uOttawa

L'Université canadienne
Canada's university

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTOCTORALES

uOttawa

L'Université canadienne
Canada's university

FACULTY OF GRADUATE AND POSDOCTORAL STUDIES

Neda Mansoorian

AUTEUR DE LA THÈSE / AUTHOR OF THESIS

M.Sc. (Biology)

GRADE / DEGREE

Department of Biology

FACULTÉ, ÉCOLE, DÉPARTEMENT / FACULTY, SCHOOL, DEPARTMENT

Analysis of a Novel myb-like Gene from Soybean

TITRE DE LA THÈSE / TITLE OF THESIS

D. A. Johnson

DIRECTEUR (DIRECTRICE) DE LA THÈSE / THESIS SUPERVISOR

CO-DIRECTEUR (CO-DIRECTRICE) DE LA THÈSE / THESIS CO-SUPERVISOR

EXAMINATEURS (EXAMINATRICES) DE LA THÈSE / THESIS EXAMINERS

C. Martin

T. Ouellet

J. Vierula

Gary W. Slater

LE DOYEN DE LA FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTDOCTORALES /
DEAN OF THE FACULTY OF GRADUATE AND POSTDOCORAL STUDIES

Analysis of a Novel myb-like Gene from Soybean

Neda Mansoorian

Thesis submitted to the
Faculty of Graduate Studies and Research
University of Ottawa
in partial fulfillment of the requirements for a Master's degree
Ottawa-Carleton Institute of Biology

Your file Votre référence
ISBN: 0-494-11339-1
Our file Notre référence
ISBN: 0-494-11339-1

Dedicated to the Baha'i youth of Iran

ABSTRACT

This thesis reports the identification and characterization of a novel gene, named 7/2, from soybean (*Glycine max* L. Merrill). Genomic clones of 7/2 from two different cultivars, cvs Maple Arrow and Resnik were isolated and sequenced, and the sequences were compared to each other. These sequences were 100% identical. Alignment with a cDNA sequence from cv. Maple Arrow revealed the structure of the gene as having 6 exons and 5 introns.

The full length cDNA contains an ORF of 798 bp encoding a novel protein of 265 amino acids with a calculated molecular weight of 29.98 kDa. The conceptual 7/2 protein is similar in structure to Myb-like transcription factors in that it has an N-terminal DNA binding domain of 52 amino acid and a potential acidic activation domain. Unlike most known plant Myb transcription factors it has one, not two, N-terminal DNA binding repeats. As determined by RT-PCR, the7/2 message is expressed in most soybean tissues but is most highly expressed in nodules. These characteristics suggest that 7/2 may represent a novel soybean regulatory protein.

RESUME

Cette thèse présente l'identification et la caractérisation d'un nouveau gène, appelé 7/2, dans le soja (*Glycine max* L. Merrill). Des clones génomiques de 7/2 de deux variétés, v. Maple Arrow et Resink, furent isolés et séquencés, puis les séquences comparées. Ces séquences étaient 100% identiques. L'alignement avec l'ADNc de v. Maple Arrow a révélé que le gène contient 6 exons et 5 introns.

L'ADNc complet contient un cadre de lecture de 798 pb codant pour une nouvelle protéine de 265 acides aminés ayant un poids moléculaire de 29.98 kDa. La protéine 7/2 codée par le gène 7/2 a une structure similaire aux facteurs de transcription de type Myb, avec un domaine de liaison à l'ADN à l'extrémité N-terminale de 52 acides aminés ainsi qu'un domaine potentiel d'activation acide. Contrairement aux facteurs de transcription Myb connus, il n'y a qu'une seule, et non deux, répétition du domaine de liaison à l'ADN à l'extrémité N-terminale. Des expériences de réaction de polymérase en chaîne ont déterminé que l'ARN messager de 7/2 est exprimé dans la plupart des tissus du soja mais est exprimé au plus haut niveau dans les nodules. Ces caractéristiques suggèrent que 7/2 est une nouvelle protéine régulatrice.

ACKNOWLEDGEMENTS

I would like to thank my supervisor Dr. Douglas Johnson for his guidance and support throughout my Masters education.

