ES7	A-YCCAWAWC	27	126.89						
ES8	A-TTGCTTKG	22	90.14						
ES9	A-GGRAAC	20	85.43						
	B-GTTYCC	24	117.21						
ES10	A-TCYTYWYC	26	121.641	B-TNNNTNAC	258	1.29 E <sup>-02</sup>			
ES11	A-GTTRGKNA	28	106.62	B-GTNANNNA	258	1.29 E <sup>-02</sup>			
				C-GTNANNAA	97	2.87 E <sup>-02</sup>			
				D-GNNANTNA	231	4.19 E <sup>-02</sup>			
				C-TTNCCT	67	1.91 E <sup>-02</sup>			
ES12	A-TYMCCT	24	108.97						
	B-TTWCCT	35	126.82						
ES13	A-TWGCTTRN	18	74.23	C-TNNNTTGG	93	7.17 E <sup>-03</sup>	D-TGTCTTGG	53	0.24
	B-TGYCTTKG	27	109.57						
ES14	A-TGWGTG	22	81.36	B-TNANTG	186	4.98 E <sup>-02</sup>			
				C-TNAGTG	33	2.37 E <sup>-02</sup>			
ES15	A-GGTGAR	31	114.98	B-GGNNA	178	3.65 E <sup>-02</sup>			
ES16	A-NAGTTYCN	28	120.66	B-TAGNTNNA	85	1.97 E <sup>-02</sup>			
ES17	A-AGCTWA	25	106.34	C-AGNTNA	231	2.04 E <sup>-02</sup>			
	B-TKAGCT	17	78.87	D-TNANCT	231	2.04 E <sup>-02</sup>			
ES18	A-AAGTNATK	29	121.69	B-AAGNNANT	115	1.06 E <sup>-03</sup>			
				C-AAGTNNTT	43	2.82 E <sup>-02</sup>			
				D-AANTGNNT	63	4.68 E <sup>-03</sup>			
				E-AAGTNNT	109	1.34 E <sup>-02</sup>			
				C-TGGNNA	179	1.38 E <sup>-02</sup>			
ES19	A-TKGGGA	16	63.20						
	B-TGKGGGA	24	91.18						

*In silico prediction of cis-regulatory elements*

Motif Identity <sup>a</sup>	MotifSampler			POCO			Weeder		
	Sequence <sup>b</sup>	Occurrences <sup>c</sup>	Log likelihood <sup>d</sup>	Sequence <sup>b</sup>	Occurrences <sup>c</sup>	P-Value <sup>e</sup>	Consensus Sequence <sup>f</sup>	Occurrences <sup>c</sup>	Significance <sup>g</sup>
ES20	<b>A</b> -RAGTTA	26	96.74	<b>C</b> -ANGTTA	83	1.26 E <sup>-02</sup>			
	<b>B</b> -ARGTTA	23	98.66						
ES21	<b>A</b> -TTGKGG	15	53.07				<b>C</b> -TTGGGG	53	0.73
	<b>B</b> -TKGNNG	16	62.60						
ES22	<b>A</b> -AGGNAA	22	86.96	<b>C</b> -AGGNAA	67	1.91 E <sup>-02</sup>			
	<b>B</b> -AGGKRA	28	106.23	<b>D</b> -AGNTNA	231	2.04 E <sup>-02</sup>			
CESP1	<b>J</b> -ACMGCT	27	113.06				<b>K</b> -ACAGCT	212	0.6
	<b>L</b> -AGCTGT	12	62.55						
	<b>M</b> -AGCTRT	26	101.99						

<sup>a</sup> Motif identity is represented by a series of letters and numbers (C= Cesa set, E= Co-expressed set, P= Primary cell wall associated gene promoters, S= Secondary cell wall associated gene promoters) the numbers were given in the order they were listed. The motif ID refers to the motif consensus sequence (Appendix 2), which was constructed from the different motifs (A, B, C...etc), listed in this table.

<sup>b</sup> Motif sequence predicted by the software (N=A/T/G/C, M=A/C, W=A/T, R=A/G, Y=C/T, S=C/G and K=G/T). The bold lettering at the beginning of each motif sequence represents the motifs position in the motif alignments from which the consensus motif sequence was constructed (Appendix 2).

<sup>c</sup> Number of times the motif was identified in Co-expressed set 2 consisting of 20 promoters sequences.

<sup>d</sup> Log likelihood score of the motifs predicted by MotifSampler, a measure of the quality of the motif and this depends on the strength of the motif and the total number of instances of the motif (Tompa et al. 2005).

<sup>e</sup> P-value of the motif predicted by POCO, a measure of the significance of each motif by estimating the probability of identifying this motif in a random set of sequences (Kankainen and Holm 2005).

<sup>f</sup> Consensus sequence for each motif as predicted by Weeder.

<sup>g</sup> Measure of significance as predicted by the Weeder algorithm that is specific for transcription factor binding sites and is used to rank the results in order to identify the best motifs. This measure of significance accounts for the number of sequences the motif appears in, how conserved the motif is and the overall number of occurrences in the input set. The higher the statistical score the more significant the predicted motif (Tompa et al. 2005).

**Table 3.7 PLACE identities and putative functions of the over-represented motifs predicted by POCO, Motifsampler and Weeder in the promoters of the *CesA* genes expressed during primary cell wall formation (CesA set 1).**

Motif Identity <sup>a</sup>	Motif Sequence <sup>b</sup>	PLACE Identity <sup>c</sup>	Putative Function <sup>d</sup>	E-Value <sup>e</sup>
CP2	CNCNNCNC	ABRECE3ZMRAB28	Abscisic response element	0.00017
CP1	GGNGGTGG	ARELIKEGHPGDFR2	Anthocyanin regulatory motif	0.0018
CP3	CCNC(A/C)CCC	ACIIPVPAL2 ACII	Vascular-specific expression	0.039
CEP3	GNCA(C/G)TGA	AS1CAMV	Activation element	0.24
CSP1	GACAGAA(G/T)N	IDE2HVIDS2	Iron deficiency response motif	0.25
CP6	GC(A/T)NGC	LEGUMINBOXLEGA5	Legumin tissue-specific motif	0.31
CP4	GACNGT(C/G)NGTGGGGC	SE1PVGRP18	Stem enhancer element 1	0.39
CESP1	G(A/T)CGGTG(A/G)AGCTGTTG(G/T)	PE3ASPHYA3	Positive photo-regulation	0.54
CEP1	NTGTCCGTG	BOXBPSAS1	Negative photoregulation	0.63
CESP2	GA(C/G)GGCAGG	PSREGIONZMZM13	Pollen specific region	0.64
CP5	ATN(A/T)ATTA	C2GMAUX28	Auxin responsive element	0.65
CEP2	(A/C)TGTCGG	CGTGTSPHZMC1	Abscisic acid response motif	0.93
CP7	NG(A/G)CNGTG	C1GMAUX28	Auxin responsive element	1
CP8	G(C/T)GCTC	GLUTEBP2OS	Glutelin nuclear factor	1
CP9	GAGCG(A/C)	ASF1ATNOS	Nopaline synthase motif	1
CP10	GTC(G/T)GT	MYBCORE	Flavonoid MYB binding site	1
CP11	ANNGA(C/T)AG	ARE1	Antioxidant response element	1
CP12	ACAGNCNG	GARE4HVEPB1	Gibberellic acid response	1
CP13	TTTTTT	PYRIMIDINEBOXHVE	Gibberellic acid induction	1
CP14	AAAAAA	3AF1BOXPSRBCS3	Light responsive element	1
CSP2	GGGGC(A/G)NGNN	EIN3ATERF1	Ethylene response element	1.7
CP15	C(C/T)C(C/G)NCCC	No Hit	No hit	-
CP16	TNNCN(G/T)NC	No Hit	No hit	-
CP17	GG(C/G)(A/T)(C/G)G(A/G)(G/C)	No Hit	No hit	-

<sup>a</sup> Motif identity is represented by a series of letters and numbers (C= CesA set, E= Co-expressed set, P= Primary cell wall associated gene promoters, S= Secondary cell wall associated gene promoters) the numbers were given in the order they were listed. The number series is interrupted when the motif was identified in more than one data set and it was allocated a unique identity

<sup>b</sup> Consensus of the motif sequences represented in Appendix 2.

<sup>c</sup> Name of motif in PLACE to which the motif was most similar.

<sup>d</sup> Putative function of the most similar motif in PLACE.

<sup>e</sup> Qualities of motif match . The lower the value the more reliable the hit.

**Table 3.8 PLACE identities and putative functions of the over-represented motifs predicted by POCO, Motifsampler and Weeder in the promoters of the Cesa genes expressed during secondary cell wall formation (Cesa set 2).**

Motif Identity <sup>a</sup>	Motif Sequence <sup>b</sup>	PLACE Identity <sup>c</sup>	Putative Function <sup>d</sup>	E-value <sup>e</sup>
CSP1	GACAGAA(G/T)N	IDE2HVIDS2	Iron deficiency responsive motif	0.058
CS2	TCCTGC(C/T)G	SORLREPSAT	Light repression element	0.2
CS1	(A/G)C(C/T)(C/G)TGCCC	RNFG2OS	Phloem-specific expression	0.25
CS4	NNGCATGC	IDE1HVIDS2	Iron deficiency responsive motif	0.32
CS3	(C/G)CTGAAGG	CGF1ATCAB2	Circadian regulation (I-Box)	0.62
CESP2	GA(C/G)GGCAGG	PSREGIONZM2M13	Pollen specific region	0.64
CESP1	G(A/T)CGGTG(A/G)AGCTGTTG(G/T)	PE3ASPHYA3	Positive photo-regulation	1
CSP2	GGGGC(A/G)NGNN	EIN3ATERF1	Ethylene response element	1
CS6	NNNT(C/G)AAG	HSE	Heat shock response element	1
CS7	GNGNAGNG	-141NTG13	Auto-regulation	1
CS5	TCCT(G/T)(C/T)TG	ANAEROBICCISZMGAPC4	Anaerobic cis-regulatory seq	1.8
CS8	(A/G)N(C/G)(C/T)T(A/G)(C/G)C	No Hit	No hit	-

<sup>a</sup> Motif identity is represented by a series of letters and numbers (C= Cesa set, E= Co-expressed set, P= Primary cell wall associated gene promoters, S= Secondary cell wall associated gene promoters) the numbers were given in the order they were listed. The number series is interrupted when the motif was identified in more than one data set and it was allocated a unique identity

<sup>b</sup> Consensus of the motif sequences represented in Appendix 2.

<sup>c</sup> Name of motif in PLACE to which the motif was most similar.

<sup>d</sup> Putative function of the most similar motif in PLACE.

<sup>e</sup> Qualities of motif match. The lower the value the more reliable the hit.

**Table 3.9 PLACE identities and putative functions of the over-represented motifs predicted by POCO, Motifsampler and Weeder in the promoters of the Cesa genes expressed during primary cell wall formation (Co-expressed set 1).**

Motif Identity <sup>a</sup>	Motif Sequence <sup>b</sup>	PLACE Identity <sup>c</sup>	Putative Function <sup>d</sup>	E-Value <sup>e</sup>
CEP3	GNCA(C/G)TGA	AS1CAMV	Activation element	0.021
EP4	GTCC(G/T)T	SUREAHVISO1	Sugar responsive element (WRKY)	0.12
EP1	(A/C)(A/G)G(C/T)GG	PREMOTIFNPCABE	Photoregulated Expression	0.14
EP2	(G/T)TGTCG	B2GMAUX28	Activation element (As-1 motif)	0.14
EP3	(G/T)TGTCG(G/T)(C/T)	GAREHVAMY1	Gibberellic acid responsive element	0.27
CESP2	GA(C/G)GGCAGG	PSREGIONZM13	Pollen specific region	0.49
EP7	(A/C)TCAAATC	ELRECOREPCR1	Elicitor response element (WRKY)	0.57
CEP1	NTGTCGGTG	BOXBPSAS1	Negative photoregulation	0.64
EP6	CG(C/G)GT(C/T)	27BPDRCONSEN	Replication fork barrier	0.65
CEP2	(A/C)TGTCGG	CGTGTSPHZMC1	Abscisic acid response Sph element	0.93
EP5	TAATTA	BOX2PVCHS15	Cell type-specific regulation	0.98
EP8	G(A/T)(C/G)AGTGA	5659BOXLELAT5659	Modulation of gene activity	1
EP9	AAG(A/T)A(A/G)AC	HSELIKENTACIDICP	Heat shock element	1
CESP1	G(A/T)CGGTG(A/G)AGCTGTTG(G/T)	PE3ASPHYA3	Positive photo-regulation	1
EP10	GATT(C/G)C	OCSGMHSP26A	Osc-element	1
EP11	TNTATTNA	23BPUASNSCYCB1	Upstream activating sequence (MYB)	1
EP12	C(A/G)GTG(A/G)	INRNTPSADB	Initiator element	1
EP13	GGTT(A/T)A	RBCSBOX2PS	G-Box	1
EP14	TA(A/C)TTA	COREOS	Coordinate regulatory element	1
EP15	(C/T)CAC(C/T)GNC	No Hit	No Hit	-
EP16	TG(C/G)(C/T)GG(C/T)G	No Hit	No Hit	-
EP17	AGT(C/G)AC	No Hit	No Hit	-

<sup>a</sup> Motif identity is represented by a series of letters and numbers (C= Cesa set, E= Co-expressed set, P= Primary cell wall associated gene promoters, S= Secondary cell wall associated gene promoters) the numbers were given in the order they were listed. The number series is interrupted when the motif was identified in more than one data set and it was allocated a unique identity

<sup>b</sup> Consensus of the motif sequences represented in Appendix 2.

<sup>c</sup> Name of motif in PLACE to which the motif was most similar.

<sup>d</sup> Putative function of the most similar motif in PLACE.

<sup>e</sup> Qualities of motif match. The lower the value the more reliable the hit.

**Table 3.10 PLACE identities and putative functions of the over-represented motifs predicted by POCO, Motifsampler and Weeder in the promoters of the Cesa genes expressed during secondary cell wall formation (Co-expressed set 2).**

Motif Identity <sup>a</sup>	Motif Sequence <sup>b</sup>	PLACE Identity <sup>c</sup>	Putative Function <sup>d</sup>	E-Value <sup>e</sup>
ES1	GG(T/A)A(A/C)NGC	SB3NPABC1	Sclareolide-specific motif	0.0051
ES7	(C/T)CCA(A/T)A(A/T)C	AGL2ATCONSENSUS	Flowering (MADS-box)	0.022
ES4	C(A/C)AA(C/G)ACA	HBOXCONSENSUSPVCHS	Elicitor induction (H-Box)	0.25
ES6	TNTA(C/T)TTA	COREOS	Coordinate regulatory element	0.34
ES9	GG(A/G)AAC	NDEGMSAUR	Auxin response element	0.42
ES2	TAACNT	BOX1PVCHS15	Organ-specific expression	0.48
ES5	CCTT(A/G)C	AS1CAMV	Activation element	0.57
ES10	TC(C/T)T(C/T)(A/T)(C/T)C	VOZATVPP	Pollen development motif	0.94
ES11	GTNAN(G/T)NA	ELRECOREPCR1	Core elicitor response element	1
ES12	TTNCCT	AGL3ATCONSENSUS	Transcriptional regulation	1
ES13	TNNCTTGG	SE2PVGRP1	Stem- specific element 2	1
ES14	TNAGTG	GT1MOTIFPSRBCS	Photo activation element	1
ES15	GGTGA(A/G)	SRENTTQ1	Stress responsive element	1
ES16	TAGTT(C/T)CA	D1GMAUX28	Possible auxin response motif	1
ES17	T(G/T)AGCT(A/T)A	ASF1ATNOS	Nopaline synthase element	1
ES18	AAGTNNTT	MARCEN3	Centromere element	1
ES19	T(G/T)GGGA	DREDRIATRD29AB	Draught responsive element	1
ES20	A(A/G)GTTA	PE1ASPHYA3	General positive element	1
CESP1	G(A/T)CGGTG(A/G)AGCTGTTG(G/T)	PE3ASPHYA3	Positive photo-regulation	1
ES3	GAGTG(A/T)	OBP1ATGST6	Stimulates protein binding	1.1
ES8	TTGCTT(G/T)G	PASNTPARA	Cadmium response element	1.4
ES21	TTG(G/T)GG	No Hit	No hit	-
ES22	AGGNAA	No Hit	No hit	-

<sup>a</sup> Motif identity is represented by a series of letters and numbers (C= Cesa set, E= Co-expressed set, P= Primary cell wall associated gene promoters, S= Secondary cell wall associated gene promoters) the numbers were given in the order they were listed. The number series is interrupted when the motif was identified in more than one data set and it was allocated a unique identity

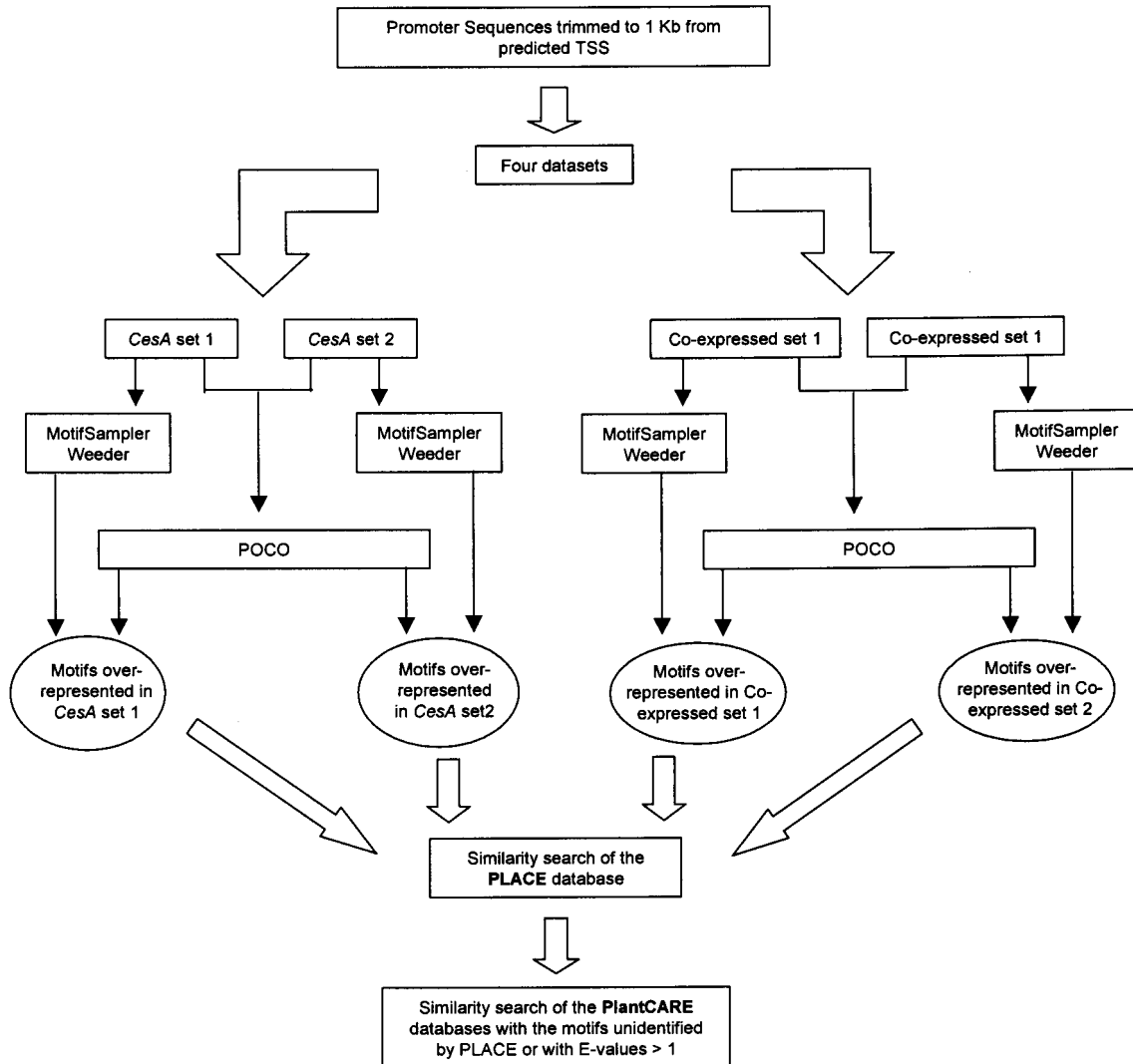
<sup>b</sup> Consensus of the motif sequences represented in Appendix 2.

<sup>c</sup> Name of motif in PLACE to which the motif was most similar.

<sup>d</sup> Putative function of the most similar motif in PLACE.

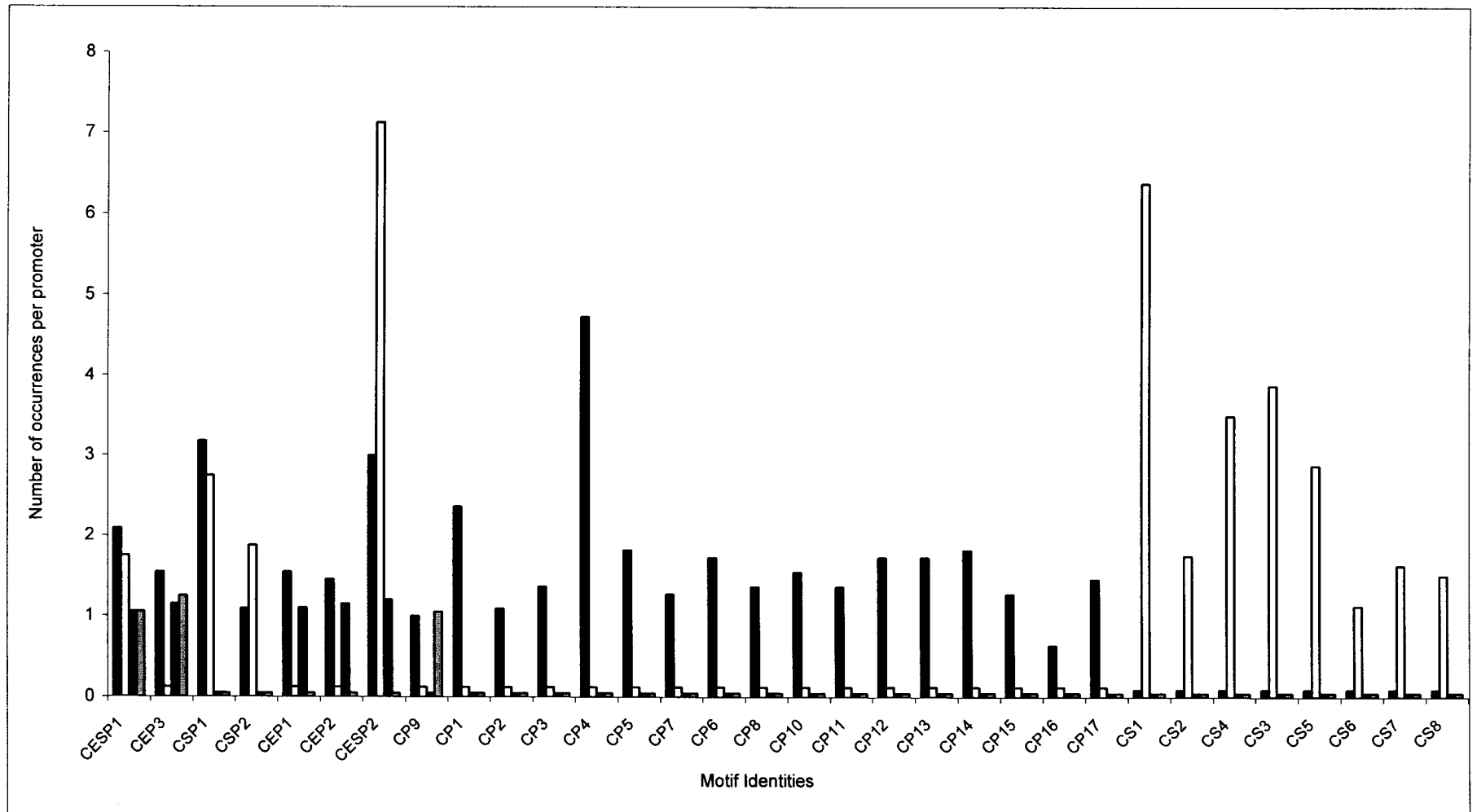
<sup>e</sup> Qualities of motif match. The lower the value the more reliable the hit.

### 3.8 Figures



**Figure 3.1** Flow diagram of the method used to identify motifs that are over-represented in the different promoter datasets. One kb of promoter sequence upstream of the predicted TSS was obtained for each gene investigated. The promoters were separated into four datasets. Three programs were used to analyze all of the datasets and motifs identified by two or more of the programs were used for further analysis. The motif sequences were used in similarity searches on the place database. Any of the motifs that had a similarity E-value of one or more, or were not similar to any motifs on PLACE, were used in searches of the PlantCARE database.



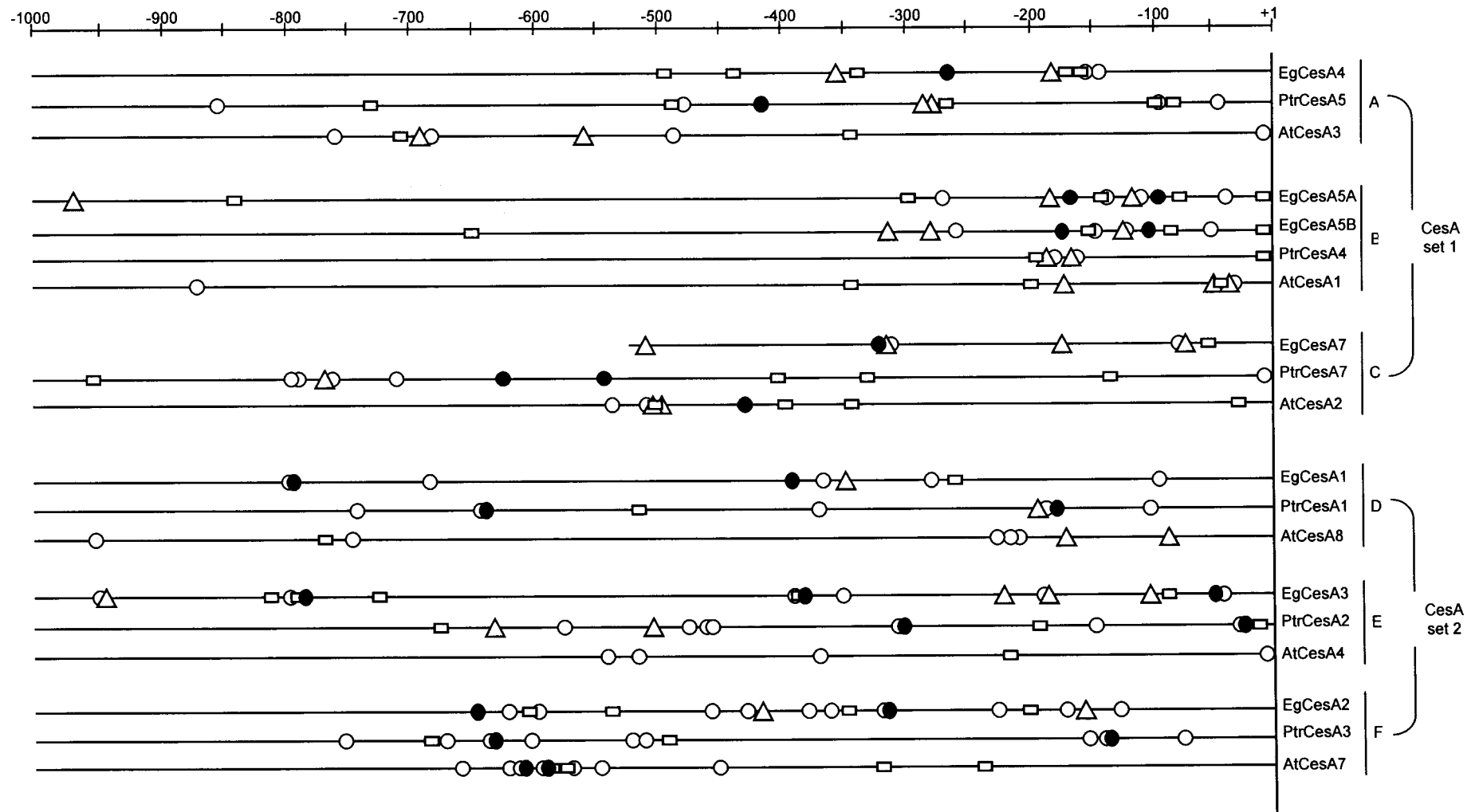


**Figure 3.2** Number of occurrences of each motif in the different data sets. Black bars indicate the motif occurrences in Cesa set 1, the white bars indicate the motif occurrences in Cesa set 2, the dark grey bars indicate the occurrences of the motif in Co-expressed set 1 and the light grey indicates the occurrences of the motifs in Co-expressed set 2. The Y-axis indicates the number of motif occurrences per promoter and the X-axis the motif identity.

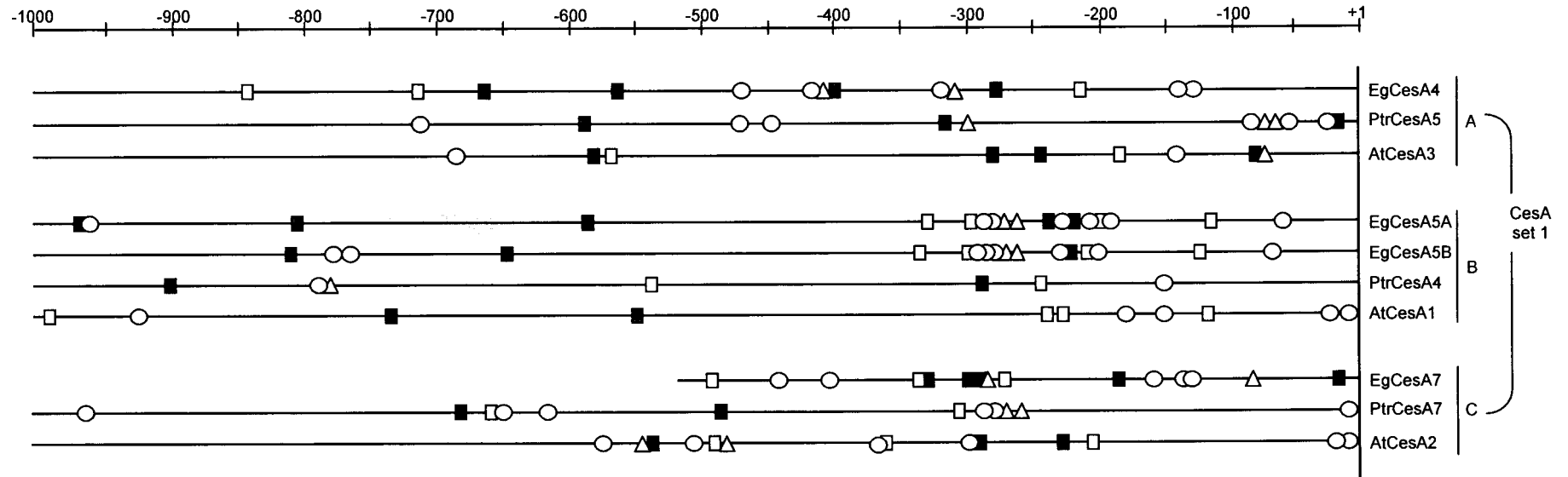
---

**Figure 3.3 Predicted motif lactations in the promoters where the motifs were identified in Cesa set 1 and 2.** Horizontal lines represent the cellulose synthase promoters, which are anchored on the right by a vertical line representing the predicted transcriptional start site (+1). Orthologous promoters are grouped together as indicated on the right. The first three sets of promoters (A, B and C) are promoters of genes involved in primary cell wall formation (Cesa set 1). The second three sets of promoters (D, E and F) are the promoters of *Cesa* genes involved in secondary cell wall formation (Cesa set 2). The gene names of the promoter regions are indicated on the right (Table 3.1). The top bar divides the promoters into 50 bp sections. The grey block on the *EgCesa5A* promoter (Group B) indicates the region in *EgCesa5A* that is deleted in *EgCesa5B*. The white, black and grey shapes correspond to the different motifs (CSP1□, CSP2●, CESP△, CESP2○)

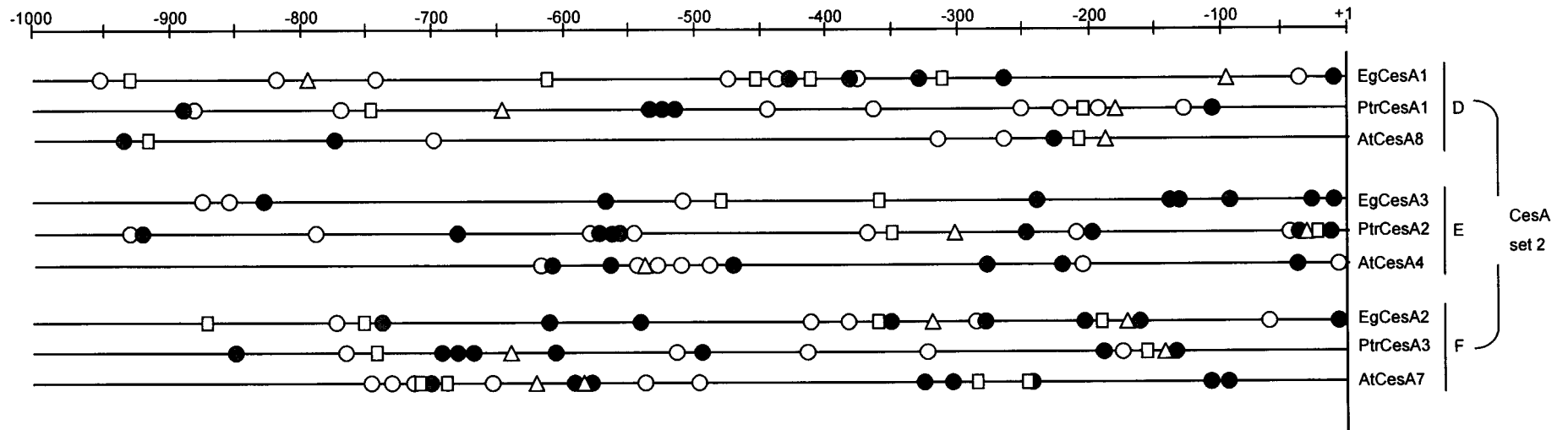
*In silico analysis of cis-regulatory elements*



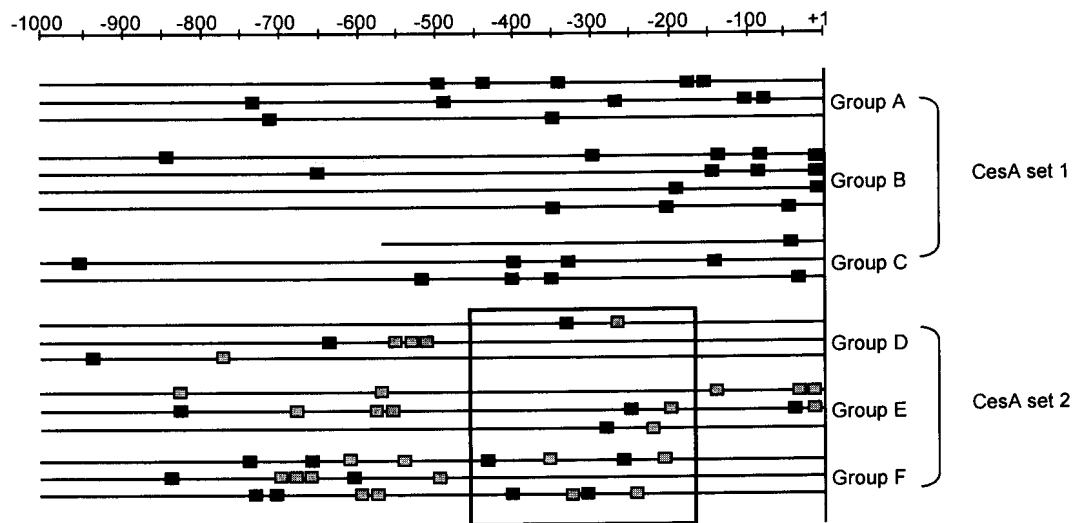
*In silico analysis of cis-regulatory elements*



**Figure 3.4 Predicted motif locations in the promoters where the motifs were identified in Cesa set 1 only.** Horizontal lines represent the cellulose synthase promoters, which are anchored on the right by a vertical line representing the predicted transcriptional start site (+1). Orthologous promoters are grouped together as indicated on the right. The promoters (A, B and C) are promoters of genes involved in primary cell wall formation (Cesa set 1). The gene names of the promoter regions are indicated on the right (Table 3.1). The top bar divides the promoters into 50 bp sections. The grey block on the *EgCesA5A* promoter (Group B) indicates the region in *EgCesA5A* that is deleted in *EgCesA5B*. The white, black and grey shapes correspond to the different motifs (CP□, CP4○, CEP1■, CEP2■, CEP3△).



**Figure 3.5 Predicted motif locations in the promoters where the motifs were identified in *CesA* set 2 only.** Horizontal lines represent the cellulose synthase promoters, which are anchored on the right by a vertical line representing the predicted transcriptional start site (+1). Orthologous promoters are grouped together as indicated on the right. The promoters (D, E and F) are the promoters of *CesA* genes involved in secondary cell wall formation (*CesA* set 2). The gene names of the promoter regions are indicated on the right (Table 3.1). The top bar divides the promoters into 50 bp sections. The white, black and grey shapes correspond to the different motifs (CS1○, CS2△, CS3●, CS4●, CS5□)



**Figure 3.6** Superimposed appendix 2 images of motifs CSP1 (**GACAGAA(G/T)N**) and CS4 (**NNGCATGC**). All the promoters are anchored at the predicted transcriptional start site (+1). The black blocks indicate the positions of CSP1 occurrences and the grey blocks indicate the positions of CS4 occurrences. The blocked off region of CesA set 2 shows conservation in occurrences of CSP1 and CS4 in this dataset. The orthologous promoter groups and data sets used are shown on the right of the figure. The top bar provides the scale and divides the promoters into 50 bp sections.

### 3.9 References

- Anderson, L. E. and Carol, A. A. (2004). Seven enzymes of carbon metabolism, including three Calvin cycle isozymes, are present in the secondary cell wall thickenings of the developing xylem tracheary elements in pea leaves. *Int J.Plant Sci.* 165:243-256
- Andersson, A., Keskitalo, J., et al. (2004). A transcriptional timetable of autumn senescence. *Genome Biol* 5:24
- Benfey, P. N., Ren, L., et al. (1989). The CaMV 35S enhancer contains at least two domains which can confer different developmental and tissue-specific expression patterns. *Embo J* 8:2195-2202
- Benfey, P. N., Ren, L., et al. (1990). Combinatorial and synergistic properties of CaMV 35S enhancer subdomains. *Embo J* 9:1685-96
- Benfey, P. N., Ren, L., et al. (1990). Tissue-specific expression from CaMV 35S enhancer subdomains in early stages of plant development. *Embo J* 9:1677-1684
- Brown, R. M. and Saxena, I. M. (2000). Cellulose biosynthesis: A model for understanding the assembly of biopolymers. *Plant Physiol Biochem.* 38:57-67
- Bruce, W. B., Deng, X. W., et al. (1991). A negatively acting DNA sequence element mediates phytochrome-directed repression of *phyA* gene transcription. *Embo J* 10:3015-24
- Burn, J. E., Hocart, C. H., et al. (2002). Functional analysis of the cellulose synthase genes *CesA1*, *CesA2*, and *CesA3* in *Arabidopsis*. *Plant Physiol.* 129:797-807
- Burn, J. E., Hurley, U. A., et al. (2002). The cellulose-deficient *Arabidopsis* mutant *rsw3* is defective in a gene encoding a putative glucosidase II, an enzyme processing N-glycans during ER quality control. *Plant J.* 32:949-960
- Burton, R. A., Shirley, N. J., et al. (2004). The *CesA* gene family of barley. Quantitative analysis of transcripts reveals two groups of co-expressed genes. *Plant Physiol* 134:224-236
- Choi, M. S., Kim, M. C., et al. (2005). Isolation of a calmodulin-binding transcription factor from rice (*Oryza sativa* L.). *J Biol Chem* 280:40820-40831
- Dai, S., Petruccioli, S., et al. (2003). Functional analysis of RF2a, a rice transcription factor. *J Biol Chem* 278:36396-36402
- Delmer, D. P., Holland, N., et al. (2000). Genes and proteins involved in cellulose synthesis in plants. *Isr J Plant Sci.* 48:165-171
- Demura, T., Tashiro, G., et al. (2002). Visualization by comprehensive microarray analysis of gene expression programs during transdifferentiation of mesophyll cells into xylem cells. *PNAS.* 99:15794-15799
- de Meaux, J., Goebel, U., Pop, A. and Mitchell-Olds, T. (2005) Allele-specific assay reveals functional variation in the chalcone synthase promoter of *Arabidopsis thaliana* that is compatible with neutral evolution. *Plant Cell* 17:676-690
- Emons, A. M. C. and Mulder, B. M. (2000). How the deposition of cellulose microfibrils builds cell wall architecture. *Trends Plant Sci.* 5:35-40

- Eyles, A., Davies, N. W., et al. (2003). Wound wood formation in *Eucalyptus globulus* and *Eucalyptus nitens*: anatomy and chemistry. *Can J Forest Res.* 33:2331-2339
- Fujimori, S., Washio, T., et al. (2003). A novel feature of microsatellites in plants: a distribution gradient along the direction of transcription. *Febs Lett.* 554:17-22
- Geffers, R., Cerff, R., et al. (2000). Anaerobiosis-specific interaction of tobacco nuclear factors with cis-regulatory sequences in the maize GapC4 promoter. *Plant Mol Biol.* 43:11-21
- Goicoechea, M., Lacombe, E., et al. (2005). EgMYB2, a new transcriptional activator from *Eucalyptus* xylem, regulates secondary cell wall formation and lignin biosynthesis. *Plant J.* 43:553-567
- Grec, S., Vanham, D., et al. (2003). Identification of regulatory sequence elements within the transcription promoter region of NpABC1, a gene encoding a plant ABC transporter induced by diterpenes. *Plant J.* 35:237-250
- Hamann, T., Osborne, E., et al. (2004). Global expression analysis of CESA and CSL genes in *Arabidopsis*. *Cellulose.* 11:279-286
- Hamilton, D. A., Schwarz, Y. H., et al. (1998). A monocot pollen-specific promoter contains separable pollen-specific and quantitative elements. *Plant Mol Biol.* 38:663-669
- Hatton, D., Sablowski, R., et al. (1995). Two classes of cis sequences contribute to tissue-specific expression of a *PAL2* promoter in transgenic tobacco. *Plant J.* 7:859-876
- Hernandez, J. M., Heine, G. F., Irani, N. G., Feller, A. et al. (2004) Different mechanisms participate in the R-dependent activity of the R2R3 MYB transcription factor C1. *J Biol Chem.* 27946:48205-48213
- Hertzberg, M., Aspeborg, H., et al. (2001). A transcriptional roadmap to wood formation. *PNAS.* 98:14732-14737
- Huala, E., Dickerman, A. W., et al. (2001). The *Arabidopsis* Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic Acids Res.* 29:102-105
- Joshi, C. P. (2003). Xylem-specific and tension stress-responsive expression of cellulose synthase genes from aspen trees. *Appl Biochem Biotechnol* 105-108:17-25.
- Joshi, C. P., Bhandari, S., et al. (2004). Genomics of cellulose biosynthesis in poplars. *New Phytol.* 164:53-61
- Kankainen, M. and Holm, L. (2005). POCO: discovery of regulatory patterns from promoters of oppositely expressed gene sets. *Nucleic Acids Res.* 33:427-431
- Kao, C. Y., Cocciolone, S. M., et al. (1996). Localization and interaction of the cis-acting elements for abscisic acid, VIVIPAROUS1, and light activation of the C1 gene of maize. *Plant Cell* 8:1171-1179
- Keller, B. and Baumgartner, C. (1991). Vascular-specific expression of the bean GRP 1.8 gene is negatively regulated. *Plant Cell.* 3:1051-1061



- Kieber, J. J. (1997). The ethylene response pathway in Arabidopsis. *Annu Rev Plant Physiol Plant Mol Biol.* 48:277-296
- Kobayashi, T., Nakayama, Y., et al. (2003). Identification of novel cis-acting elements, IDE1 and IDE2, of the barley IDS2 gene promoter conferring iron-deficiency-inducible, root-specific expression in heterogeneous tobacco plants. *Plant J.* 36:780-793
- Kobayashi, T., Suzuki, M., et al. (2005). Expression of iron-acquisition-related genes in iron-deficient rice is co-ordinately induced by partially conserved iron-deficiency-responsive elements. *J Exp Bot.* 56:1305-1316
- Kuhn, E. (2001). From Library Screening to Microarray Technology: Strategies to Determine Gene Expression Profiles and to Identify Differentially Regulated Genes in Plants. *Ann Botany.* 87:139-155
- Lacombe, E., Van Doorselaere, J., et al. (2000). Characterization of cis-elements required for vascular expression of the cinnamoyl CoA reductase gene and for protein-DNA complex formation. *Plant J.* 23:663-676
- Luscher, B. and Eisenman, R. N. (1990). New light on Myc and Myb. Part I. *Myc. Genes Dev.* 4:2025-2035
- Luscher, B. and Eisenman, R. N. (1990). New light on Myc and Myb. Part II. *Myb. Genes Dev.* 4:2235-2241
- Mitsuda, N., Seki, M., et al. (2005). The NAC transcription factors NST1 and NST2 of Arabidopsis regulate secondary wall thickenings and are required for anther dehiscence. *Plant Cell.* 17: 2993-3006
- Nellesen, D. T., Lai, E. C., et al. (1999). Discrete enhancer elements mediate selective responsiveness of enhancer of split complex genes to common transcriptional activators. *Dev Biol* 213:33-53
- Ngai, N., Tsai, F. Y., et al. (1997). Light-induced transcriptional repression of the pea AS1 gene: identification of cis-elements and transactors. *Plant J.* 12:1021-1034
- Persson, S., Wei, H. R., et al. (2005). Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets. *PNAS.* 102:8633-8638
- Quandt, K., Frech, K., et al. (1995). MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucl. Acids Res.* 23:4878-4884
- Ranik, M., Creux, N. M., et al. (2006). Within-tree transcriptome profiling in wood-forming tissues of a fast-growing Eucalyptus tree. *Tree Physiol.* 26:365-375
- Ranik, M. and Myburg, A. A. (2006). Six new cellulose synthase genes from Eucalyptus are associated with primary and secondary cell wall biosynthesis. *Tree Physiol.* 26: 545-556
- Rastogi, R., Bate, N. J., et al. (1997). Footprinting of the spinach nitrite reductase gene promoter reveals the preservation of nitrate regulatory elements between fungi and higher plants. *Plant Mol Biol.* 34:465-476