Thanks to the members of my research committee, Dr. C. Martin and Dr. S. Gleddie for their helpful comments and advice throughout this course of study. I would also like to thank Dr. L. Bonen and Dr. G. Drouin for their advices and help as well.

I appreciate and I am thankful to my Lab mate, Candace Webb, for her support, patient, guidance and friendship and thanks to the next door lab members, Jennifer, Sophie and Tom for providing a very happy and friendly working environment beside their help when ever was needed.

Sincere thank to my dear family and friends for their thoughtfulness, kindness and support during my stay in Canada and finally my special thanks go to the world wide Baha'i community, in particular the Baha'i communities of Iran and Canada, for their outstanding and generous support.

TABLE OF CONTENTS

ABSTRACT i
RESUME ii
ACKNOWLEDGEMENTS iii
TABLE OF CONTENTS iv
LIST OF FIGURES vii
LIST OF TABLES viii
LIST OF ABBREVIATIONS ix
CHAPTER ONE 1
INTRODUCTION 1
1.1 Gene Regulation in Eukaryotes 1
1.1.1 Gene Regulation at the Chromatin Level 1
1.1.2 Gene Regulation at the Transcriptional Level 3
1.1.3 Gene Regulation at the Post-Transcriptional Level 4
1.1.4 Gene Regulation at the Translational and Post-Translational level 5
1.2 Gene Regulation in Plants 6
1.3 Transcription Factors 6
1.3.1 Transcription Factors in Plants 8
1.3.2 Classification of Transcription Factors in Plants 9
1.4 Myb Family of Transcription Factors 13
1.4.1 Myb Transcription Factor Structure in Plants 14
1.4.2 Single Myb-repeat Proteins 17
1.4.3 Myb Transcription Factor Function in Plants 18
1.5 SANT Domain 23
1.6 Transcription Factors in Soybean 23
1.7 Symbiosis Interaction and Nitrogen Fixation 25
1.8 Genes Implicated in Symbiotic Nitrogen Fixation 29
1.9 Nodulins 30
1.9.1 Methods for Identifying nodulins 32
1.10 Hypothesis and Objectives 33

CHAPTER TWO **34**
MATERIALS AND METHODS **34**
2.1 Plant Materials and Growth Conditions 34
2.2 Nucleic Acid Isolation 35
2.2.1 Isolation of Plant RNA 35
2.2.2 Isolation of Plant DNA 35
2.2.3 Isolation of Plasmid DNA from Bacteria 36
2.2.4 Isolation of DNA Fragments from Gels 36
2.3 General Molecular Methods 36
2.3.1 Restriction Digestion 36
2.3.2 Agarose Gel Electrophoresis of DNA 37
2.4 Cloning 37
2.4.1 Subcloning of DNA Fragments 37
2.4.2 TA Cloning 37
2.4.3 Transformation with Plasmid DNA 38
2.5 Polymerase Chain Reaction (PCR) Amplification 39
2.5.1 Oligonucleotide Primers for PCR 39
2.5.2 DNA PCR 44
2.5.3 RT- PCR 44
2.6 Sequencing Reaction 45
2.7 Southern Blot Analysis 45
2.7.1 Transfer of Nucleic acid DNA to Biotrans 45
2.7.2 Preparation of Radioactively Labeled Probe 46
2.7.3 Hybridization of DNA Membranes 46
2.8 Molecular Biology Data Bases, Analytical Tools and Software Programs 47
CHAPTER THREE **49**
RESULTS **49**
3.1 Approaches Used to Obtain cDNA and Genomic Sequences 49
3.2 Analysis of the 7/2 cDNA and Genomic Sequences 53
3.3 7/2 Conceptual Protein 63
3.4 Expression of 7/2 in Soybean Plants 79

3.4.1 Tissue Distribution 79
3.4.2 Partial Splicing 83
CHAPTER FOUR 87
DISCUSSION 87
4.1 Analysis of 7/2 DNA Sequence 87
4.2 7/2 Conceptual Protein 89
4.3 Expression of 7/2 96
4.3.1 Tissue Distribution 102
4.4 Detection and Analysis of Partially Spliced 7/2 mRNA 105
4.5 Conclusions and Future Work 106
REFRENCES 109
APPENDIX 128