- Rombauts, S., Florquin, K., et al. (2003). Computational approaches to identify promoters and cis-regulatory elements in plant genomes. *Plant Physiol.* 132:1162-1176
- Samuga, A. and Joshi, C. P. (2002). A new cellulose synthase gene (*PtrCesA2*) from aspen xylem is orthologous to *Arabidopsis AtCesA7* (*irx3*) gene associated with secondary cell wall synthesis. *Gene* 296:37-44
- Scudiero, R., Carginale, V., et al. (2001). Structural and functional analysis of metal regulatory elements in the promoter region of genes encoding metallothionein isoforms in the Antarctic fish *Chionodraco hamatus* (icefish). *Gene* 274:199-208
- Shen, B., Li, C. J., et al. (2002). High free-methionine and decreased lignin content result from a mutation in the *Arabidopsis* S-adenosyl-L-methionine synthetase 3 gene. *Plant J.* 29:371-380
- Solano, R., Stepanova, A., et al. (1998). Nuclear events in ethylene signaling: a transcriptional cascade mediated by ETHYLENE-INSENSITIVE3 and ETHYLENE-RESPONSE-FACTOR1. *Genes Dev.* 12:3703-3714
- Srikanth, C. V., Vats, P., et al. (2005). Multiple cis-regulatory elements and the yeast sulphur regulatory network are required for the regulation of the yeast glutathione transporter, *Hgt1p*. *Curr Genet* 47:345-58
- Sterky, F., Regan, S., et al. (1998). Gene discovery in the wood-forming tissues of poplar: Analysis of 5,692 expressed sequence tags. *PNAS.* 95:13330-13335
- Sun, C., Palmqvist, S., et al. (2003). A novel WRKY transcription factor, *SUSIBA2*, participates in sugar signaling in barley by binding to the sugar-responsive elements of the *isol1* promoter. *Plant Cell* 15:2076-2092
- Terzaghi, W. B. and Cashmore, A. R. (1995). Photomorphogenesis. Seeing the light in plant development. *Curr Biol.* 5:466-468
- Thijs, G., Lescot, M., et al. (2001). A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics* 17:1113-1122
- Tompa, M., Li, N., et al. (2005). Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotechnol.* 23:137-144
- Turner, S. R. and Hall, M. (2000). The Gapped Xylem Mutant Identifies a Common Regulatory Step in Secondary Cell Wall Deposition. *Plant J.* 24:477-488
- Turner, S. R. and Somerville, C. R. (1997). Collapsed xylem phenotype of *Arabidopsis* identifies mutants deficient in cellulose deposition in the secondary cell wall. *Plant Cell* 9:689-701
- Villain, P., Clabault, G., et al. (1994). SIF binding site is related to but different from the light-responsive GT-1 binding site and differentially represses the spinach *rps1* promoter in transgenic tobacco. *J Biol Chem.* 269:16626-16630
- von Gromoff, E. D., Schroda, M., et al. (2006). Identification of a plastid response element that acts as an enhancer within the *Chlamydomonas* HSP70A promoter. *Nucleic Acids Res.* 34:4767-4779

- Wurschum, T., Gross-Hardt, R., et al. (2006). APETALA2 regulates the stem cell niche in the Arabidopsis shoot meristem. *Plant Cell* 18:295-307
- Yin, Y. and Beachy, R. N. (1995). The regulatory regions of the rice tungro bacilliform virus promoter and interacting nuclear factors in rice (*Oryza sativa* L.). *Plant J.* 7:969-980
- Yin, Y., Chen, L., et al. (1997). Promoter elements required for phloem-specific gene expression from the RTBV promoter in rice. *Plant J.* 12:1179-1188
- Yin, Y., Zhu, Q., et al. (1997). RF2a, a bZIP transcriptional activator of the phloem-specific rice tungro bacilliform virus promoter, functions in vascular development. *Embo J.* 16:5247-5259
- Zhou, D. X. (1999). Regulatory mechanism of plant gene transcription by GT-elements and GT-factors. *Trends Plant Sci.* 4:210-214



## **Concluding Remarks**

### **There is some conservation among the *CesA* promoters of *Arabidopsis*, *Populus* and *Eucalyptus***

One of the over all conclusions of this study was that the *Eucalyptus CesA* promoters, and the orthologous regions from *Arabidopsis* and *Populus*, show conservation of the core promoter elements and some cis-regulatory elements. The process of cellulose biosynthesis emerged early in the evolution of plants and has been highly conserved among higher plant species (Nobles et al. 2001). Within the coding sequence of orthologous *CesA* genes from different plant species there are highly conserved regions (Ranik and Myburg 2006) and when clustered by phylogenetic studies the orthologous *CesA* genes form very distinct clades (Figure 2.1). It appears that the conservation in the *CesA* genes is also carried through to the elements within the promoter regions. Sun et al. (2006) showed that among vertebrates such as mammals, chickens and fish even the distance between motifs might be highly conserved in the promoters of orthologous genes. Therefore it is logical that the promoters of genes in a highly conserved pathway such a cellulose biosynthesis will have a number of conserved features among distantly related species.

The conservation of motifs and their positions within promoters can be used as a tool for identifying motifs in orthologous promoters. A study by De Bodt et al. (2006) brought to the fore a program that has been in existence since 2003 but we did not come across till now. De Bodt et al. (2006) compared the orthologous promoter of different plant species to look for conserved motifs. Phylogenetic analysis of promoters has been known for some time but the majority of the software packages are human or mammalian specific (Loots et al. 2002; Solovyev and Shahmuradov 2003). The software package used by de Bodt et al. (2006) is different because it uses a phylogenetic tree that is supplied by the

user and would have been very useful in the comparison of the *Arabidopsis*, *Populus* and *Eucalyptus CesA* gene promoters.

### **Software available for the *in silico* prediction of cis-regulatory elements**

Overlooking the Footprinter software package highlights an important issue that arises with the vast number of motif prediction softwares. There is no database or web interface that lists all the motif prediction software packages available. Also this topic requires a constant stream of reviews such as the one by Tompa et al. (2005) because the different software packages should be tested against each other in order to ensure the development of these packages and to inform readers of the latest most accurate software available. Only through continual use of these software packages will the accuracy of *in silico* motif prediction increase.

*In silico* motif prediction is a relatively new field and although a number of breakthroughs (Tompa et al. 2005; Fiedler and Rehmsmeier 2006; Jensen et al. 2006; Pavesi et al. 2006) have improved the accuracy of motif identification it is still far from accurate. The accuracy of the prediction varies from organism to organism because some organisms such as humans and *Arabidopsis* have been extensively studied. The extensive research on these organisms has led to far more accurate gene and promoter models. The recent completion of the poplar genome sequence will no doubt lead to the production of a number of poplar-specific motif prediction algorithms and the modification of a number already existing algorithms. This will be useful in the study and identification of motifs involved in wood formation.

### **Identification of motifs that may play a role in cellulose biosynthesis**

This study identifies a number of cis-regulatory elements that may play a role in the spatio-temporal expression of the *CesA* genes. The motifs were used in similarity searches in order to give the motifs putative identities. A number of motifs were found to have similarity to well-known cis-regulatory motifs, which are involved in stem specific expression. Some of the most pertinent motifs identified in *CesA* set 2 that should be studied further are CS1, CS7 and CSP1. CS1 showed similarity to an ACII element, which is bound by a MYB transcription factor and plays a major role in the xylem specific expression of lignin (Lauvergeat et al. 2002; Goicoechea et al. 2005). Lignin genes are co-expressed with the *CesA* genes that are associated with secondary cell wall formation and thus this motif may play a role in the tissue specific expression of the *CesA* genes.

The motifs CS7 and CSP1 showed similarity to IDE1 (Iron deficiency response element) and IDE2 respectively. IDE1 and IDE2 are interesting because they work together to respond to iron deficiency and confer a vascular specific expression pattern (Kobayashi et al. 2003; Kobayashi et al. 2005). This is interesting because one of the major enzymes involved in the iron deficiency response pathway is SAMS (S-adenosylmethionine synthase). SAMS is also a major player in the lignin biosynthetic pathway and is co-expressed with the *CesA* genes associated with secondary cell wall formation (Shen et al. 2002). Identifying well-known secondary cell wall motifs, such as above, adds support to the experimental method. Indicating that motifs involved in the tissue-specific expression patterns of the *CesA* genes are indeed being identified.

The *CesA* genes have a very specific spatio-temporal expression pattern and not only are different *CesA* genes expressed during primary and secondary cell wall formation but

*CesA* genes also show a nocturnal expression pattern (Solomon and Myburg unpublished results). In agreement with these findings this study identified a number of motifs (CEP1, CESP1, CS4 and CS6) that showed similarity to motifs involved in the light regulated expression of the genes they control. Motifs CS4 and CEP1 are both elements that have been found to repress gene expression in the presence of light while CEP1 induces gene expression in the absence of light (Appendix 2). Therefore it is possible that these motifs act in unison to produce the relevant expression patterns. Both the *CesA* genes involved in primary and secondary cell wall formation have an elevated nocturnal expression pattern which is in keeping with the identification of these motifs in both *CesA* set 1 and *CesA* set 2. These motifs are important as they play a crucial role in fully understanding cellulose biosynthesis and thus should be further investigated.

A number of motifs identified in this study had poor or no similarity to previously identified cis-regulatory elements. These are of great interest because this is the first documented study of the *CesA* promoters. Motifs specific to cellulose biosynthesis have likely not been identified as yet. The weak similarity could indicate the binding site of a different family member to the transcription factor documented in PLACE. Through deletion studies and other molecular tool, the motifs identified in the Chapter 3 can be tested for functionality. This will aid in the identification of cis-regulatory elements involved in the regulation of the cellulose biosynthesis.

A problem with scanning databases such as PLACE and PlantCARE is that they are updated at particular intervals and thus elements that have recently been identified may not be present in these databases yet. This means that when conducting studies of this nature, the databases should be re-scanned at regular intervals to see if other motifs have been added. Ko et al. (2006) recently compiled a core group of genes thought to play a



role in xylem formation in *Arabidopsis*. The promoters of these genes were scanned for similar motifs and ACAAAGAA was identified in a number of these promoters. The Co-expressed set 2 dataset of this study shared a number of promoters with the dataset scanned by Ko et al. (2006). Re-visiting the over-represented motifs identified in Co-expressed set 2 it was clear that motif ES4 (Table 3.10) was similar to the ACAAAGAA motif. There was no similar motif identified in *CesA* set 2. This suggests that ACAAAGA is *Arabidopsis* specific or is not involved in the regulation of the *CesA* genes. The dataset scanned by Ko et al. (2006) contained three *AtCesA* genes, but only one of these had the ACAAAGAA motif.

The study presented in this dissertation, documents the first the *CesA* promoters to have been studied using *in silico* analyses to identify putative cis-regulatory elements conserved among distantly related plant genera (*Arabidopsis*, *Populus* and *Eucalyptus*). This preliminary study will act as a starting point for a number of other studies pertaining to the regulation of wood formation and cellulose biosynthesis. Using this study as a starting point it will be possible to work up the regulatory network to identify the core regulatory mechanisms of cellulose biosynthesis.

## References

- De Bodt, S., Theissen, G. and Van de Peer, Y. (2006) Promoter analysis of *MADS-Box* genes in eudicots through phylogenetic footprinting. *Mol. Biol. Evol.* 23:1293-1303
- Fiedler, T. and Rehmsmeier, M. (2006) jPREdictor: a versatile tool for the prediction of cis-regulatory elements. *Nucl. Acids Res.* 34(suppl\_2):546-550
- Goicoechea, M., Lacombe, E., Legay, S., Mihaljevic, S., Rech, P., Jauneau, A., Lapierre, C., Pollet, B., Verhaegen, D., Chaubet-Gigot, N. et al. (2005) *EgMYB2*, a new transcriptional activator from *Eucalyptus* xylem, regulates secondary cell wall formation and lignin biosynthesis. *Plant J.* 43:553-567
- Jensen, K. L., Styczynski, M. P., Rigoutsos, I. and Stephanopoulos, G. N. (2006) A generic motif discovery algorithm for sequential data. *Bioinformatics* 22:21-28
- Ko, J. H., Beers, E. P. and Han, K. H. (2006) Global comparative transcriptome analysis identifies gene network regulating secondary xylem development in *Arabidopsis thaliana*. *Mol. Genet. Genomics.* 10:1007-1021
- Kobayashi, T., Nakayama, Y., Itai, R. N., Nakanishi, H., Yoshihara, T., Mori, S. and Nishizawa, N. K. (2003) Identification of novel cis-acting elements, IDE1 and IDE2, of the barley *IDS2* gene promoter conferring iron-deficiency-inducible, root-specific expression in heterogeneous tobacco plants. *Plant J.* 36:780-793
- Kobayashi, T., Suzuki, M., Inoue, H., Itai, R. N., Takahashi, M., Nakanishi, H., Mori, S. and Nishizawa, N. K. (2005) Expression of iron-acquisition-related genes in iron-deficient rice is co-ordinately induced by partially conserved iron-deficiency-responsive elements. *J. Exp. Bot.* 56:1305-1316
- Lauvergeat, V., Rech, P., Jauneau, A., Guez, C., Coutos-Thevenot, P. and Grima-Pettenati, J. (2002) The vascular expression pattern directed by the *Eucalyptus gunnii* cinnamyl alcohol dehydrogenase *EgCAD2* promoter is conserved among woody and herbaceous plant species. *Plant Mol. Biol.* 50:497-509
- Loots, G. G., Ovcharenko, I., Pachter, L., Dubchak, I. and Rubin, E. M. (2002) rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Res.* 12:832-839
- Nobles, D. R., Romanovicz, D. K. and Brown, Jr., R. M. (2001) Cellulose in cyanobacteria: Origin of vascular plant cellulose synthase? *Plant Physiol.* 127:529-542
- Pavesi, G., Mereghetti, P., Zambelli, F., Stefani, M., Mauri, G. and Pesole, G. (2006) MoD Tools: regulatory motif discovery in nucleotide sequences from co-regulated or homologous genes. *Nucl. Acids Res.* 34(suppl\_2):566-570
- Ranik, M. and Myburg, A. A. (2006) Six new cellulose synthase genes from *Eucalyptus* are associated with primary and secondary cell wall biosynthesis. *Tree Physiol.* 26:545-556

- Shen, B., Li, C. J. and Tarczynski, M. C. (2002) High free-methionine and decreased lignin content result from a mutation in the *Arabidopsis* S-adenosyl-L-methionine synthetase 3 gene. *Plant J.* 29:371-380
- Solovyev, V. V. and Shahmuradov, I. A. (2003). PromH: Promoters identification using orthologous genomic sequences. *Nucl. Acids Res.* 31:3540-3545
- Sun, H., Skogerbo, G. and Chen, R. (2006). Conserved distances between vertebrate highly conserved elements. *Hum. Mol. Genet.* 15:2911-2922
- Tompa, M., Li, N., Bailey, T. L., Church, G. M., De Moor, B., Eskin, E., Favorov, A. V., Frith, M. C., Fu, Y. T., Kent, W. J. et al. (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.* 23:137-144



## **Summary**

# Isolation and characterization of the cellulose synthase promoters of *Eucalyptus* trees

*Nicky Creux*

Supervised by *Prof A.A. Myburg and Prof D.K. Berger*

Submitted in partial fulfilment of the requirements for the degree *Magister Scientiae*

*Department of Genetics*

*University of Pretoria*

---

## Summary

Cellulose is one of the most abundant biopolymers on earth and is an important commodity for industries such as the pulp and paper industry. Cellulose is deposited into the plant cell walls by a complex of membrane bound enzymes known as cellulose synthases. A number of cellulose synthase (CesA) genes, which encode for different cellulose synthase proteins, have been identified from plant species such as *Eucalyptus*, *Populus* and *Arabidopsis*. Mutant and expression profile analysis of the CesA genes indicated that a set of three CesA genes are associated with secondary cell wall formation, while a different set of CesA genes are associated with primary cell wall formation. The aim of this study was to investigate the transcriptional regulation of the different members of the CesA gene family in *Eucalyptus*. The promoter regions were comparatively analysed with the orthologous regions in *Arabidopsis* and *Populus* using bioinformatics tools to identify putative regulatory motifs that play a role in CesA genes regulation.

Six *Eucalyptus* CesA gene promoters were isolated using genome walking. The *Eucalyptus* promoter regions and the orthologous promoter regions from *Populus* and *Arabidopsis* were analysed using TSSP (Transcriptional start site plant promoter prediction) and NNPP (Neural network promoter prediction) software packages. The

software packages predicted the transcriptional start sites of the genes and the core regulatory elements such as the TATA-box and initiator elements. The *in silico* results were compared among species and it was found that the predicted transcriptional start sites and the core elements of the Cesa gene promoters showed substantial structural conservation.

The promoter regions were used in a comparative *in silico* analysis with the orthologous promoter regions from *Arabidopsis* and *Populus* to identify putative regulatory motifs. This is the first study in which the promoters of the Cesa gene family are characterized in *Arabidopsis*, *Populus* and *Eucalyptus*. Three software packages (Weeder, POCO and MotifSampler) were used to analyse the promoter regions and identify over-represented motif sequences. A number of key stem-specific and xylem-specific motifs such as the AC-motif and G-box motif were identified as well as a number of novel motifs. Although all of the predicted motifs identified here will have to be functionally tested, the results of this study provide a good map for directed deletion studies and functional testing of the Cesa promoters.

Molecular testing of the predicted motifs may lead to the identification of cis-regulatory elements involved in the differential Cesa gene expression, which will aid in a better understanding the mechanisms underlying gene regulation in the cellulose biosynthetic pathway.



## **APPENDIX 1**

### **Cellulose Synthase Promoter Sequences**

```

SEQ ID NO 1
LOCUS      EgCesA1_Promoter          2000 bp      DNA      linear
DEFINITION Eucalyptus grandis
ACCESSION  EgCesA1_Promoter
VERSION
KEYWORDS   .
SOURCE     Eucalyptus grandis
  ORGANISM Eucalyptus grandis
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; eudicotyledons; core eudicots;
            rosids; Myrtales; Myrtaceae; Eucalyptus.
FEATURES   Location/Qualifiers
  source    1..2000
            /organism="Eucalyptus grandis"
            /mol_type="genomic DNA"
  5'UTR     1788..1997
BASE COUNT 638 a      366 c      368 g      628 t
ORIGIN
1 aatttccatc aattctcttc caattgagtc caaattaag tccacaaaa ttatcccatg
61 atcctaataa gatgtgtgac atgaccgagc ttccgacttc taaattcaat ctccattttt
121 tcggttgaat tcaatctggt gcgatttcag cgcgccta at cactaagtg gcttttaggg
181 aaattttctt gcatcgacc gagggtgatc agtcccaag tctgacgagg taaaattcct
241 taattacctc actcaataga ctagtccctc cgtgagtggc caactctgag agacgaaacta
301 tatgcacaat tgaattgccc gactcaattg actgaaaata aaattgagaa aaatcgggat
361 gtcacagttt acgccttctt tactttcaaa tttacattag attttttcaa atgtttgctt
421 taagcatatt ttaaattttt aatccacttt atatgagatt gaatctacaa tcaatagtaa
481 ttgctttctt cataatagga gtagggcaaa aaaaagtttc tttttttttt tccatcatgt
541 gcttatattt catcagtatt tagatatgat ataatagaa atatgtttga ggttgatcg
601 tgcttctgag aatttaacct attttcatat tttgatcaat tttgatttc cattttttat
661 caagattggc tagataatta aattgaattt taatttcgaa attgaaatat agctttgttc
721 aaaatatttg acaaaaactag gaaattaatt gtgtgtttat aagctgtcat aaatagaaat
781 atcttgtgac tgaaaatata gtgccataaa ttaaaattga caacattatc taggcagagt
841 tatggtaaga tagatgataa aatgatattt tgttgagtgt gatgaagtat ttgttttgtg
901 aaattcatca aaaaaattat gggttaagcct gattataaaa gaaacaagaa tttgaagaga
961 gagagggagg gtaagttcac taatgtttta aaaatcgggt gaaaataggg cctccctaaa
1021 ttagaattga caacatttct taggcaaagt taatgtaagt tacatgaaaa aaaaatttga
1081 tagtttgttg gaagtaatgg agcatttgta ttgtgaaatt cacgatagag ctaacaaaaa
1141 taaaggtagt tgggtgggta acctagttaa aaaagaacaa taatttgaag agagaagaga
1201 gagagagagg agggggagag catttcgata aattcactag aaaaaatggg tgttttagta
1261 taaatgagag tggaaatagg gccatctagg gaacgatcga ttgccctgc acccggccat
1321 ctggagagtc tgatttatac ttctctcgg gctgaaggag ggaaggaaga acaactctt
1381 caatcgggta gtcagacttt gatttcgagg gagggaattg aggttgacaa gaccaaagga
1441 gctaataaac cacctgctga aattctcgag gaagttgaga ggttccagat tagatcttta
1501 ccaaacaaaa aaaaaactat tgcttatgct aaattggcca ttataataag attttttagaa
1561 tactcgttga gtatactcaa ctcaagatat tataagtttt ctcaattggt ttttctccat
1621 ttcttatgat ccgtccacga gcttgaggtc gcttttgaag atgtagccag cccaacagaa
1681 ccgtttcctt catcttcccg cgaaagtttc atgtcatctc cctcctctgc atcacgaacc
1741 aaacctctgc tctctctctc tctctctctc tgcttcaaca caatgacacc aacatcgac
1801 cctcctcacc ttcccaacca ccgccatacc atctcttcta agcattccga tgagtccctg
1861 atccaccgcc ttctcactga gccttcccgc tctcctctt ctcgtctcac tttctcatat
1921 aaagaagtga aagaatacga ggatactcca ctggggtatc gccagaact cattgggtcg
1981 cgagaagatt ggccaacatg
//

```





```

SEQ ID NO 2
LOCUS      CesA2_Promoter          1142 bp    DNA     linear
DEFINITION Eucalyptus grandis
ACCESSION  CesA2_Promoter
VERSION
KEYWORDS   .
SOURCE     Eucalyptus grandis
  ORGANISM Eucalyptus grandis
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; eudicotyledons; core eudicots;
            rosids; Myrtales; Myrtaceae; Eucalyptus.
FEATURES   Location/Qualifiers
  source    1..1142
            /organism="Eucalyptus grandis"
            /mol_type="genomic DNA"
BASE COUNT 281 a    303 c    241 g    317 t
ORIGIN
  1 tggagcatcg agcttcaagg ctttttgggt cttcattttt catcatctct ttgctgcaat
  61 tagatatatg tatatccgct gactaattat tgctactgaa gtggcggctt tttggattag
 121 ctgaactgaa catggcttcg gccctggcgc atcggttttt ctgtctggct actagagaat
 181 ttgacagtct cttataacaa taacatgctg gaacgctgca cttgatgtgt ttcatattgca
 241 ttgaatgta tcatcttttg tctgctcctc ggatatgaat gccagacgaa gaagaagcgc
 301 tgcttgatt ttgccattta ttgaaatata gatgagagat aatcatctta ctcttctttt
 361 gagatggctg aacaataccc caactcgtaa cgggaagtca gcaaaaagga tgcatagtta
 421 gaaaaacaaa gaaaggtaac ccaaatcaaa ggagatgcta gtagttttgc cgggctgccc
 481 gtattttcct gtgtctccgt tctgagaaaa cttaggtggg cgccgcctac gttcatttga
 541 gtttaacttg cacgaaagca tcgctcatct ttcgaacttg ccaaagagtt cggcatgcga
 601 agattcggtc ctatcgattt atgatgctgg ttcagtcatt ggttgcgttt cgctcgacct
 661 gattcgacc  gccccgccc gaaccaatt  cgatgccaca cctgaaccgg ataatcgagc
 721 tacgcgatac attgtaaaga gctgtatag cacgtgttca aggagggtgt gaaagcaaa
 781 tttcccccca taaacaacgc tgtcatgggt tggaaagagg ttagcatcag cgctttcaag
 841 acaaccttgg attcgaagta accggggcgg cagcagcagc agcagcagag ctaacaagca
 901 atctctctct ctctctctcc accaaaacat tgccaaaaaa ttgacgtctc ttcaactcaa
 961 acccagcacc tcccacctca caccgtttgg tagagagatt acaccacttc cacacacaca
1021 cacatactct ctctctacgc ccctctcttc ttatatgatg cagccctagc aagcacctct
1081 ctcgtaccgt tcttctctcc ggcgcctccc cctcgcgatc gtttcccgct cggcccgtgg
1141 cc
//

```

```

SEQ ID NO 3
LOCUS      EgCesA3_Promoter          1312 bp    DNA      linear
DEFINITION Eucalyptus grandis
ACCESSION  EgCesA3_Promoter
VERSION
KEYWORDS   .
SOURCE     Eucalyptus grandis
  ORGANISM Eucalyptus grandis
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; eudicotyledons; core eudicots;
            rosids; Myrtales; Myrtaceae; Eucalyptus.
FEATURES   Location/Qualifiers
  source    1..1312
            /organism="Eucalyptus grandis"
            /mol_type="genomic DNA"
  5'UTR     1265..1309
BASE COUNT 348 a    303 c    293 g    368 t
ORIGIN
1 tctagattaa ggaggacgat tggccaacct gatcgggtaca tatgtaaaat gtctgacgta
61 aggagcattc caggcgtgat atacacactg agtaggagtg gctgcctctg cggagatgga
121 gagctctcaa ctcaactcacc taacgactta aaaaattgtt gacttggatt tcaagctgga
181 gtgcttagat ctcaagcata atatgactca cccagctctc tctctctctc tctctctctt
241 cacaaccagc caaccgctgt tggaaagcaa tgttttctca cagattaatg ctaaaatgtt
301 tgggtgggga tattttctta aatgctgcac ggaaaatggg atctcctttt ggtgctaatt
361 actttcaatc cttccaagat ccgatgtatt ttacgctagt ttgccccgt atttccttat
421 aatcgagcgc ataaagcaaa tccaagtaag atggagcaat aagctcgtgc agtcgactac
481 gacgtcgggc ccgtcctcca cgtaacgctt caattgtgca tgccttatgt gacgttcgga
541 tccatacatg agcccgaatc tcatgtaatt aatacgcagt cctaattcct aagataattg
601 cttctgtttt tactcttggg cggatcgaga atacgggtgtt acatgtccaa cttaatcctc
661 gacacgatta gctgactata cccaaaagga atgattttgg acgtgtaaaa tatcatgctc
721 agtagtgtaa aaacctagtg agcggacatc agcgggtgctg aaatttatac atgtttatac
781 gatggtcgtg gttagcttga tctttcagtt catctgctaa acttacgaag gcatgagaaa
841 cggctaataca aaccggacga cgactatfff aattcgctcg agccttgcca acaggagtcc
901 tcttaaaactg tacaggtaaa aagagtgaaa ctcaatgatc ctaggagtgt catacaagca
961 caatcggacc cgtaggacgc agaagagcct ctgctcggat ttcagcatca aggggctgca
1021 agaccggtcg ttgcagataa gcctgtagag ggatggcaga tgggtggttg agaaaggaaa
1081 tcttggacga agcttagctt caaggcaatg gggactttgc aaccgacgt tttgtaccat
1141 gtgcgtgctg gtgcgtttgt gacattaggt gagatcggct tagaggtttg aaatgctgtg
1201 tcctactttc tctcatttaa aagaaccttt ttccctccat tgcaccacca ccaccttgag
1261 ctaagtcctg ttctagcacc accgccatcc tctcctcctc cctcctccca tg

```

//

```

SEQ ID NO 4
LOCUS      EgCesA4_Promoter          1537 bp    DNA      linear
DEFINITION Eucalyptus grandis
ACCESSION  EgCesA4_Promoter
VERSION
KEYWORDS   .
SOURCE     Eucalyptus grandis
  ORGANISM Eucalyptus grandis
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; eudicotyledons; core eudicots;
            rosids; Myrtales; Myrtaceae; Eucalyptus.
FEATURES   Location/Qualifiers
  source    1..1537
            /organism="Eucalyptus grandis"
            /mol_type="genomic DNA"
  5'UTR     801..1534
  intron    1083..1376
BASE COUNT 400 a    352 c    323 g    462 t
ORIGIN
1 tcaaatatatt tatgtatccg agactaaatt gaatatttac aaaaaaatg taaggataac
61 attgaataaa ttaaaaactt aaaaatgata ttataagtcg gggctaaaat tcagagatca
121 tttttgtcat tcttcaatth tattttccca gaaaatggta gatcatcgcc acaaattgcc
181 tccaaatgac caaaaagaaa ctcccatctc ctccagctct tctctttccc caaccgcga
241 cgatcccgcc gtccccaca catggcgaa acaattaaaa ataccctttt tataaatatc
301 tegtgtcttct gcttcccttc tctgcgattt tttaacaggg aaaaattgag tcaatgatgg
361 ggggtggttg cccgcgtttt ctttgcaag cgtcatttca gtggttagct gatttagacg
421 ggatgcaaaa ggcaaggaga gagcagatgat gattggtagg gatgtcttaa cgtcagatct
481 cgccagttcc tctcgttgcc gtttgatctg tccggtgagg ctccgtcctg gaaataccaa
541 tgcccttcca gggcagctgc tgctctgtcg ctttcgtcgc tccaagcaa tttccttccc
601 cttccccctt ttgcaaagtc acgaaattag gaattcccaa ctacagctgg actccgacac
661 tcggacagtg ggggagtaaa acggaaaaaa aaaaaaaaaat tgaatttttt ttctccttct
721 tttacaaact ctaactaaag agattaaaaat aaagaaacgg aggaattata tacacattgc
781 cgctagtggg tacatcccca tcacatogca tctacaaagc gcgagcgaga gaaccagagg
841 agagacagct agcgtttccc cgcacaccac tctctctctc tctctctctc tctctctgct
901 catcctcttc tctctttcag ctccgggtcag tttcgatctg cattttttca tgctttccct
961 ctgggttcgg ttccggttctg ttggattcga ttccgatggag agttgaagaa agtgctcttc
1021 tttgtgcagg aactgagcgt ttccgctacc gtcctccgtc gttctatccg gtcaagatcg
1081 gatttgaggg tgagtgcact gcttaatttc ccattttctc ccgggggtca tgatctgtga
1141 ttgccgaagt aggacggata atctggaagg tttggatctc gtcactcgag attcgacttt
1201 gttctttgac tggggaacaa gggggaatg agatttattc agttaagctg aagaatcgtt
1261 gcctattatg cagtccttat agaagctagt ccaagtctct aaatttcgct cgcaattgtc
1321 gttcaattgg agcgatcttg ggcttggtga aggagaataa cttatttttc gcaacttttt
1381 caggaagtta ctcacggatc tgtgttttta ctggaaaaca agttgcttct gaatgcaaca
1441 ctagagatct ctacagcttc tgctaagtc acatcaagtt cggaatcagt gaagtcatcc
1501 tctcttagca tccgagccag gaggagctat tgcgatg

```

//

SEQ ID NO 5  
 LOCUS Cesa5\_A\_Promoter 1559 bp DNA linear  
 DEFINITION Eucalyptus grandis  
 ACCESSION Cesa5\_A\_Promoter  
 VERSION  
 KEYWORDS .  
 SOURCE Eucalyptus grandis  
 ORGANISM Eucalyptus grandis  
 Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;  
 Spermatophyta; Magnoliophyta; eudicotyledons; core eudicots;  
 rosids; Myrtales; Myrtaceae; Eucalyptus.  
 FEATURES Location/Qualifiers  
 source 1..1559  
 /organism="Eucalyptus grandis"  
 /mol\_type="genomic DNA"  
 5'UTR 1262..1556  
 BASE COUNT 513 a 296 c 285 g 465 t  
 ORIGIN  
 1 gctggttctg cttgacgaac tgtttgaaca atccaagcct cttgctcagc agctcagccg  
 61 tgactcgtca ttcatgctaa agtatcaaaa tgcgcgtgca ctaaaaaaga tggttaatac  
 121 catacaaaaa atcataaatt tatatacttg tcttgatatt atctcaaatt acttttatat  
 181 cccaaaaatt tcatgcaatt ctgtcatcat ttgctctgaa ttaattttta cataacaaaa  
 241 actcgaaatt gatacaacaa tgtcacattt atatcaaatt gtgcattcat cccgtattta  
 301 tccaagaaat cccaaattca ccataaatta gggaaaatat aaccatcagt ttgatgtaac  
 361 aaatatatta gcttgagata ttttacaaca caaaaattag tttttgtaa gtgtgacaaa  
 421 agtgtactaa tttgaaattt ttttatgatg gaaaaattag tttgcagtaa atgcaatgca  
 481 agtatgtcgg ttggagtttt tatgtcacia aaattaactt aaagtagatg tgggtgtaat  
 541 gttataattt ggggggtttt tttttgtag tattagtctt gacaaaaata gatataata  
 601 gaaataaaat aattcagatt ttgatggaga gagaaggaat ggttcaaatt gcaacattta  
 661 tgagttaaac ttcatctaac aactagctaa tttagatgca tagatgatt tccggtgata  
 721 tgaacttctc aacaacgaag atgaagccga cttgatttcg taatattggt atttcacatt  
 781 tattatgtaa caagtaactc catctagcgt tcataatctc tttttaataa atgcacggta  
 841 tcttgagtta tgttctatgt gatgggcctt tcatttccat aatctaagca aaaaaaatca  
 901 atacaactca tgacttttta gcatatgcat attaaagaaa ttcgcacatt catatattaa  
 961 taattcaaac tttatctcct acttcaogtg ctattatcta tcaactagggt gttatttttt  
 1021 agtatcttca aacatgaaca gaaagaaatc acagatcadc gttgattctg aaagtaactc  
 1081 aagcatatgc atcttctaaa gcaatcgaac tagatgaccg attgacaact ctatttacta  
 1141 atgtaatgct gcaaattgca atccagacag tgcctagcca caggcagccg agccgagccg  
 1201 agccgagcca ccgggtaact cgctcagccg gtgaagcgac tcctcaacct ccccgatta  
 1261 aaaaaagaga ggtaagtgc gggagcgtc aatcatggaa aacgaaggcc caaacgataa  
 1321 aacgacagcg atgggggcg atgtagctgg actgaacaag cttattacag atccagaagc  
 1381 cgagcgacag tgagcgtggt tcagaggcaa gtagcatggc gtgctgagaa aggcgaagaa  
 1441 gaactcggtc cctcctctct ctctcctctc tctcctcctc ctcctccgcc agatcctctc  
 1501 gcttccgcct tcgatctcgg ggagaaggaa ggaaggaaga ggacgacgat ggaggcatg

//



```

SEQ ID NO 6
LOCUS      EgCesA5_B_Promoter          1363 bp    DNA      linear
DEFINITION Eucalyptus grandis
ACCESSION  EgCesA5_B_Promoter
VERSION
KEYWORDS   .
SOURCE     Eucalyptus grandis
  ORGANISM Eucalyptus grandis
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; eudicotyledons; core eudicots;
            rosids; Myrtales; Myrtaceae; Eucalyptus.
FEATURES   Location/Qualifiers
  source    1..1363
            /organism="Eucalyptus grandis"
            /mol_type="genomic DNA"
  5'UTR     1279..1360
BASE COUNT 450 a    254 c    257 g    402 t
ORIGIN
1  tgtcatcatt  tgtctgaa  taattttac  ataacaaaa  ctcgaaattg  atacaacaat
61  gtcacattta  tatcaaattg  tgcattcatc  ccgtatttat  ccaagaaatc  ccaaattcac
121  cataaattag  ggaaaatata  accatcagtt  tgatgtaaca  aatataattag  cttgagatat
181  tttacaacac  aaaaattagt  ttttgtaag  tgtgacaaaa  gtgtactaat  ttgaaatfff
241  tttatgatgg  aaaaattagt  ttgcagtaaa  tgcaatgcaa  gtatgtcggg  tggagttfff
301  atgtcacaaa  aattaactta  aagtagatgt  ggtgtgaatg  ttataatttg  gggggttfff
361  tttttgtagt  attagtcttg  acaaaaatag  atatatacag  aaataaaata  attcagattt
421  tgatggagag  agaaggaatg  gttcaaattg  caacatttat  gagtttaact  tcacttaaca
481  actagctaat  ttgatgcat  agatgatttt  ccggtgatat  gaacttctca  acaacgaaga
541  tgaagccgac  ttgatttcgt  aatattgtta  tttcacattt  attatgtaac  aagtaactcc
601  atctagcgtt  cataatctct  ttttaataaa  tgcacgggat  cttgagttat  gttctatgtg
661  atgggccttt  ctttccata  atctaagcaa  aaaaaatcaa  tacaactcat  gactttttag
721  catatgcata  ttaaagaaat  tcgcacattc  atatattaat  aattcaaact  ttatctccta
781  cttcacgtgc  tattatctat  cactaggggtg  ttatttttta  gtatcttcaa  acatgaacag
841  aaagaaatca  cagatcatcg  ttgattctga  aagtaactca  agcatatgca  tcttctaaag
901  caatcgaaact  agatgaccga  ttgacaactc  tatttactaa  tgtaatgctg  caaattgcaa
961  tccagacagt  gcctagccac  aggcagccga  gccgagccga  gccgagccac  cgggtaactc
1021  gctcagccgg  tgaagcgact  cctcaacctc  cccgaattaa  aaaaagagag  gtaagtgacg
1081  ggagcgctca  atcatggaaa  acgaaggccc  aaacgataaa  acgacagcga  tgggggcgga
1141  tgtagctgga  ctgaacaagc  ttattacaga  tccagaagcc  gagcgacagt  gagcgtgfff
1201  cagaggcaag  tagcatggcg  tgctgagaaa  gccgaagaag  aactcgggtcc  ctcctctctc
1261  tctcctctct  ctctctctcc  tctctcogcca  gatcctctcg  cttccgcctt  cgatctcggg
1321  gagaaggaag  gaaggaagag  gacgacgatg  gaggcaatct  atg

```

//



```

SEQ ID NO 7
LOCUS      Cesa7_Promoter          781 bp    DNA      linear
DEFINITION Eucalyptus grandis
ACCESSION  Cesa7_Promoter
VERSION
KEYWORDS   .
SOURCE     Eucalyptus grandis
  ORGANISM Eucalyptus grandis
            Eukaryota; Viridiplantae; Streptophyta; Embryophyta; Tracheophyta;
            Spermatophyta; Magnoliophyta; eudicotyledons; core eudicots;
            rosids; Myrtales; Myrtaceae; Eucalyptus.
FEATURES   Location/Qualifiers
  source    1..781
            /organism="Eucalyptus grandis"
            /mol_type="genomic DNA"
  5'UTR     526..779
BASE COUNT 132 a    271 c    153 g    225 t
ORIGIN
  1 aaaggaaaga cgcgacagcc agaaccctaaa aacccccggcc cggaatcgcc attaaagcgc
  61 ccttaaatta attcccctac tcctcctccc aaaaaataaa actgaaatcc cccctctctc
  121 tctctctact ttctctcact accaactcga ccttcctcct cctgtttttt ttttttggg
  181 tttttgtcgt ggcgggtggg ggcacccagc ccagcgccac caccacattg tgggtccccc
  241 ctccctctct ctctatgact gacttttctc attctctctc tctatctctt ctggagtctc
  301 gattagattg gattcactcg tacaacaagc accagcacat ctctgcatct gctcctccct
  361 cctctcactc ctcccactt cccttcctt ccttcctctg tcaogcctct cccctctctc
  421 tctctctaga cgctcgcgaa tacgcaggcg agaccattt cctcccttcc tttctctctc
  481 tgtgaatcta cccgtctaaa aaaggctgtc cgcagcacat tgatcgagat cgagagcgca
  541 gcagagcatc cccgctcga caagcattct cccccgccag atcggccgct gcattcctcg
  601 tcgtagaggg ggagggcagc tttcttggtg ggtggctccg ggccggcaatg cggagatccg
  661 ggtctgttct gaagagctga gactgctgct gggtttctct tctttcttct ctttcttggtg
  721 ccggttcgctt ccttgcgctt ttgtcggtgg tgggtgagtc gggtcctctc gttctggtat
  781 g
  //

```



## **APPENDIX 2**

# **Primary and secondary associated Cesa motif datasheets**

## Appendix 2: Table of Contents

<b>APPENDIX 2</b> .....	<b>182</b>
PRIMARY AND SECONDARY ASSOCIATED CESA MOTIF DATASHEETS .....	182
APPENDIX 2 LEGEND .....	184
<b>MOTIFS IDENTIFIED IN MORE THAN ONE DATASET</b> .....	<b>186</b>
<i>APPENDIX 2.1: CESP1</i> .....	186
<i>APPENDIX 2.3: CESP2</i> .....	187
<i>APPENDIX 2.3: CEP1</i> .....	188
<i>APPENDIX 2.4: CEP2</i> .....	189
<i>APPENDIX 2.5: CEP3</i> .....	190
<i>APPENDIX 2.6: CSP1</i> .....	191
<i>APPENDIX 2.7: CSP2</i> .....	192
<b>MOTIFS IDENTIFIED IN CESA SET 1</b> .....	<b>193</b>
<i>APPENDIX 2.8: CP1</i> .....	193
<i>APPENDIX 2.9: CP2</i> .....	194
<i>APPENDIX 2.10: CP3</i> .....	195
<i>APPENDIX 2.11: CP4</i> .....	196
<i>APPENDIX 2.12: CP5</i> .....	197
<i>APPENDIX 2.13: CP6</i> .....	198
<i>APPENDIX 2.14: CP7</i> .....	199
<i>APPENDIX 2.15: CP8</i> .....	200
<i>APPENDIX 2.16: CP9</i> .....	201
<i>APPENDIX 2.17: CP10</i> .....	202
<i>APPENDIX 2.18: CP11</i> .....	203
<i>APPENDIX 2.19: CP12</i> .....	204
<i>APPENDIX 2.20: CP13</i> .....	205
<i>APPENDIX 2.21: CP14</i> .....	206
<i>APPENDIX 2.22: CP15</i> .....	207
<i>APPENDIX 2.23: CP16</i> .....	208
<i>APPENDIX 2.24: CP17</i> .....	209
<b>MOTIFS IDENTIFIED IN CESA SET 2</b> .....	<b>210</b>
<i>APPENDIX 2.25: CS1</i> .....	210
<i>APPENDIX 2.26: CS2</i> .....	211
<i>APPENDIX 2.27: CS3</i> .....	212
<i>APPENDIX 2.28: CS4</i> .....	213
<i>APPENDIX 2.29: CS5</i> .....	214
<i>APPENDIX 2.30: CS6</i> .....	215
<i>APPENDIX 2.31: CS7</i> .....	216
<i>APPENDIX 2.32: CS8</i> .....	217



## Appendix 2 Legend

Appendix 2 presents a data sheet for each of the motifs identified in Cesa set 1 and Cesa set 2. The motif identity assigned to each motif indicates which dataset it was identified in (C = Cesa set, E = Co-Expressed set, P = promoters of genes associated with primary cell wall formation: set 1, S = promoters of genes associated with secondary cell wall formation: set 2). Each data sheet contains the motif consensus sequences, the reverse complement consensus sequence, the datasets the motif was over-represented in, the software packages it was identified by, the statistical significance and the motif logos provided by the different programs. The motif alignment from which the motif consensus was constructed is provided. The base most often present, in a specific position, was used in the consensus sequence. The motifs in the alignment are labelled A-Z, which refers back to the motifs listed in Tables 3.3-3.6. The consensus motifs were used in similarity searches on the PLACE and PlantCARE databases. The element, which is most similar to the motif, is presented here and the level of similarity is given (E-value and Z-score).


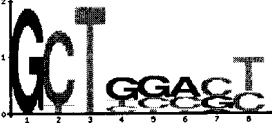
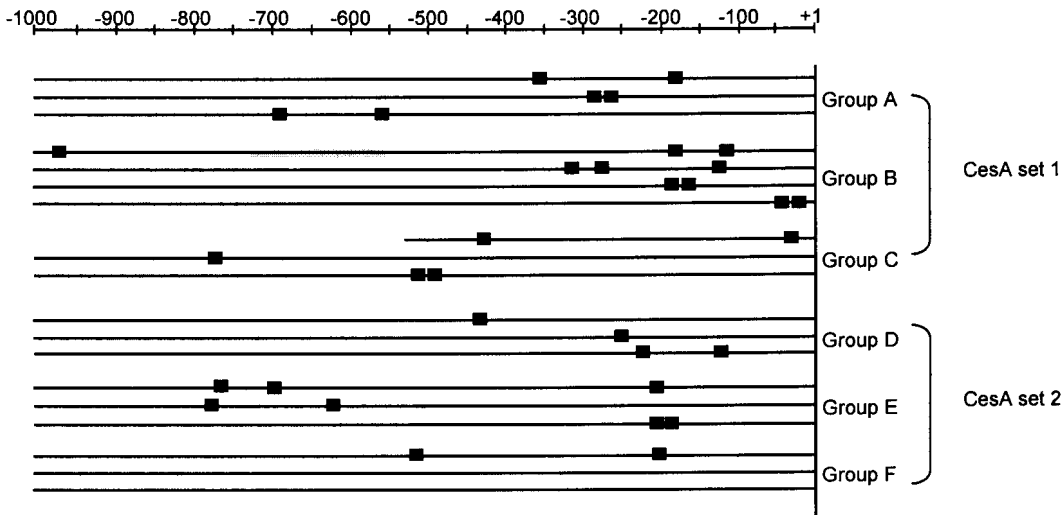
The schematic at the bottom of the table maps the motifs, from which the consensus sequence was constructed to the promoters in Cesa set 1 and Cesa set 2. In the schematic the orthologous promoters are grouped together: Group A contains the promoters of *EgCesA4*, *PtrCesA5* and *AtCesA3*; Group B contains *EgCesA5A*, *EgCesA5B*, *PtrCesA4* and *AtCesA1*; Group C contains *EgCesA7*, *PtrCesA7* and *AtCesA2*; Group D contains *EgCesA1*, *PtrCesA1* and *AtCesA8*; Group E contains *EgCesA3*, *PtrCesA2* and *AtCesA4*; Group F contains *EgCesA2*, *PtrCesA3* and *AtCesA7* (Table 3.8). The order of the promoters in each group are as stated here, where *Eucalyptus* is always listed first, *Populus* second and the *Arabidopsis* promoters last. The promoters in each schematic are anchored at the predicted transcriptional start site and the top bar provides an

approximation of the motif position. The grey block in Group B indicates the region of *EgCesA5A* promoter that is deleted in the *EgCesA5B* promoter (Figure 2.5, Group B).



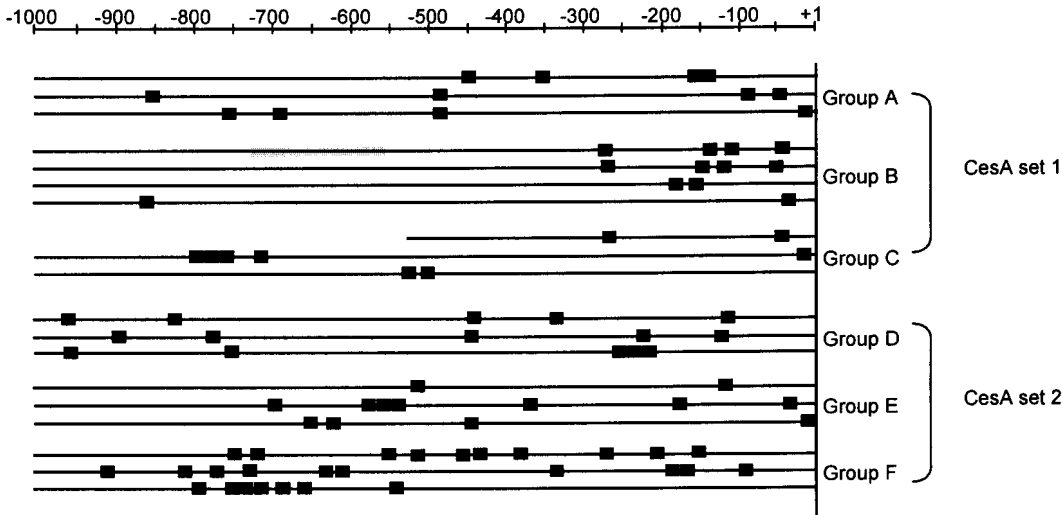
In some cases the reverse motif has been included as a separate motif (E.g. CP8/CP9 and CP13/CP14); because when these motifs were used in the similarity searches they were found to be highly similar to a couple of different elements on the PLACE database. This is due to the fact that the motifs isolated in this study are relatively short and the elements contained on the PLACE database range from 6 bp to approximately 30 bp. A motif may therefore be highly similar to more than one element and without molecular studies to confirm the motifs function it will be difficult to assign a single function or identity to the motifs.