LIST OF FIGURES

Figure 1: Schematic showing functional domains of prototypic Myb proteins..... 15
Figure 2: Early stage of nodule formation..... 27
Figure 3: Position of the primers on the gDNA and cDNA sequences in this study..... 42
Figure 4: The cDNA and gDNA structure of 7/2..... 51
Figure 5: Alignment of cDNA and gDNA sequences..... 56
Figure 6: Sequence of the 7/2 conceptual protein..... 64
Figure 7: Presence of single repeat Myb-DNA binding domain in the 7/2 protein..... 66
Figure 8: Alignment of 7/2 protein with CDPKS and PSRR..... 69
Figure 9: Alignment of 7/2 protein sequence with CSP1..... 73
Figure 10: The mRNA expression of 7/2 in various tissues (Set A1)..... 81
Figure 11: Schematic showing the structures of the 7/2 gene and the RT-PCR products. 85
Figure 12: 5' region of the 7/2 genomic DNA..... 100
Figure 1 in Appendix: The mRNA expression of 7/2 in various tissues (Set A2)..... 129
Figure 2 in Appendix: The mRNA expression of 7/2 in various tissues (Set B1) 131
Figure 3 in Appendix: The mRNA expression of 7/2 in various tissues (Set B2) 133
Figure 4 in Appendix: Pedigree relation ships giving ancestors of Maple Arrow and Resnik cultivars 135

LIST OF TABLES

Table 1: Structural features of conserved domains that are used to classify plant transcription factors. 11

Table 2: List of R2R3 myb genes. 21

Table 3: Primers used for the amplification of gDNA and cDNA in this study. 40

Table 4: The motifs on the 7/2 protein identified by PROSITE. 77

Table 1 in Appendix: Summary of mismatched positions on the gDNA and cDNA Sequences 137

LIST OF ABBREVIATIONS

A1E HM	medium containing 0.1% yeast extract, 3.9mM HEPES, 0.88mM Na_2HPO_4, 0.15%(w/v) arabinose, 1X salts, pH 6.8
BAC	Bacterial Artificial Chromosome
bHLH	basic helix-loop-helix
BLAST	Basic Local Alignment Search Tool
BLASTP	Basic Local Alignment Search Tool for Proteins
BLASTN	Basic Local Alignment Search Tool for Nucleic acids
CCGB	Center for Computational Genomics and Bioinformatics
Ccd	Cortical cell division
cDNA	Complementary DNA
CDPK	Calcium dependent protein kinase
CLUSTAL	Multiple sequence alignment program for DNA or proteins
cv	Cultivar
DEPC	Diethylpyrocarbonate
DTT	dithiothreitol
dpi	Days post inoculation
EDTA	Ethylene diamine tetra acetic acid
ENTREZ	A database giving access to DNA and protein sequences as well as the Medline references
EST	Expressed sequence taq
F	Forward primer
GBF	G-box binding factor
GTF(s)	General Transcription Factor(s)
H	Histone
Had	root hairs deformation
Hac	root hair curling
HAT(s)	Histone acetyl transferase(s)
HDAC(s)	Histone deacetylase(s)

HMG	high mobility group
Inf	Infection threads
kDa	Kilodaltons
MA	Mapple Arrow
MADS	A group of transcription factors with a conserved region first found in MCM1, AG,DEFA, and SRF.
myb	A DNA binding domain (from avian **my**elo**b**lastosis virus)
Myb	The protein domain enconded by myb.
McCDPK	*Mesembryanthemum crystallinum* CDPK
NCBI	National center for biotechnology information
NLS	Nuclear Localization Signal
ORF	Open reading frame
PLACE	Plant Cis-acting Regulatory DNA Elements data base
PROSITE	Data base of protein families and domains
PSORT	Computer program for Prediction of **P**rotein **Sort**ing Signals and Localization Sites in Amino Acid Sequences
Res	Resnik
R	Reverse primer
RACE	Rapid amplification of cDNA ends
SANT	DNA binding domain first found in **S**wi3, **A**da2, **N**-Cor,**T**FIIB transcription factors
SDS	Sodium dodecyl sulfate
SOC	A growth broth containing 2%(w/v) tryptone, 0.5%(w/v)yeast extract, 10mM NaCl, 2.5 mM KCl, pH 7.0
TAFs	TBP-associated factors
TBP	TATA-binding protein
TF	Transcription factors
TIGR	The Institute for Genomic Research
UV	Ultra Violet