## Motifs identified in more than one dataset



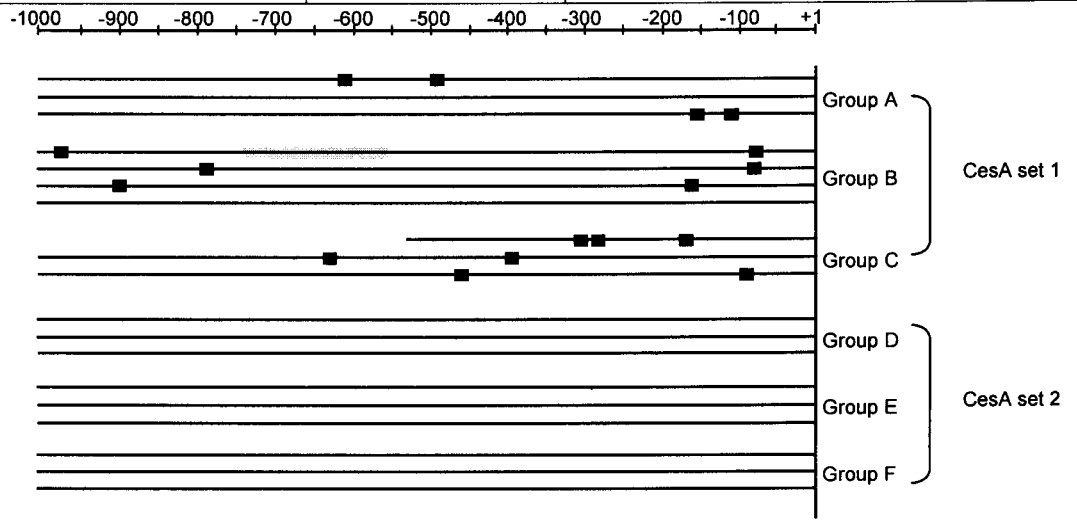
### Appendix 2.1: CESP1

<b>Motif Identity:</b>	CESP1	
<b>Consensus sequence:</b>	G(AT)CGGTG(A/G)AGCTGTTG(G/T)	
<b>Reverse compliment:</b>	(A/C)CAACAGCT(C/T)CACCG(A/T)C	
<b>Data sets:</b>	Identified in Cesa set 1, Cesa set 2, Co-expressed set 1 and Co-expressed set 2	
<b>Consensus alignment:</b> A- GCTNNMSY B- GCNNCACC C- GCTCCACC D- YCACCGWC E- YSACCGWC F- YCACCGNC H- NCACYSWC I- TNNCGNC J- ACMGCT K- ACAGCT L- ACAGCT M- AYAGCT N- ACAGCT O- CCATCAGC P- MCATCAGC <u>MACAACAGCTYCACCGWC</u>	<b>POCO:</b>	<b>Number of motif occurrences:</b> 21 <b>Number of promoters with motif:</b> 10/11 <b>Motif representation in the data sets:</b> This motif is represented 3 times more in the Cesa set 1 than in Cesa set 2 and is 6 times more abundant in Cesa set 1 as compared to the background <b>Motif P-value:</b> 0.0038
	<b>Weeder:</b>	<b>Weeder value:</b> 0.35 <b>Motif signature:</b> 
	<b>MotifSampler:</b>	<b>Log-likelihood:</b> 88.52 <b>Information content:</b> 1.4 <b>Consensus score:</b> 1.1 <b>Motif signature:</b> 
	<b>PLACE Hit:</b>	<b>Z-score:</b> 78.3 <b>E-value:</b> 0.54 <b>Place ID:</b> PE3ASPHYA3 <b>Place Motif:</b> <b>CAGCTCCCATGGCTCTCCCATCCGCGCCGGT</b> <b>Putative function:</b> Cis-acting elements involved in the photo-regulation of phytochromes
		

*Appendix 2.3: CESP2*



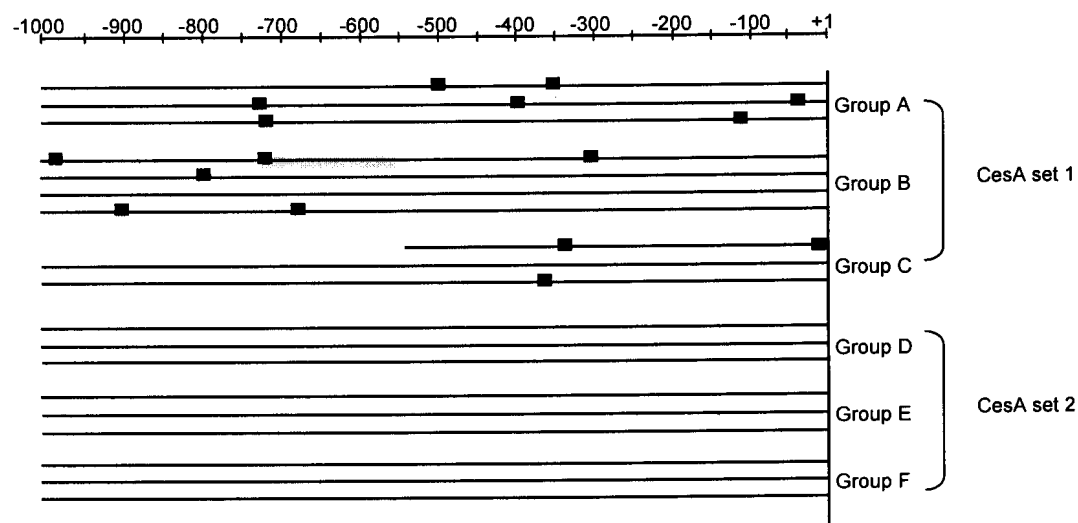
<u>Motif Identity:</u>	CESP2	
<u>Consensus sequence:</u>	GA(C/G)GGCAGG	
<u>Reverse compliment:</u>	CCTGCC(C/G)TC	
<u>Data sets:</u>	Identified in CesA set 1, CesA set 2 and Co-expressed set 1	
<u>Consensus alignment:</u>  A- AGNSSAGN B- MGCMSAGY C- AGCCGAGC D- GGGCTG E- GGCAGG F- GGCAGR G- GGCWGG H- GGCAGG I- SSGCWG J- SNGGWWG K- GGGCTG L- GGCWGY M- GGCAYG N- YGCAGY O- GGCAGG P- GGGCWS Q-GTCGKY R-GTCGGC _____ _GASGGCAGG	<u>POCO:</u>	<u>Number of motif occurrences:</u> Co-expressed set 1=24, CesA set 1 = 33 and CesA set 2=62  <u>Number of promoters with motif:</u> Co-expressed set 1= 14/20 promoters, CesA set 1= 11/11 promoters, CesA set 2= 8/8 promoters  <u>Motif representation in the data sets:</u> 2 fold higher representation in CesA set 2 than CesA set 1.
	<u>Weeder:</u>	<u>Weeder value:</u> 0.72  <u>Motif signature:</u> 
	<u>MotifSampler:</u>	<u>Log-likelihood:</u> 77.85 <u>Information content:</u> 1.5 <u>Consensus score:</u> 1.2 <u>Motif signature:</u> 
	<u>PLACE Hit:</u>	<u>Z-score:</u> 69.3 <u>E-value:</u> 0.64 <u>Place ID:</u> PSREGIONZM13 <u>Place Motif:</u> <b>TCGGCCACTATTTCTACGGGCAGCCAGACAAA</b> <u>Putative function:</u> Pollen specific region in maize contains separable pollen-specific and quantitative elements
		

*Appendix 2.3: CEP1*

<b>Motif Identity:</b>	CEP1	
<b>Consensus sequence:</b>	NTGTCGGTG	
<b>Reverse compliment:</b>	CACCGACAN	
<b>Data sets:</b>	Identified CesA set 1 and Co-expressed set 1	
<b>Consensus alignment:</b>  A- TGTSNSTN B- TGNCNGTG C- TGTCGGTG D- NTGTCGKT E- KTGTGCKT F- TGNNGGTC G-TNGTNNNT H-CTGTGCGGT I- GTGTGCGGT J- TGTCGGTG K- TGACGGTG — NTGTCGGTG	<b>POCO:</b>	<b>Number of motif occurrences:</b> 6 <b>Number of promoters with motif:</b> 8/11 <b>Motif representation in the data sets:</b> This motif is six times more represented in the CesA set1 than in CesA set 2 and the background <b>Motif P-value:</b> 0.00242
	<b>Weeder:</b>	<b>Weeder value:</b> 0.46 <b>Motif signature:</b> 
	<b>MotifSampler:</b>	<b>Log-likelihood:</b> 63.43 <b>Information content:</b> 1.4 <b>Consensus score:</b> 1.2 <b>Motif signature:</b> 
	<b>PLACE Hit:</b>	<b>Z-score:</b> 80.5 <b>E-value:</b> 0.63 <b>Place ID:</b> BOXBPSAS1 <b>Place Motif:</b> AAACGACACCGTTT <b>Putative function:</b> Is known as the box-B and has been identified in the Pea Asparagine synthase gene and is involved in light-induced transcriptional repression of the genes.
		



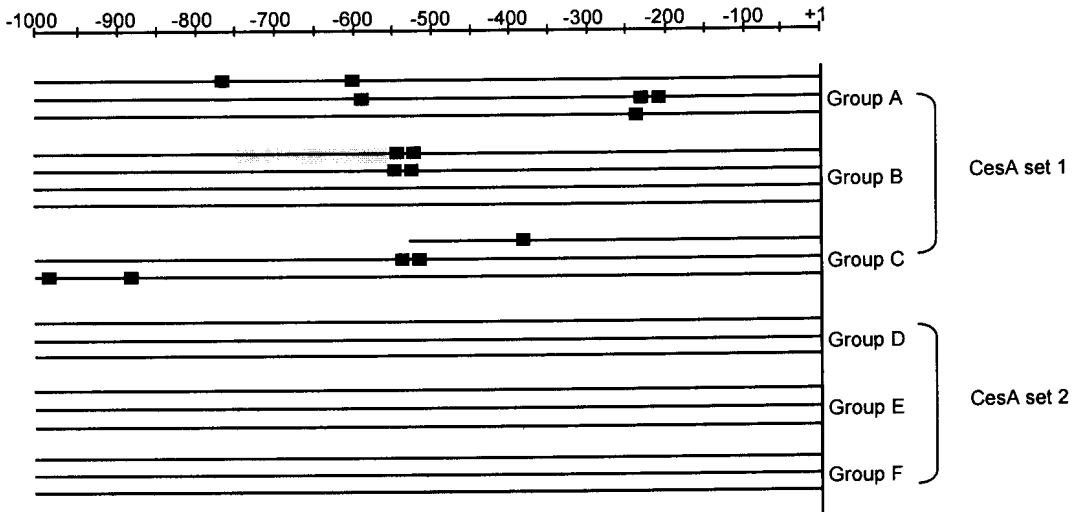


*Appendix 2.4: CEP2*


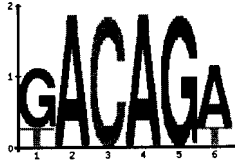
<b>Motif Identity:</b>	CEP2	
<b>Consensus sequence:</b>	(A/C)TGTCGG	
<b>Reverse compliment:</b>	CCGACA(G/T)	
<b>Data sets:</b>	Identified in Cesa set 1 and Co-expressed set 1	
<b>Consensus alignment:</b> A- TGCKG B- TGTGRG C- TGTCGG D- TGTCGG E- MTGTCG F- TGTCGK H- GTGTCG MTGTCGG	<b>POCO:</b>	No POCO Prediction
	<b>Weeder:</b>	<b>Weeder value:</b> 1.12 <b>Motif signature:</b> 
	<b>MotifSampler:</b>	<b>Log-likelihood:</b> 55.49 <b>Information content:</b> 1.8 <b>Consensus score:</b> 1.6 <b>Motif signature:</b> 
	<b>PLACE Hit:</b>	<b>Z-score:</b> 73.4 <b>E-value:</b> 0.93 <b>Place ID:</b> CGTGTSPHZMC1 <b>Place Motif:</b> CGTGTCTGTCATGCAT <b>Putative function:</b> Required for abscisic acid responsiveness
		



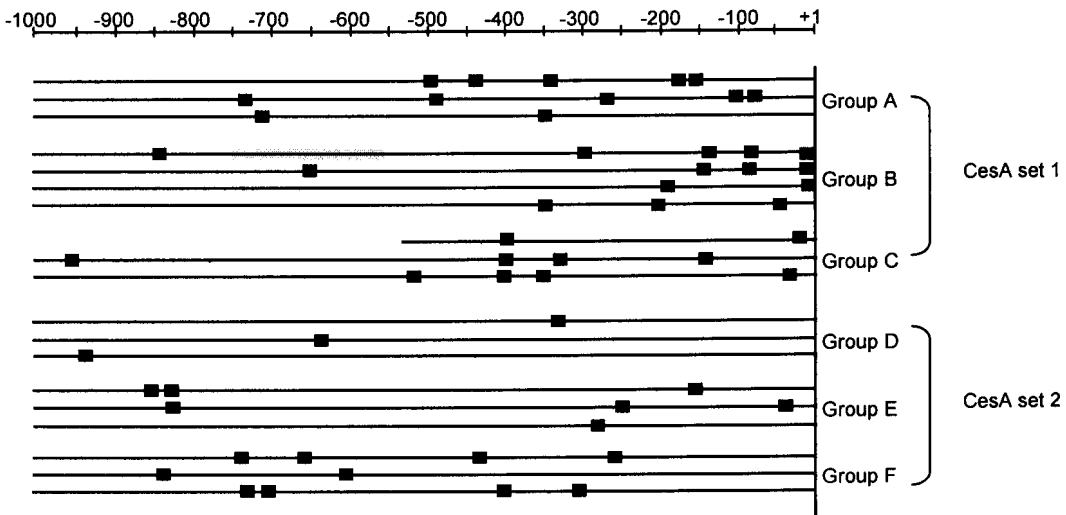
*Appendix 2.5: CEP3*

<b>Motif Identity:</b>	CEP3	
<b>Consensus sequence:</b>	GNCA(C/G)TGA	
<b>Reverse compliment:</b>	TCA(C/G)TGNC	
<b>Data sets:</b>	Identified in CesA set 1, Co-expressed set 1 and Co-expressed set 2	
<b>Consensus alignment:</b>  A- GNCASTGN B- GACAGTGG C- CACTGA D- CANTNA GNCAS <del>T</del> GA	<b>POCO:</b>	No POCO Prediction
	<b>Weeder:</b>	Weeder value: 0.45 Motif signature: 
	<b>MotifSampler:</b>	Log-likelihood: 73.31 Information content: 1.5 Consensus score: 1.26 Motif signature: 
	<b>PLACE Hit:</b>	Z-score: 85.9 E-value: 0.24 Place ID: AS1CAMV Place Motif: CCACTGACGTAAGGGATGACGCACAATCC Putative function: Is similar to an activation sequence in the CaMV 35s promoter and directs expression in Roots and leaves.
		

*Appendix 2.6: CSP1*


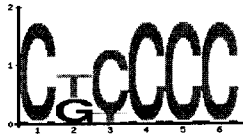
<u>Motif Identity:</u>	CSP1	
<u>Consensus sequence:</u>	GACAGAA(G/T)N	
<u>Reverse compliment:</u>	N(A/C)TTCTGTC	
<u>Data sets:</u>	Identified in CesA set 1 and CesA set 2	
<u>Consensus alignment:</u>  A- KACAGA B- GACAGA C- GACAGA D- GASAGA E- SACAGA F- AGAGMTKS G- WMRGCAKG H- WCAGMAKN I- ANGNCATG J- ACAGCAGG K- ACAGCAGC _____ GACAGAAKN	<u>POCO:</u>	No POCO Prediction
	<u>Weeder:</u>	Weeder value: 0.96 Motif signature: 
	<u>MotifSampler:</u>	Log-likelihood: 68.98 Information content: 1.7 Consensus score: 1.7 Motif signature: 
	<u>PLACE Hit:</u>	Z-score: 80.7 E-value: 0.25 Place ID: IDE2HVIDS2 Place Motif: TTGAACGGCAAGTTTCACGCTGTCCT Putative function: One of two elements identified in Barley that confer iron-deficiency-inducible root specific expression.

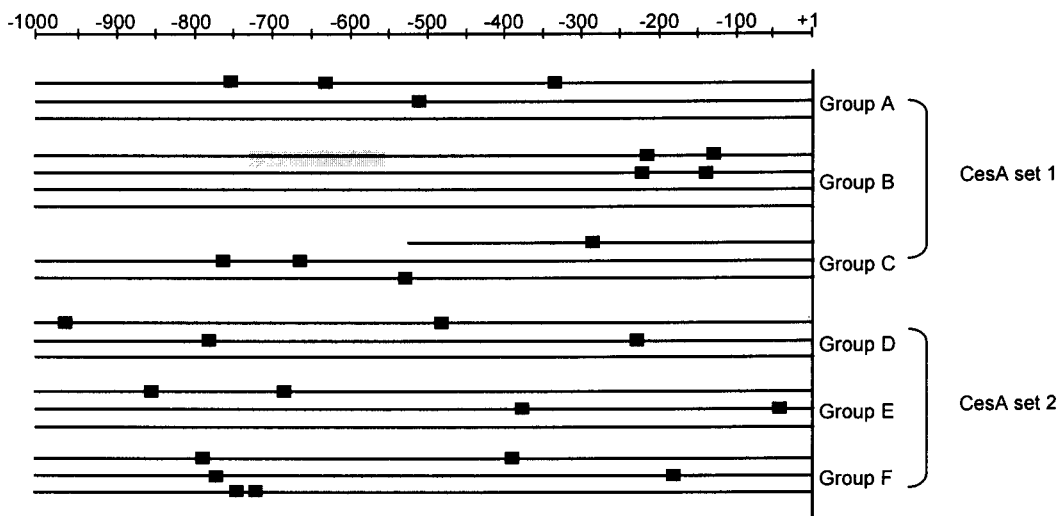
  







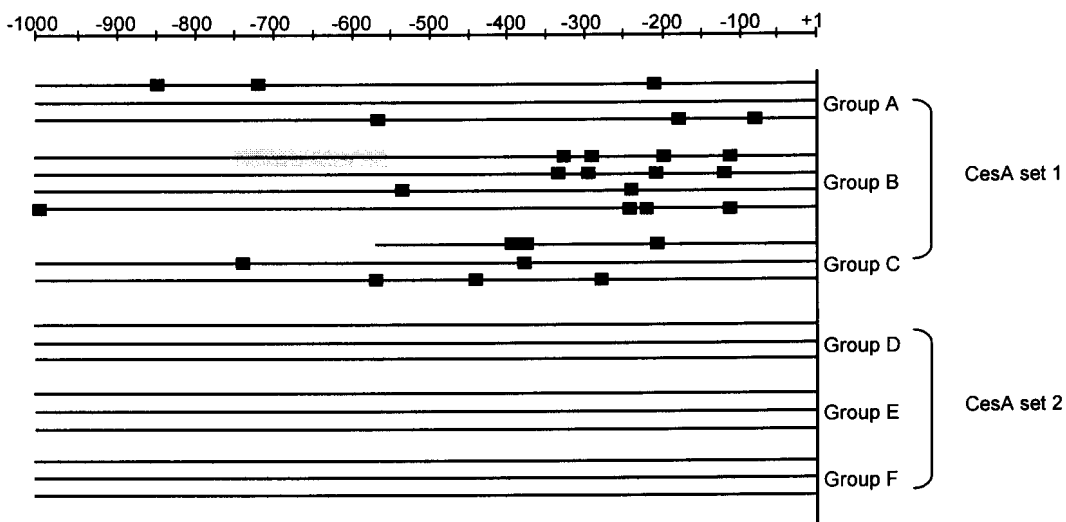
*Appendix 2.7: CSP2*

<u>Motif Identity:</u>	CSP2	
<u>Consensus sequence:</u>	GGGGC(A/G)NGNN	
<u>Reverse compliment:</u>	NNCN(C/T)GCCCC	
<u>Data sets:</u>	Identified in Cesa set 1 and Cesa set 2	
<u>Consensus alignment:</u> A-GGGGCG B-GGGGKG C- GGCANGKN D- GNNANGNT GGGCRNGNN	<u>POCO:</u>	<u>Number of motif occurrences:</u> 12 <u>Number of promoters with motif:</u> 6/8 Promoters <u>Motif representation in the data sets:</u> Represented 2 fold higher in the Cesa set 2 than in Cesa set 1 <u>Motif P-value:</u> 0.0009
	<u>Weeder:</u>	<u>Weeder value:</u> 1.40 <u>Motif signature:</u> 
	<u>MotifSampler:</u>	<u>Log-likelihood:</u> 45.86 <u>Information content:</u> 2.1 <u>Consensus score:</u> 1.5 <u>Motif signature:</u> 
	<u>PLACE Hit:</u>	<u>Z-score:</u> 63 <u>E-value:</u> 1.7 <u>Place ID:</u> EIN3ATERF1 <u>Place Motif:</u> GGATTCAAGGGGCATGTATCTTGAATCC <u>Putative function:</u> Ethylene-insensitive binding site is necessary and sufficient for Ethylene response factor 1 expression which activates a variety of ethylene response genes



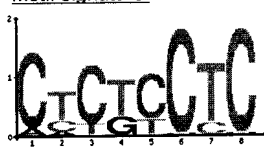
## Motifs identified in Cesa set 1

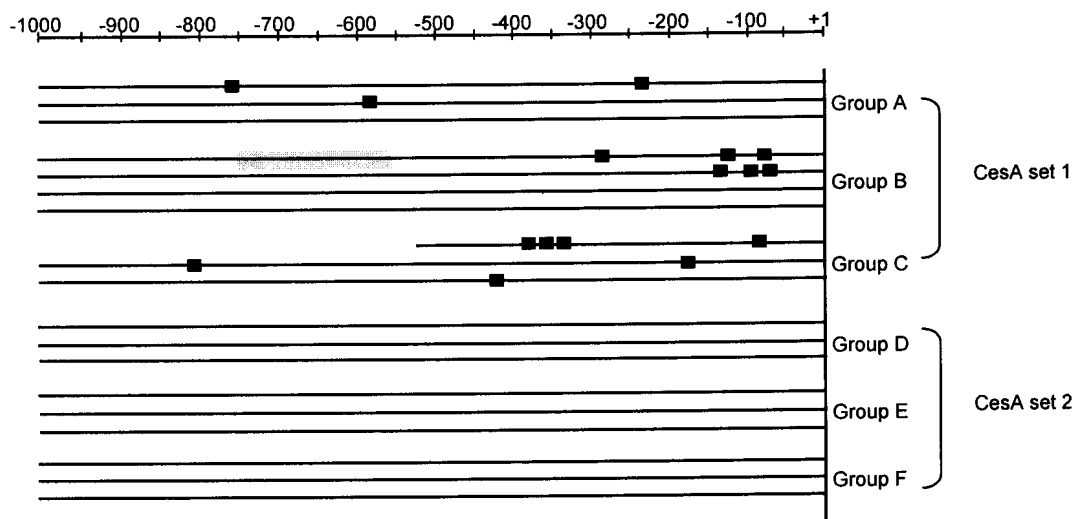
### Appendix 2.8: CPI

<b>Motif Identity:</b>	CP1	
<b>Consensus sequence:</b>	GGNGGTGG	
<b>Reverse compliment:</b>	CCACCNCC	
<b>Data sets:</b>	Only identified in Cesa set 1	
<b>Consensus alignment:</b>  A- GSKGGYKS B- GGRKGYKG C- GGGGGTGG D- GGGTGG E- SGSTGS F- NGCTGG G- SGGTGG H- GGNTGG I- GGGTGG GGNGGTGG	<b>POCO:</b>	No POCO Prediction
	<b>Weeder:</b>	Weeder value: 1.08 Motif signature: 
	<b>MotifSampler:</b>	Log-likelihood: 71.75 Information content: 1.7 Consensus score: 1.37 Motif signature: 
	<b>PLACE Hit:</b>	Z-score: 123.7 E-value: 0.0018 Place ID: ARELIKEGHPGDFR2 Place Motif: AGTTGAATGGGGGTGCA Putative function: Anthocyanin regulatory element found in the maize anthocyanin promoter. It is a R2R3 MYB binding site
		





Appendix 2.9: CP2

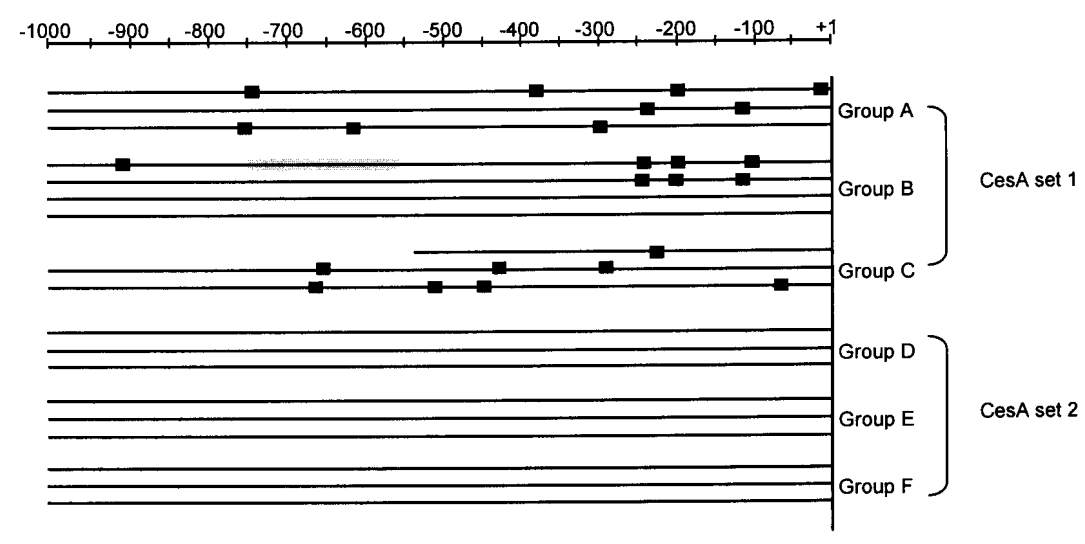
<b>Motif Identity:</b>	CP2	
<b>Consensus sequence:</b>	CNCNNCNC	
<b>Reverse compliment:</b>	GNGNNGNG	
<b>Data sets:</b>	Only identified in Cesa set 1	
<b>Consensus alignment:</b> A- CNCNCCTC B- CNCNNCNC C- CNCTNCNC CNCNNCNC	<b>POCO:</b>	<b>Number of motif occurrences:</b> 22 <b>Number of promoters with motif:</b> 6/11 <b>Motif representation in the data sets:</b> This motif is over represented 2.4 fold higher in Cesa set 1 when compared to the background <b>Motif P-value:</b> 0.0272
	<b>Weeder:</b>	No Weeder Prediction
	<b>MotifSampler:</b>	<b>Log-likelihood:</b> 50.37 <b>Information content:</b> 1.5 <b>Consensus score:</b> 1.3 <b>Motif signature:</b> 
<b>PLACE Hit:</b>	<b>Z-score:</b> 113.2 <b>E-value:</b> 0.00017 <b>Place ID:</b> ABRECE3ZMRAB28 <b>Place Motif:</b> ACGCGCCTCCTC <b>Putative function:</b> Abscisic Acid response element	




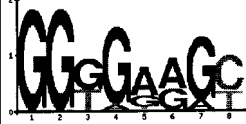
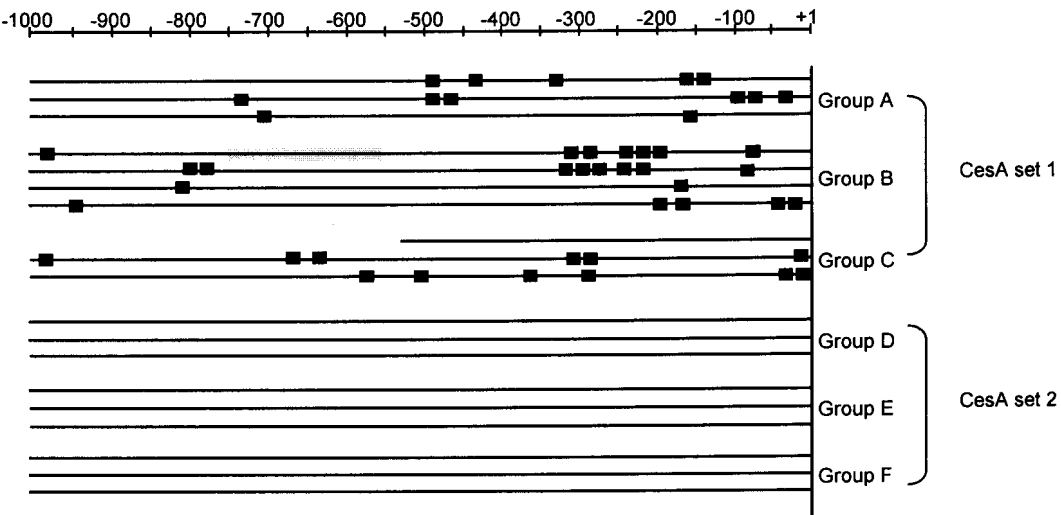
*Appendix 2.10: CP3*

<b>Motif Identity:</b>	CP3	
<b>Consensus sequence:</b>	CCNC(A/C)CCC	
<b>Reverse compliment:</b>	GGG(G/T)GNGG	
<b>Data sets:</b>	Only identified in Cesa set 1	
<b>Consensus alignment:</b> A- CCNNMCCC B- CCACCCCC C- CCCMCY D- CCCCCC CCNCMCCC	<b>POCO:</b>	No POCO Prediction
	<b>Weeder:</b>	Weeder value: 0.7 Motif signature: 
	<b>MotifSampler:</b>	Log-likelihood: 62.01 Information content: 1.5 Consensus score: 1.1 Motif signature: 
	<b>PLACE Hit:</b>	Z-score: 102.6 E-value: 0.039 Place ID: ACIIPVPAL2 ACII Place Motif: CCACCAACCCCC Putative function: Is an ACII-element, which is required for vascular specific gene expression and is a possible MYB binding site. This may interact with the G-Box to direct the complex patterns of tissue-specific expression of <i>PAL2</i> gene.

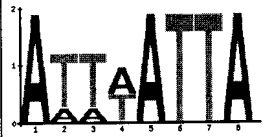
  

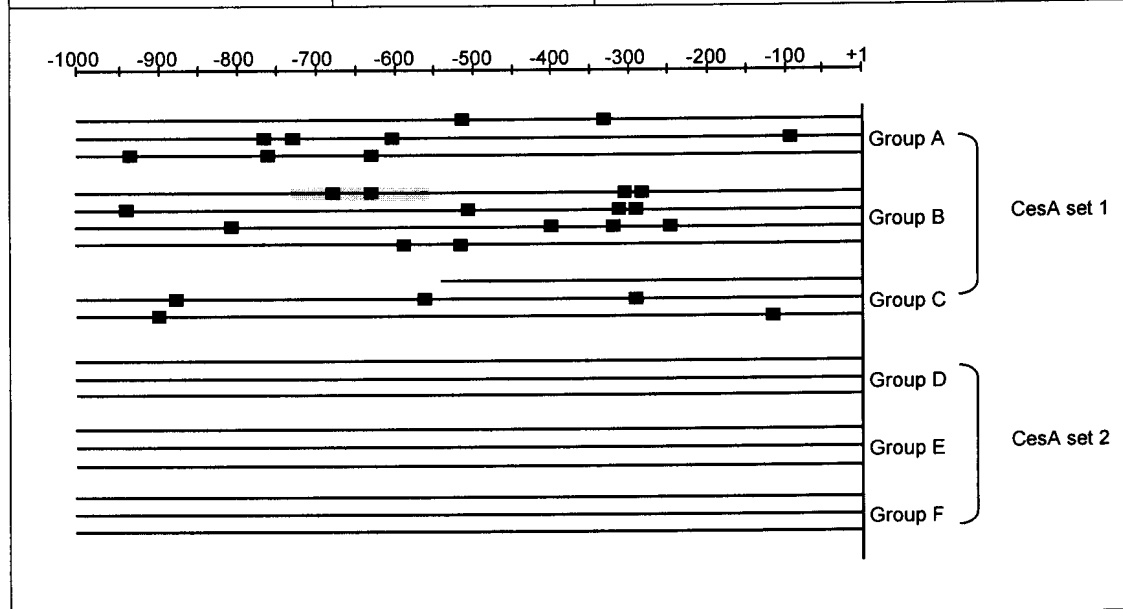


Appendix 2.11: CP4

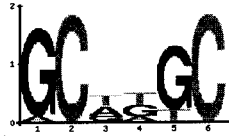
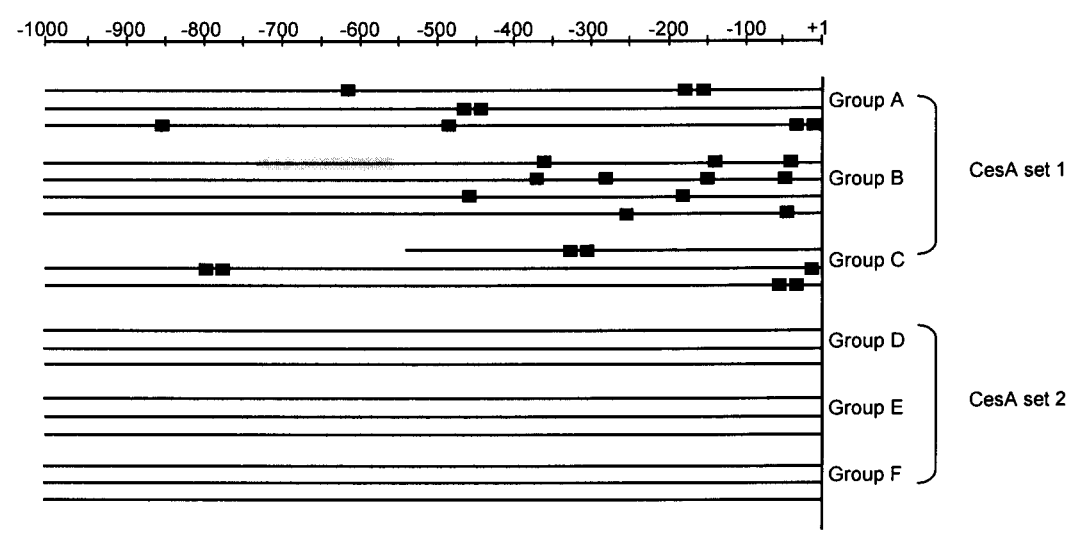
<b>Motif Identity:</b>	CP4	
<b>Consensus sequence:</b>	GACNGT(C/G)NGTGGGGC	
<b>Reverse compliment:</b>	GCCCCACN(C/G)ACNGTC	
<b>Data sets:</b>	Identified in only CesA set 1	
<b>Consensus alignment:</b> A- GNCAS <del>T</del> GN B- GACAGTGC C- CWGKCTGT D- ACMGACWN E- ACNGNCNG F- ANNGACNG G- GGTGGGGC H- GGKGAGGY I- GGTGGNGC <u>GACNGT</u> SNGTGGGGC	<b>POCO:</b>	<b>Number of motif occurrences:</b> 52 <b>Number of promoters with motif:</b> 11/11 <b>Motif representation in the data sets:</b> The motif is 4 time more over represented in CesA set 1 when compared to CesA set 2 and 3 times more represented than in the background set <b>Motif P-value:</b> 0.00274
	<b>Weeder:</b>	<b>Weeder value:</b> 0.52 <b>Motif signature:</b> 
	<b>MotifSampler:</b>	<b>Log-likelihood:</b> 73.17 <b>Information content:</b> 1.5 <b>Consensus score:</b> 1.3 <b>Motif signature:</b> 
	<b>PLACE Hit:</b>	<b>Z-score:</b> 81.8 <b>E-value:</b> 0.39 <b>Place ID:</b> SE1PVGRP18 <b>Place Motif:</b> ATAATGGGCCACACTGTGGGGCAT <b>Putative function:</b> A stem element found in bean to enhance vascular expression strongly but non-specifically
		

*Appendix 2.12: CP5*


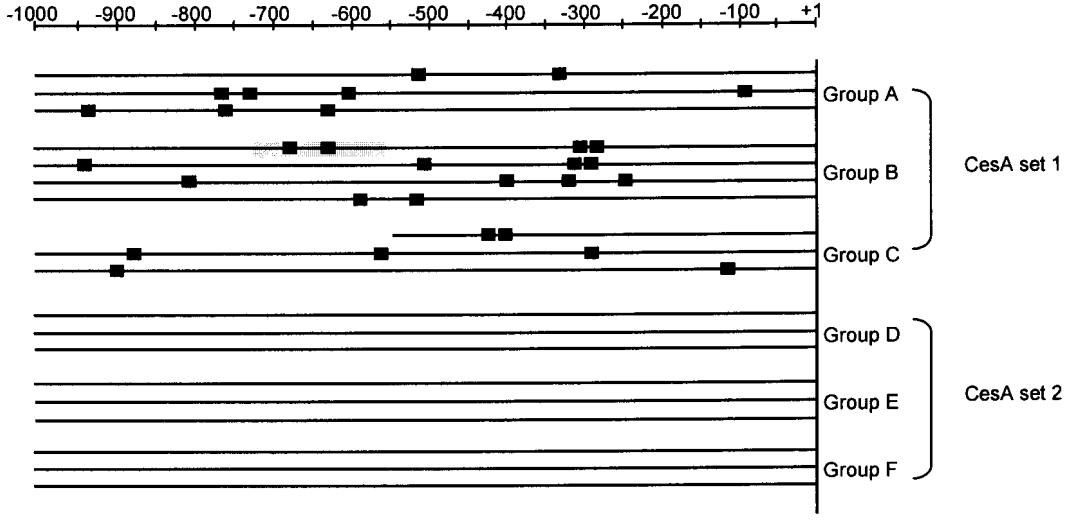
<b>Motif Identity:</b>	CP5	
<b>Consensus sequence:</b>	ATN(AT)ATTA	
<b>Reverse compliment:</b>	TAAT(AT)NAT	
<b>Data sets:</b>	Only identified in Cesa set 1	
<b>Consensus alignment:</b> A- ATTWATTA B- ATNTNTTA ATNWTTA	<b>POCO:</b>	<b>Number of motif occurrences:</b> 39 <b>Number of promoters with motif:</b> 11/11 <b>Motif representation in the data sets:</b> This motif is represented 4 times more in Cesa set 2 than in Cesa set1 and is 2 times higher in the Cesa set 1 than in the background. <b>Motif P-value:</b> 0.00418
	<b>Weeder:</b>	No Weeder Prediction
	<b>MotifSampler:</b>	<b>Log-likelihood:</b> 67.27 <b>Information content:</b> 1.2 <b>Consensus score:</b> 1.6 <b>Motif signature:</b> 
	<b>PLACE Hit:</b>	<b>Z-score:</b> 80.3 <b>E-value:</b> 0.65 <b>Place ID:</b> C2GMAUX28 <b>Place Motif:</b> AATAATAATAATAATAATA <b>Putative function:</b> This motif has been found in the promoters of auxin responsive genes.



*Appendix 2.13: CP6*


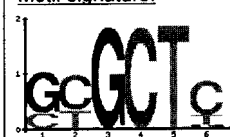
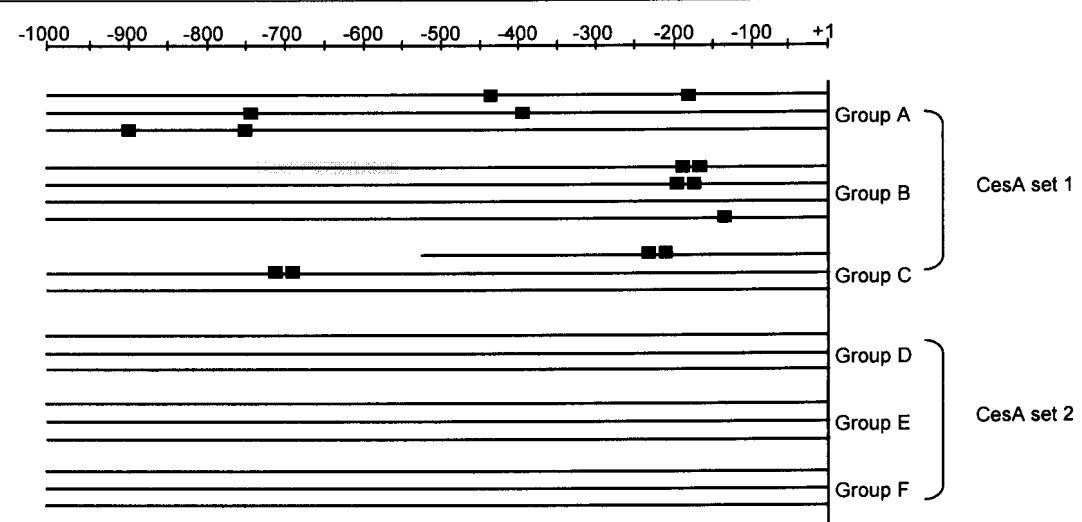
<b>Motif Identity:</b>	CP6	
<b>Consensus sequence:</b>	GC(A/T)NGC	
<b>Reverse compliment:</b>	GCN(A/T)GC	
<b>Data sets:</b>	Only identified in Cesa set 1	
<b>Consensus alignment:</b>  A- GCNTGC B- GCWWGC C- GCWNGC D- GCANGC E- GCAWGC F- GCWKGC G- GCNNGC H- GNANGC I- GCNTNC GCWNGC	<b>POCO:</b>	Number of motif occurrences: 42 Number of promoters with motif: 11 Motif representation in the data sets: This motif is identified in both Cesa set 1 and Cesa set 2 2-fold higher than compared to the background Motif P-value: 0.0287
	<b>Weeder:</b>	No Weeder Prediction
	<b>MotifSampler:</b>	Log-likelihood: 75.63 Information content: 1.7 Consensus score: 1.4 Motif signature: 
	<b>PLACE Hit:</b>	Z-score: 53.2 E-value: 0.31 Place ID: LEGUMINBOXLEGA5 Place Motif: TCCATAGCCATGCAWRCTGMAGAATGTC Putative function: Sequence responsible for the tissue specific expression of the Pea legume gene
		

*Appendix 2.14: CP7*



<u>Motif Identity:</u>	CP7	
<u>Consensus sequence:</u>	NG(A/G)CNGTG	
<u>Reverse compliment:</u>	CACNG(C/T)CN	
<u>Data sets:</u>	Only identified in Cesa set 1	
<u>Consensus alignment:</u> A- NGRCA GTG B- TGNCNGTG NGRCNGTG	<u>POCO:</u>	<u>Number of motif occurrences:</u> 6 <u>Number of promoters with motif:</u> 6/11 <u>Motif representation in the data sets:</u> This motif is represented six times more in Cesa set 1 than in Cesa set 2 and the background <u>Motif P-value:</u> 0.00242
	<u>Weeder:</u>	No Weeder Prediction
	<u>MotifSampler:</u>	<u>Log-likelihood:</u> 64.08 <u>Information content:</u> 1.3 <u>Consensus score:</u> 1.5 <u>Motif signature:</u> 
	<u>PLACE Hit:</u>	<u>Z-score:</u> 15.2 <u>E-value:</u> 1 <u>Place ID:</u> C1GMAUX28 <u>Place Motif:</u> TGAAAACAGTGAGTTA <u>Putative function:</u> This motif has been found in the promoters of auxin responsive genes.
		



*Appendix 2.15: CP8*

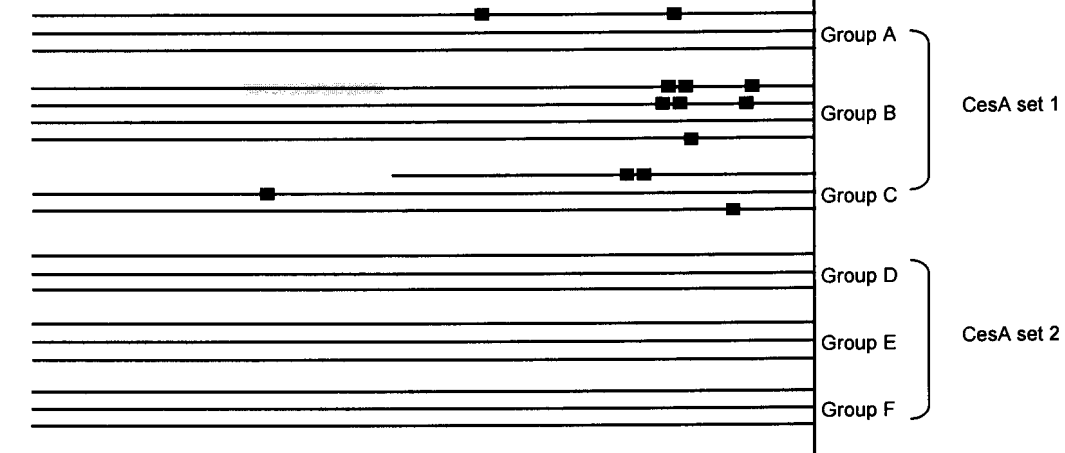
<u>Motif Identity:</u>	CP8	
<u>Consensus sequence:</u>	G(C/T)GCTC	
<u>Reverse compliment:</u>	GAGC(A/G)C	
<u>Data sets:</u>	Only identified in Cesa set 1	
<u>Consensus alignment:</u>  <b>A-</b> GYGCTC <b>B-</b> GCGCTC <u>    </u> GYGCTC	<u>POCO:</u>	No POCO Prediction
	<u>Weeder:</u>	<u>Weeder value:</u> 1.05 <u>Motif signature:</u> 
	<u>MotifSampler:</u>	<u>Log-likelihood:</u> 65.50 <u>Information content:</u> 1.8 <u>Consensus score:</u> 1.4 <u>Motif signature:</u> 
	<u>PLACE Hit:</u>	<u>Z-score:</u> 15.2 <u>E-value:</u> 1 <u>Place ID:</u> GLUTEBP2OS <u>Place Motif:</u> ATGCTCAATAGATATAAGT <u>Putative function:</u> Cis-element identified in the rice gluten genes involved in the gluten gene expression
		

*Appendix 2.16: CP9*

<u>Motif Identity:</u>	CP9	
<u>Consensus sequence:</u>	GAGCG(A/C)	
<u>Reverse compliment:</u>	(G/T)CGCTC	
<u>Data sets:</u>	Identified in CesA set 1	
<u>Consensus alignment:</u>  A- GAGCGM B- GAGCGC GAGCGM	<u>POCO:</u>	No POCO Prediction
	<u>Weeder:</u>	Weeder value: 1.05 Motif signature: 
	<u>MotifSampler:</u>	Log-likelihood: 52.74 Information content: 1.9 Consensus score: 1.6 Motif signature: 
	<u>PLACE Hit:</u>	Z-score: -0.6 E-value: 1 Place ID: ASF1ATNOS Place Motif: TGAGCTAAGCACATACGTCAG Putative function: One of two motifs needed for nopaline synthase genes

-1000   -900   -800   -700   -600   -500   -400   -300   -200   -100   +1



Group A } CesA set 1

Group B } CesA set 1


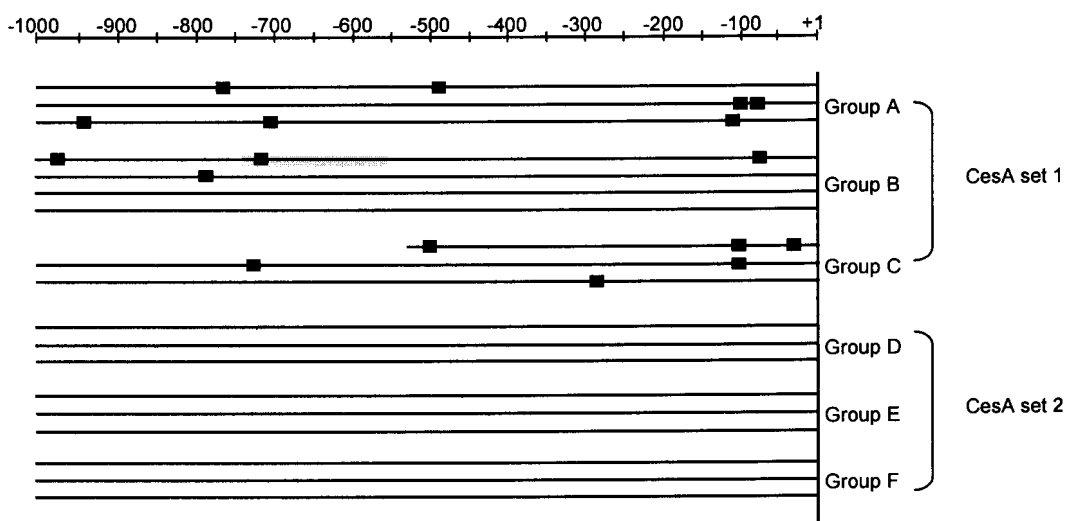
Group C } CesA set 1

Group D } CesA set 2


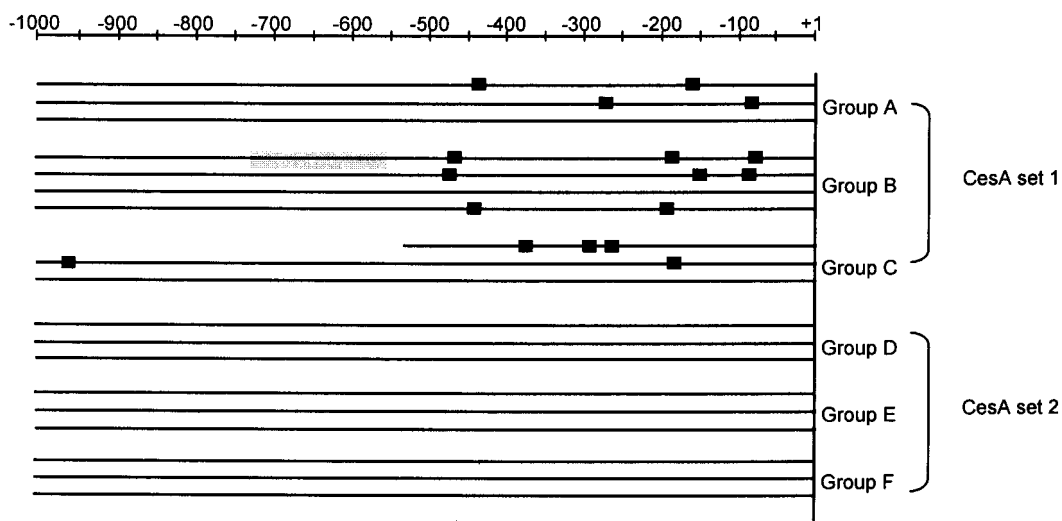
Group E } CesA set 2

Group F } CesA set 2


*Appendix 2.17: CP10*

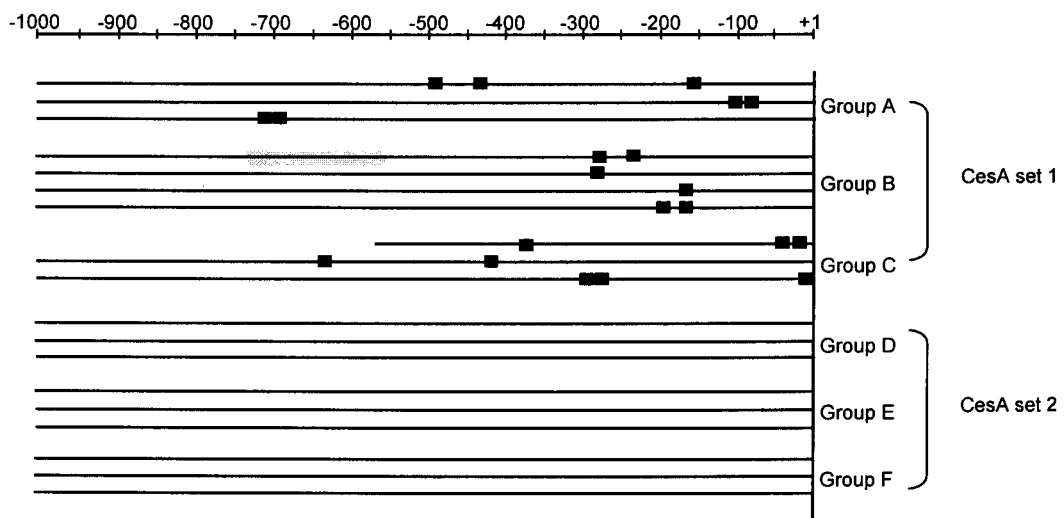
<b>Motif Identity:</b>	CP10	
<b>Consensus sequence:</b>	GTC(G/T)GT	
<b>Reverse compliment:</b>	AC(A/C)GAC	
<b>Data sets:</b>	Only identified in CesaA set 1	
<b>Consensus alignment:</b>  A- GTCCKGT B- GTCNTG GTCKGT	<b>PQQQ:</b>	Number of motif occurrences: 20 Number of promoters with motif: 11/11 Motif representation in the data sets: This motif is represented 20 times more in CesaA set 1 than in CesaA set 2 and is at least 2 times higher than the background Motif P-value: 0.000413
	<b>Weeder:</b>	No Weeder Prediction
	<b>MotifSampler:</b>	Log-likelihood: 72.27 Information content: 1.8 Consensus score: 1.5 Motif signature: 
	<b>PLACE Hit:</b>	Z-score: -3.8 E-value: 1 Place ID: MYBCORE Place Motif: CNGTTR Putative function: The core sequence of a MYB binding site similar to the <i>Arabidopsis</i> MYBs involved in stress and similar to the <i>Patunia</i> MYB involved in flavinoid biosynthesis.
		

*Appendix 2.18: CP11*

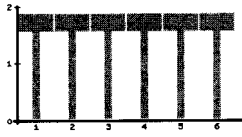
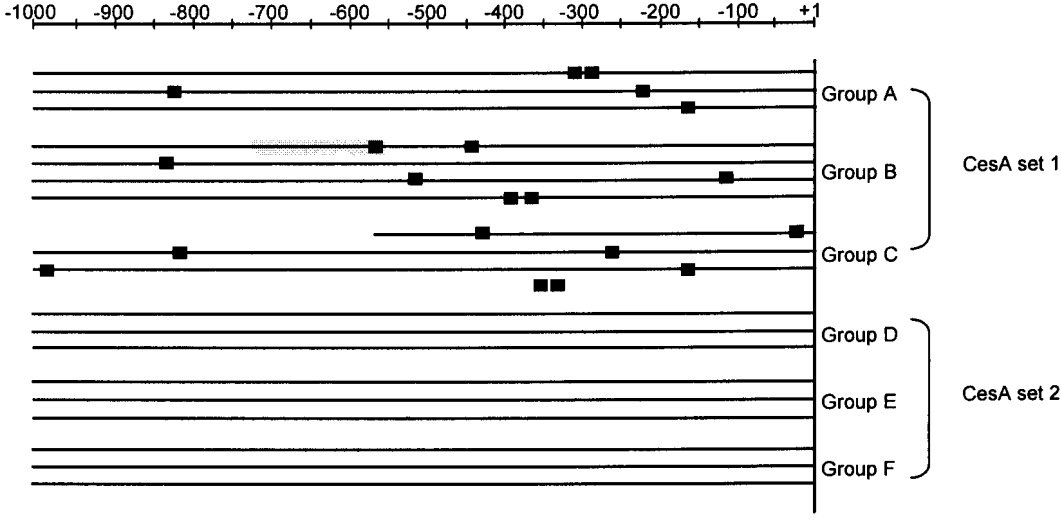
<b>Motif Identity:</b>	CP11	
<b>Consensus sequence:</b>	ANNGA(C/T)AG	
<b>Reverse compliment:</b>	CT(A/G)TCNNT	
<b>Data sets:</b>	Identified in CesA set 1	
<b>Consensus alignment:</b>  A- AGNGAYAG B- ANTGNAG C- ANNGACNG <u>ANNGAYAG</u>	<b>POCO:</b>	<b>Number of motif occurrences:</b> 21 <b>Number of promoters with motif:</b> 11/11 <b>Motif representation in the data sets:</b> This motif is 2 times more represented in CesA set 1 than in the CesA set 2 and the background <b>Motif P-value:</b> 0.0124
	<b>Weeder:</b>	No Weeder Prediction
	<b>MotifSampler:</b>	<b>Log-likelihood:</b> 61.08 <b>Information content:</b> 1.5 <b>Consensus score:</b> 1.4 <b>Motif signature:</b> 
	<b>PLACE Hit:</b>	<b>Z-score:</b> -3.8 <b>E-value:</b> 1 <b>Place ID:</b> ARE1 <b>Place Motif:</b> RGTGACNNGC <b>Putative function:</b> Similar to the rat anti oxidant response element of the glutathione s-transferase gene
		

Appendix 2.19: CP12


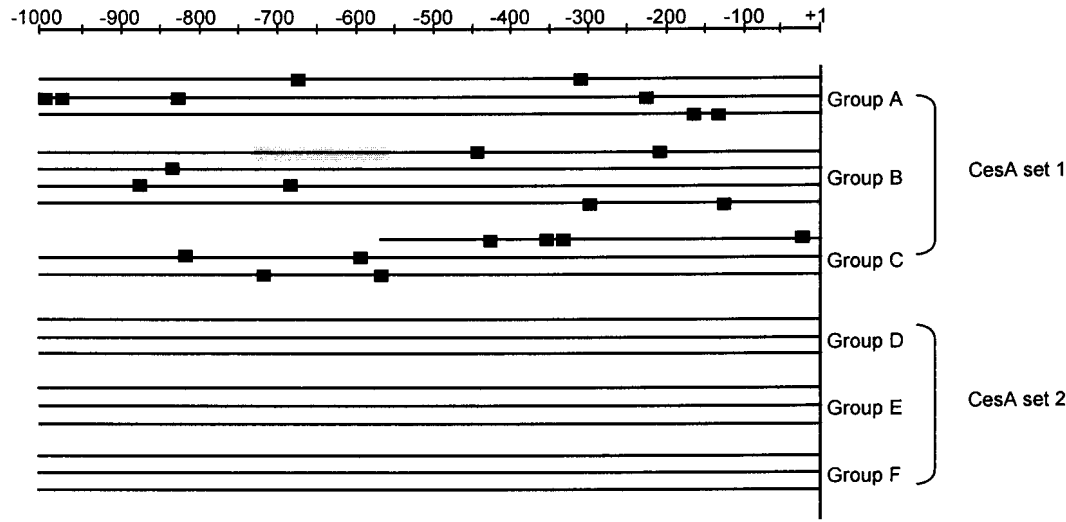
<b>Motif Identity:</b>	CP12	
<b>Consensus sequence:</b>	ACAGNCNG	
<b>Reverse compliment:</b>	CNGNCTGT	
<b>Data sets:</b>	Only identified in Cesa set 1	
<b>Consensus alignment:</b> A- ASAGNCTG B- ACAGRCWG C- ACNGNCNG <u>ACAGNCNG</u>	<b>POCO:</b>	<b>Number of motif occurrences:</b> 12 <b>Number of promoters with motif:</b> 8/11 <b>Motif representation in the data sets:</b> This motif is represented 4 times in Cesa set 1 more than in the Cesa set 2 and the back ground <b>Motif P-value:</b> 0.000394
	<b>Weeder:</b>	No Weeder Prediction
	<b>MotifSampler:</b>	<b>Log-likelihood:</b> 89.97 <b>Information content:</b> 1.5 <b>Consensus score:</b> 1.3 <b>Motif signature:</b> 
	<b>PLACE Hit:</b>	<b>Z-score:</b> -3.8 <b>E-value:</b> 1 <b>Place ID:</b> GARE4HVEPB1 <b>Place Motif:</b> GTAACAGAATGCTGG <b>Putative function:</b> Sequence plays a role in the coordinate gene expression regulated by gibberellins and abscisic acid




*Appendix 2.20: CP13*

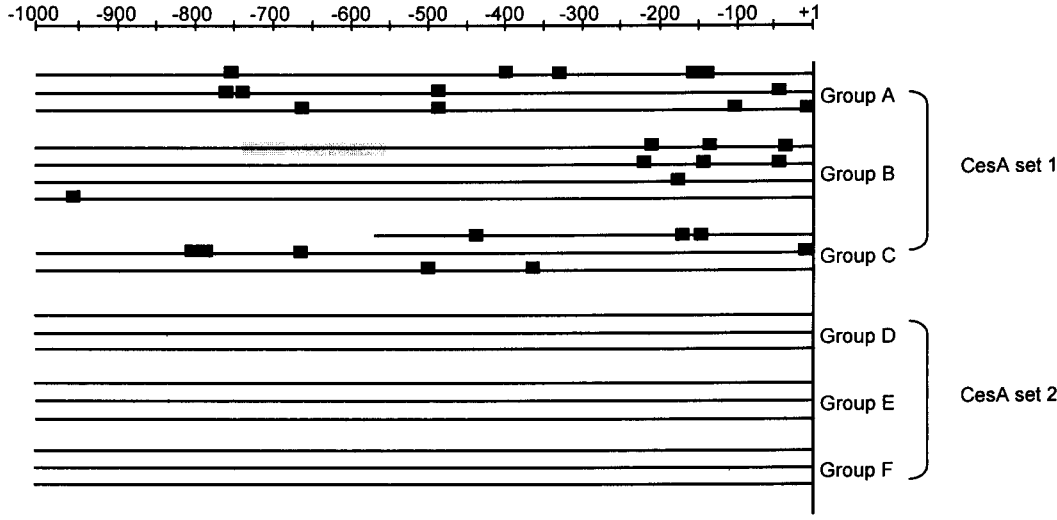
<u>Motif Identity:</u>	CP13	
<u>Consensus sequence:</u>	TTTTTT	
<u>Reverse compliment:</u>	AAAAAA	
<u>Data sets:</u>	Only identified in Cesa set 1	
<u>Consensus alignment:</u>  A- TTTTTT B- TTTTTT C- TTTTNT TTTTTT	<u>POCO:</u>	<u>Number of motif occurrences:</u> 265 <u>Number of promoters with motif:</u> 11/11 <u>Motif representation in the data sets:</u> This motif is present two fold higher in Cesa set 1 than in Cesa set 2 or the background <u>Motif P-value:</u> 0.0409
	<u>Weeder:</u>	No Weeder Prediction
	<u>MotifSampler:</u>	<u>Log-likelihood:</u> 58.10 <u>Information content:</u> 1.5 <u>Consensus score:</u> 1.9 <u>Motif signature:</u> 
	<u>PLACE Hit:</u>	<u>Z-score:</u> -29.1 <u>E-value:</u> 1 <u>Place ID:</u> PYRIMIDINEBOXHVEPB1 <u>Place Motif:</u> TTTTTTCC <u>Putative function:</u> A pyrimidine box required for gibberellic acid induction
		

*Appendix 2.21: CP14*

<b>Motif Identity:</b>	CP14	
<b>Consensus sequence:</b>	AAAAAA	
<b>Reverse compliment:</b>	TTTTTT	
<b>Data sets:</b>	Only identified in Cesa set 1	
<b>Consensus alignment:</b> A- AAAAAA B- AAAAAA C- ANAAAA AAAAAA	<b>POCO:</b>	<b>Number of motif occurrences:</b> 265 <b>Number of promoters with motif:</b> 11/11 <b>Motif representation in the data sets:</b> This motif is present two fold higher in Cesa set 1 than in Cesa set 2 or the background <b>Motif P-value:</b> 0.0409
	<b>Weeder:</b>	No Weeder Prediction
	<b>MotifSampler:</b>	<b>Log-likelihood:</b> 51.56 <b>Information content:</b> 1.4 <b>Consensus score:</b> 1.9 <b>Motif signature:</b> 
	<b>PLACE Hit:</b>	<b>Z-score:</b> -29.1 <b>E-value:</b> 1 <b>Place ID:</b> 3AF1BOXPSRBCS3 <b>Place Motif:</b> AAATAGATAAATAAAAAACATT <b>Putative function:</b> An At rich region found to play a role in the expression of light responsive genes
		

*Appendix 2.22: CP15*


<b>Motif Identity:</b>	CP15	
<b>Consensus sequence:</b>	C(C/T)C(C/G)NCCC	
<b>Reverse compliment:</b>	GGGN(C/G)G(A/G)G	
<b>Data sets:</b>	Only identified in Cesa set 1	
<b>Consensus alignment:</b> A- SYCSWSCC B- CNCCMCCN C- CNCNNCNC CYCSNCCC	<b>POCO:</b>	<b>Number of motif occurrences:</b> 47 <b>Number of promoters with motif:</b> 11/11 <b>Motif representation in the data sets:</b> This motif occurs 2 times more in Cesa set 1 and Cesa set2 than the background <b>Motif P-value:</b> 0.02806
	<b>Weeder:</b>	No Weeder Prediction
	<b>MotifSampler:</b>	<b>Log-likelihood:</b> 68.39 <b>Information content:</b> 1.6 <b>Consensus score:</b> 1.1 <b>Motif signature:</b> 
<b>PLACE Hit:</b>	No PLACE Hit	

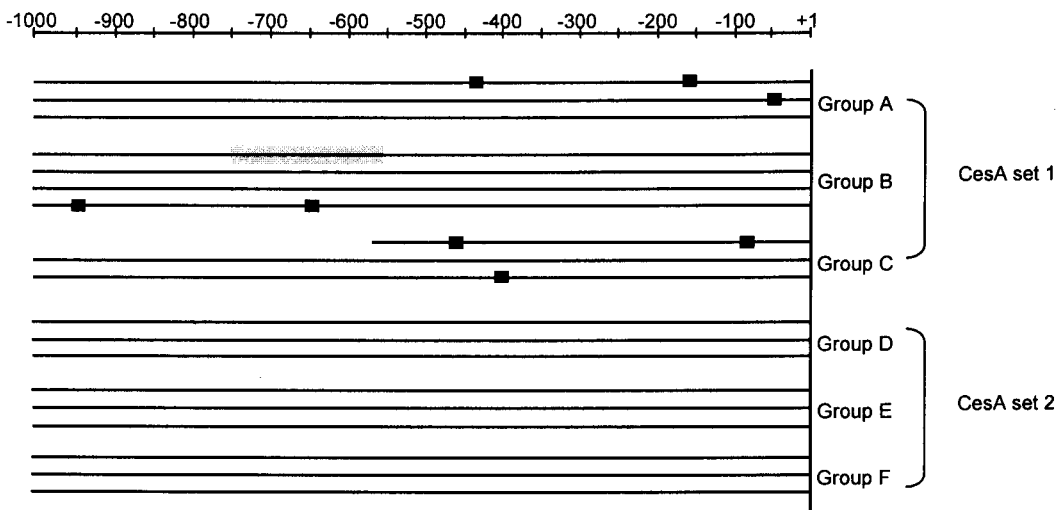
  


-1000   -900   -800   -700   -600   -500   -400   -300   -200   -100   +1



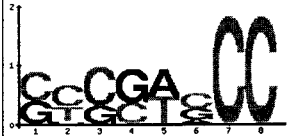
Appendix 2.23: CP16

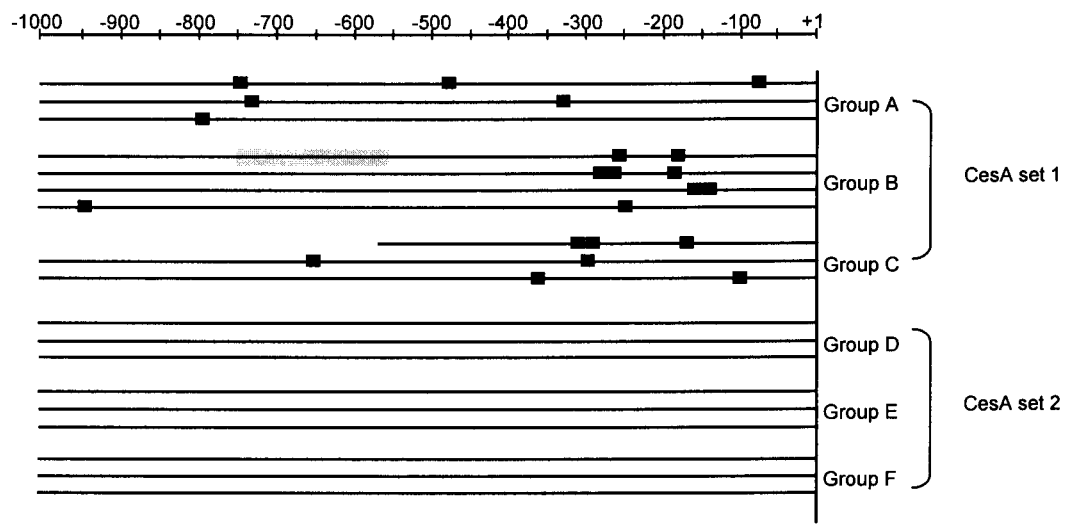
<b>Motif Identity:</b>	CP16	
<b>Consensus sequence:</b>	TNNCN(G/T)NC	
<b>Reverse compliment:</b>	GN(A/C)NGNNA	
<b>Data sets:</b>	Only identified in Cesa set 1	
<b>Consensus alignment:</b> A- TNGCTKTC B- TNNCNTNC TNNCNKNC	<b>POCO:</b>	<b>Number of motif occurrences:</b> 68 <b>Number of promoters with motif:</b> 11/11 <b>Motif representation in the data sets:</b> This motif is 1.5 time more represented in the Cesa set 1and Cesa set 2 than the background <b>Motif P-value:</b> 0.021
	<b>Weeder:</b>	No Weeder Prediction
	<b>MotifSampler:</b>	<b>Log-likelihood:</b> 39.36 <b>Information content:</b> 1.4 <b>Consensus score:</b> 1.3 <b>Motif signature:</b> 
<b>PLACE Hit:</b>	No PLACE Hit	




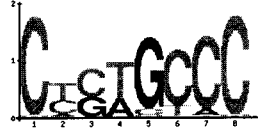
Appendix 2.24: CP17

<b>Motif Identity:</b>	CP17	
<b>Consensus sequence:</b>	GG(C/G)(A/T)(C/G)G(A/G)(G/C)	
<b>Reverse compliment:</b>	(C/G)(C/T)C(C/G)(A/T)(C/G)CC	
<b>Data sets:</b>	Identified in CesA set 1	
<b>Consensus alignment:</b> A- GGSWSGGS B- GGSWSGRS C- GNGNNGNG GGWSGRS	<b>POCO:</b>	<b>Number of motif occurrences:</b> 47 <b>Number of promoters with motif:</b> 11/11 <b>Motif representation in the data sets:</b> This motif is two time more represented in the two CesA data sets than the <i>Arabidopsis</i> background <b>Motif P-value:</b> 0.0281
	<b>Weeder:</b>	No Weeder Prediction
	<b>MotifSampler:</b>	<b>Log-likelihood:</b> 82.37 <b>Information content:</b> 1.6 <b>Consensus score:</b> 1.2 <b>Motif signature:</b> 
	<b>PLACE Hit:</b>	No PLACE Hit

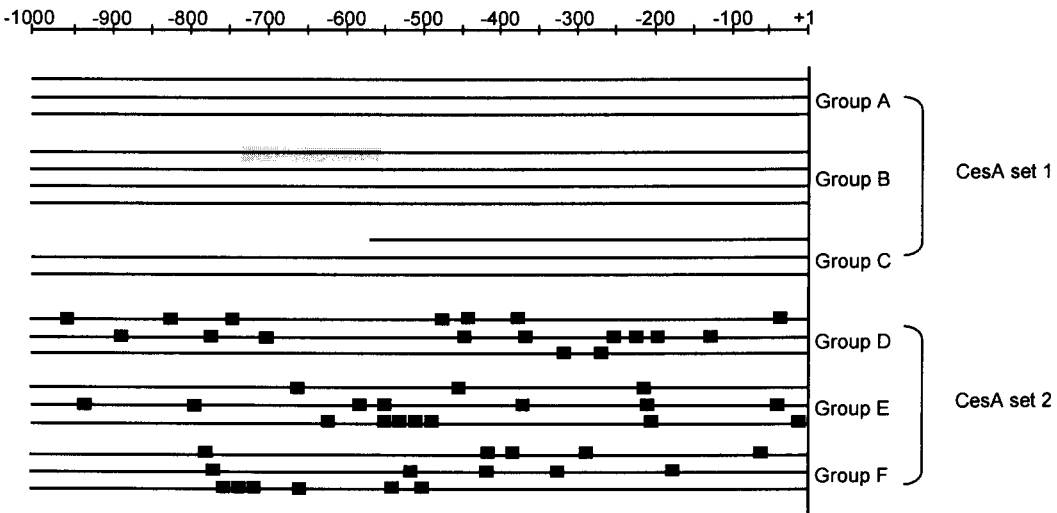
  


## Motifs identified in Cesa set 2


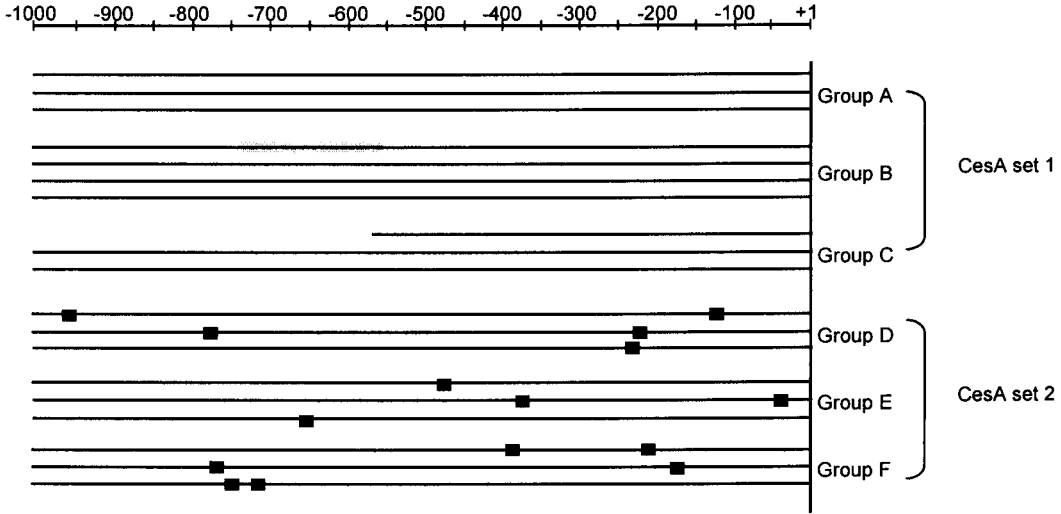
### Appendix 2.25: CS1

<b>Motif Identity:</b>	CS1		
<b>Consensus sequence:</b>	(A/G)C(C/T)(C/G)TGCCC		
<b>Reverse complement:</b>	GGGCA(C/G)(A/G)G(C/T)		
<b>Data sets:</b>	Only found in Cesa set 2		
<b>Consensus alignment:</b>  <b>A-</b> RNYSTGCC <b>B-</b> RMCNTGCC <b>C-</b> RCYSTGCC <b>D-</b> RCYNGGCC <b>E-</b> RCNNGGCC <b>F-</b> NMYNNGCC <b>G-</b> RCYSTGCC <b>H-</b> CYSTGCCR <b>I-</b> CTNNGCCC <b>J-</b> CNSTGCC <b>K-</b> CNSWGCCC <b>L-</b> CYSWGCCC <b>M-</b> STGCC <b>N-</b> ANCNTNNC <b>O-</b> CNCNNCNC <b>P-</b> CNCTNCNC <b>Q-</b> GCTGTGCC <b>R-</b> CTGTGCC <b>S-</b> GTGCC <b>RCYSTGCC</b>	<b>POCO:</b>  	<b>Number of motif occurrences:</b> 48 <b>Number of promoters with motif:</b> 8/8 promoters <b>Motif representation in the data sets:</b> 2 fold more occurrences in Cesa set 2 than in Cesa set 1 <b>Motif P-value:</b> 0.0021	
	<b>Weeder:</b>  	<b>Weeder value:</b> 0.7 <b>Motif signature:</b> 	
	<b>MotifSampler:</b>  	<b>Log-likelihood:</b> 59 <b>Information content:</b> 1.5 <b>Consensus score:</b> 1.2 <b>Motif signature:</b> 	
	<b>PLACE Hit:</b>  	<b>Z-score:</b> 93.6 <b>E-value:</b> 0.25 <b>Place ID:</b> RNFG2OS <b>Place Motif:</b> CCAGTGTGCCCTGG <b>Putative function:</b> GATA-motif binds the GATA motif binding factor. The GATA motif is required for phloem-specific gene expression.	



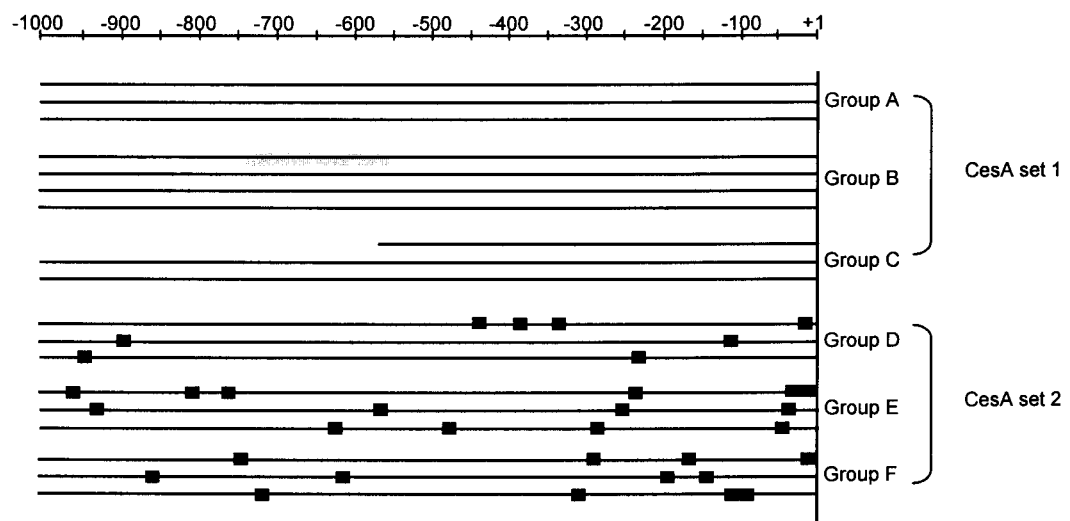
  



*Appendix 2.26: CS2*



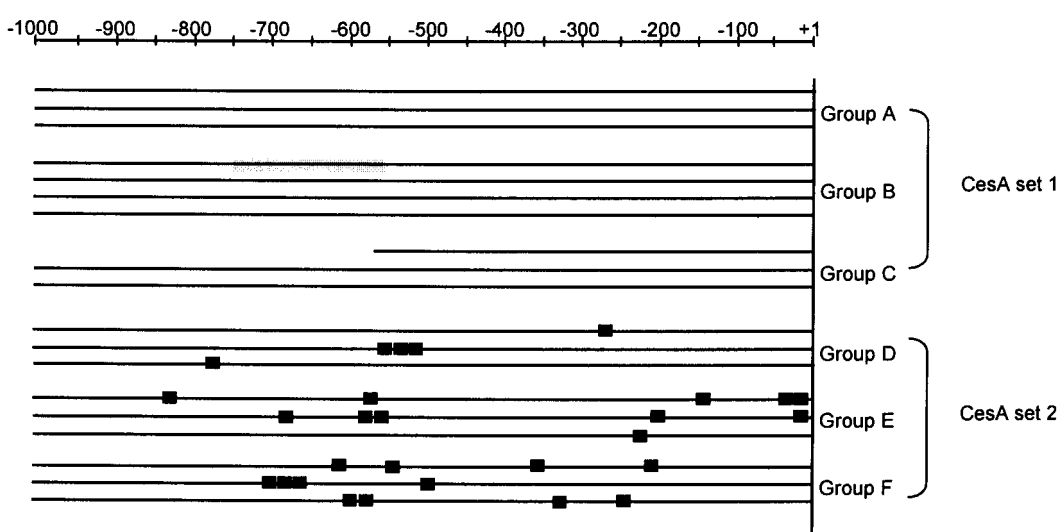
<u>Motif Identity:</u>	CS2	
<u>Consensus sequence:</u>	TCCTGC(C/T)G	
<u>Reverse compliment:</u>	G(A/G)GCAGGA	
<u>Data Set:</u>	Only identified in Cesa set 2	
<u>Consensus alignment:</u> <b>A-</b> TCCTGCTG <b>B-</b> TCCTKYTG <b>C-</b> YNTGCC <b>D-</b> CATGNC <u>TCCTGCTG</u>	<u>POCO:</u>	<u>Number of motif occurrences:</u> 21 <u>Number of promoters with motif:</u> 8/8 promoters <u>Motif representation in the data sets:</u> Motif is represented 3.5 folds more in Cesa set 2 compared to Cesa set 1 <u>Motif P-value:</u> 0.04
	<u>Weeder:</u>	Motif was not identified by Weeder
	<u>MotifSampler:</u>	<u>Log-likelihood:</u> 64 <u>Information content:</u> 1.8 <u>Consensus score:</u> 1.4 <u>Motif signature:</u> 
	<u>PLACE Hit:</u>	<u>Z-score:</u> 79.7 <u>E-value:</u> 0.2 <u>Place ID:</u> SORLREP5AT <u>Place Motif:</u> TTGCATGACT <u>Putative function:</u> A sequence found to be over-represented <i>Arabidopsis</i> light-repressed promoters
		

Appendix 2.27: CS3



<b>Motif Identity:</b>	CS3	
<b>Consensus sequence:</b>	(C/G)CTGAAGG	
<b>Reverse compliment:</b>	CCTTCAG(C/G)	
<b>Data Set:</b>	Only identified in Cesa set 2	
<b>Consensus alignment:</b> A- NCNGMAGG B- SYTSAAGN C- GCTGAAGG SCTGAAGG	<b>POCO:</b>	<b>Number of motif occurrences:</b> 28 <b>Number of promoters with motif:</b> 8/8 Promoters <b>Motif representation in the data sets:</b> Motif not predicted by POCO <b>Motif P-value:</b> Motif not predicted by POCO
	<b>Weeder:</b>	<b>Weeder value:</b> 0.3 <b>Motif signature:</b> 
	<b>MotifSampler:</b>	<b>Log-likelihood:</b> 59 <b>Information content:</b> 1.5 <b>Consensus score:</b> 1.3 <b>Motif signature:</b> 
	<b>PLACE Hit:</b>	<b>Z-score:</b> 74.3 <b>E-value:</b> 0.62 <b>Place ID:</b> CGF1ATCAB2 <b>Place Motif:</b> GATAAAGATTACTTCAGATATAACAAACGTTAC <b>Putative function:</b> I-Box, Part of the GATA binding element involved in Phytochrome and circadian gene regulation
		

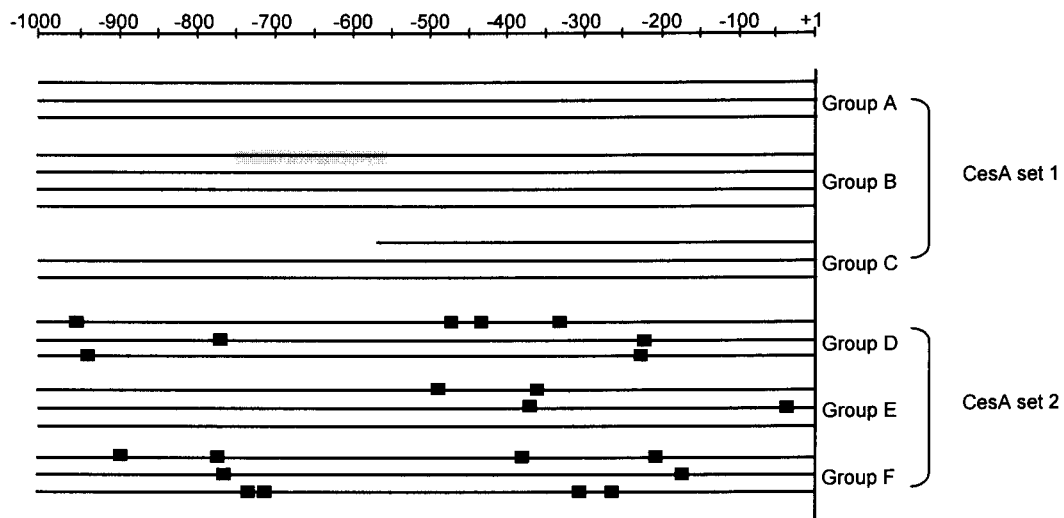


Appendix 2.28: CS4


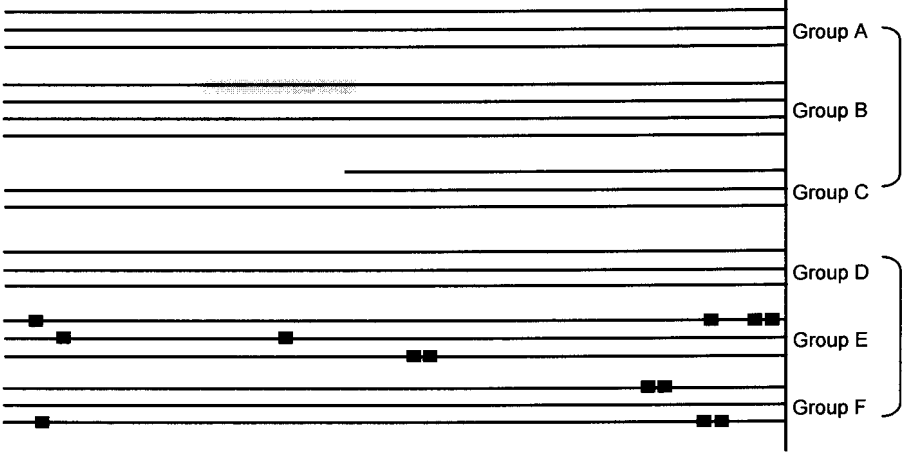
<b>Motif Identity:</b>	CS4	
<b>Consensus sequence:</b>	NNGCATGC	
<b>Reverse compliment:</b>	GCATGCNN	
<b>Data Set:</b>	Identified in Cesa set 2	
<b>Consensus alignment:</b> A- NNGCATGC B- GCATGC C- GCANGC D- GAGCATGC E- GCATGC NNGCATGC	<b>POCO:</b>	<b>Number of motif occurrences:</b> 10 <b>Number of promoters with motif:</b> 5/8 promoters <b>Motif representation in the data sets:</b> Motif not predicted by POCO <b>Motif P-value:</b> Motif not predicted by POCO
	<b>Weeder:</b>	<b>Weeder value:</b> 0.25 <b>Motif signature:</b> 
	<b>MotifSampler:</b>	<b>Log-likelihood:</b> 50 <b>Information content:</b> 1.7 <b>Consensus score:</b> 1.4 <b>Motif signature:</b> 
	<b>PLACE Hit:</b>	<b>Z-score:</b> 82.9 <b>E-value:</b> 0.32 <b>Place ID:</b> IDE1HVIDS2 <b>Place Motif:</b> ATCAAGCATGCTTCTTGC <b>Putative function:</b> Plays a role in the induction of genes expressed during iron deficiency in Barley.
		

*Appendix 2.29: CS5*

<b>Motif Identity:</b>	CS5	
<b>Consensus sequence:</b>	TCCT(G/T)(C/T)TG	
<b>Reverse compliment:</b>	CA(A/G)(A/C)AGGA	
<b>Data Set:</b>	Only identified in Cesa set 2	
<b>Consensus alignment:</b>  A-TCCTKYTG B-TCCTGCTG TCCTKYTG	<b>POCO:</b>	<b>Number of motif occurrences:</b> 20 <b>Number of promoters with motif:</b> 7/8 Promoters <b>Motif representation in the data sets:</b> This motif was not predicted by POCO <b>Motif P-value:</b> This motif was not predicted by POCO
	<b>Weeder:</b>	<b>Weeder value:</b> 0.37 <b>Motif signature:</b> 
	<b>MotifSampler:</b>	<b>Log-likelihood:</b> 50 <b>Information content:</b> 1.5 <b>Consensus score:</b> 1.3 <b>Motif signature:</b> 
	<b>PLACE Hit:</b>	<b>Z-score:</b> 68.6 <b>E-value:</b> 1.8 <b>Place ID:</b> ANAEROBICCISZMGAPC4 <b>Place Motif:</b> CGAAACCAGCAACGGTCCAG <b>Putative function:</b> Required for anaerobic gene expression

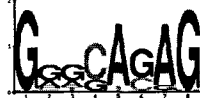


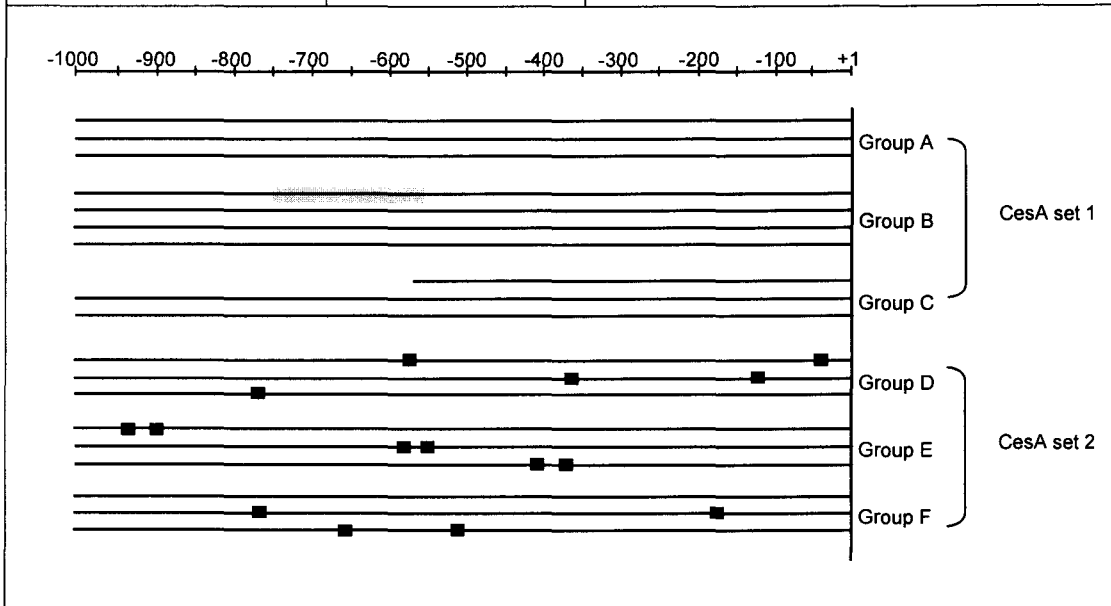
*Appendix 2.30:CS6*

<u>Motif Identity:</u>	CS6	
<u>Consensus sequence:</u>	NNNT(C/G)AAG	
<u>Reverse compliment:</u>	CTT(C/G)ANNN	
<u>Data Set:</u>	Only identified in CesA set 2	
<u>Consensus alignment:</u> A- NNTTSAAG B- AANTNAAG NNNTSAAG	<u>POCO:</u>	<u>Number of motif occurrences:</u> 32 <u>Number of promoters with motif:</u> 11/11 <u>Motif representation in the data sets:</u> represented more than 2 folds higher in CesA set 1 than in CesA set 2 <u>Motif P-value:</u> 0.012
	<u>Weeder:</u>	Motif was not predicted by Weeder
	<u>MotifSampler:</u>	<u>Log-likelihood:</u> 35 <u>Information content:</u> 1.4 <u>Consensus score:</u> 1.4 <u>Motif signature:</u> 
	<u>PLACE Hit:</u>	<u>Z-score:</u> 2.6 <u>E-value:</u> 1 <u>Place ID:</u> HSE <u>Place Motif:</u> CTNGAANN TTCNAG <u>Putative function:</u> Heat shock response element consensus sequence found in the promoter regions of heat shock proteins
<p style="text-align: center;">-1000   -900   -800   -700   -600   -500   -400   -300   -200   -100   +1</p> 		




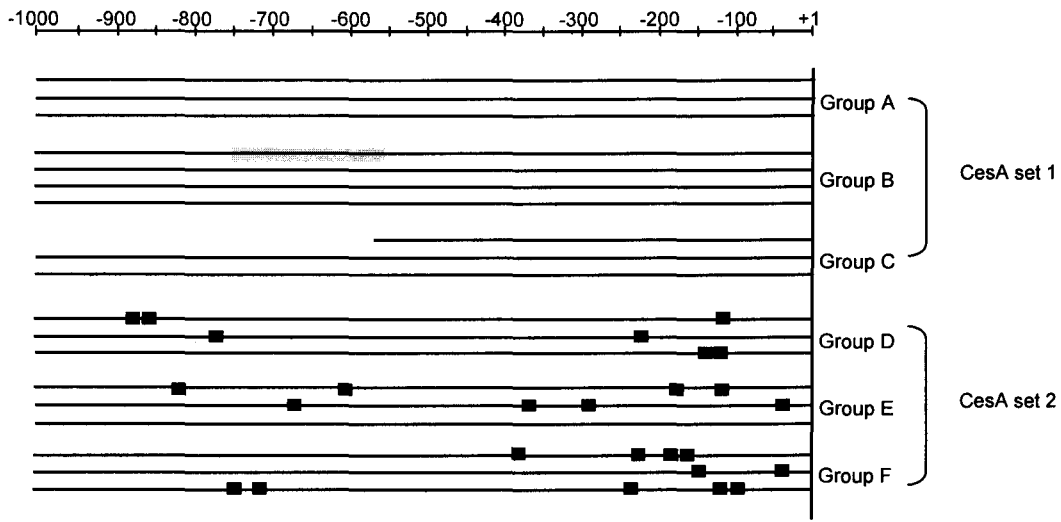
*Appendix 2.31: CS7*

<b>Motif Identity:</b>	CS7	
<b>Consensus sequence:</b>	GNGNAGNG	
<b>Reverse compliment:</b>	CNCTNCNC	
<b>Data Set:</b>	Only identified in Cesa set 2	
<b>Consensus alignment:</b> A- GNNCAGAG B- GNGNNGNG C- GNGNAGNG GNGNAGNG	<b>POCO:</b>	<u>Number of motif occurrences:</u> Cesa set 1= 44 and Cesa set 2 = 30 <u>Number of promoters with motif:</u> Cesa set 1= 9/11 and Cesa 8/8 <u>Motif representation in the data sets:</u> represented 2 fold higher than the background <u>Motif P-value:</u> 0.03
	<b>Weeder:</b>	This motif was not predicted by Weeder
	<b>MotifSampler:</b>	<u>Log-likelihood:</u> 1.2 <u>Information content:</u> 1.5 <u>Consensus score:</u> 58 <u>Motif signature:</u> 
	<b>PLACE Hit:</b>	<u>Z-score:</u> 2.6 <u>E-value:</u> 1 <u>Place ID:</u> -141NTG13 <u>Place Motif:</u> GCTTTTGATGACTTCAAACAC <u>Putative function:</u> Auto regulation of transcription in root tip meristems



*Appendix 2.32: CS8*

<b>Motif Identity:</b>	CS8	
<b>Consensus sequence:</b>	(A/G)N(C/G)(C/T)T(A/G)(C/G)C	
<b>Reverse compliment:</b>	G(C/G)(C/T)A(G/A)(C/G)N(C/T)	
<b>Data Set:</b>	Only identified in Cesa set 2	
<b>Consensus alignment:</b>  <b>A- AGSTWANC</b> <b>B- RNCYTRCC</b> <b>C- ANGNCATG</b> <b>  ANSYWAYC</b>	<b>POCO:</b>	<b>Number of motif occurrences: 78</b> <b>Number of promoters with motif: 8/8 Promoters</b> <b>Motif representation in the data sets: Represented 2 fold higher in the Cesa set 2 than in Cesa set 1</b> <b>Motif P-value: 0.0009</b>
	<b>Weeder:</b>	Motif not identified by Weeder
	<b>MotifSampler:</b>	<b>Log-likelihood: 64</b> <b>Information content: 1.4</b> <b>Consensus score: 1.1</b> <b>Motif signature:</b> 
	<b>PLACE Hit:</b>	No PLACE Hit

The genomic map displays motif occurrences across six groups (A-F). The x-axis represents distance from -1000 to +1. Groups A, B, and C are associated with Cesa set 1, while groups D, E, and F are associated with Cesa set 2. Black squares indicate motif hits. In Cesa set 1, hits are sparse. In Cesa set 2, hits are more frequent, particularly in groups D, E, and F, with multiple hits per promoter region.