WRKY	A conserved motif, found in a group of transcription factors which contain a zinc finger structure
2XYT	A growth broth containing 1.6% (w/v) tryptone, 1% (w/v) yeast extract, 0.5% (w/v) NaCl, 7mM KPi, pH 7.0

CHAPTER ONE
INTRODUCTION

The present study describes the characterization of a novel gene, first isolated from soybean (*Glycine max* L. Merril) nodules, whose structure suggests that it is a transcription factor. The objectives of this study are to describe the gene sequence and structure, and to measure the tissue-specific mRNA distribution. This chapter will first provide a general review of gene regulation with emphasis on plants and transcription factors containing the Myb domain(s), then provide an overview of the biological system- symbiotic nitrogen fixation- that was the genesis of the novel cDNA, and finally end with a summary of the hypothesis and objectives of the thesis.

1.1 Gene Regulation in Eukaryotes

In eukaryotes, gene expression is regulated at different levels. There are regulatory systems for the control of transcription, precursor-RNA processing, transport of the mature RNA out of the nucleus, translation of the mRNAs, degradation of the mature RNAs, and degradation of the protein products (Brown, 2002).

1.1.1 Gene Regulation at the Chromatin Level

In eukaryotic cells, nuclear DNA is packaged as chromatin which is a well defined complex composed of repeating subunits called nucleosomes. Nucleosomes consist of two left handed superhelical turns of DNA. Each turn contains 165 base pairs wounded around a protein core histone octamer that consist of two copies each of H2A, H2B, H3, and H4. Strings of nucleosomes are further folded into chromatin fibers

stabilized by binding of histone H1 to the outside of the complex. Besides histones, chromatin contains a heterogeneous group of proteins that bind to it, including DNA polymerase, regulatory proteins and others collectively referred to as non-histone proteins (Brown, 2002).

One of chromatin's major roles is to facilitate the packaging of the very large amount of DNA into the nucleus where it is found as euchromatin and heterochromatin. Euchromatin is less densely packed than the more highly compact heterochromatin. Heterochromatization leads to general suppression of gene activity, therefore, chromatin in the vicinity of the gene must be remodeled to allow for easy access of transcription factors and the recruitment of the RNA polymerase II transcription-initiation complex necessary for gene activation and transcription (Brown, 2002).

High nucleosome density is associated with transcriptionally inactive chromatin, whereas transcriptionally active chromatin exhibits a lower nucleosome density (Brown, 2002). Regulation of many nuclear genes in eukaryotic cells are facilitated by proteins that modify nucleosomal proteins and cause local changes in chromatin structure such as acetylation of histones by histone acetyl transferases (HATs) (Berger, 2002).

Histone acetylation at amino acids such as Lysine changes their charge and, as a result, the DNA associated with the nucleosome becomes more accessible to transcription factors (Anderson *et al.,* 2001). Hyperacetylated histones are mostly associated with activated genomic regions and deacetylation by histone deacetylases (HDACs) mainly results in repression and gene silencing (Grunstein, 1997; Lusser *et al.* 2001; Turner, 2000). Phosphorylation, methylation and ubiquitination are other forms of histone modification which have important role in gene regulation both in activation or

suppression. All of these modifications work through direct electrostatic effects or by creating altered surfaces on nucleosomes as recognition sites for the recruitment of transcription factors or regulatory complexes (Berger, 2002; Cheung *et al.*, 2000; Fischle *et al.*, 2003).

1.1.2 Gene Regulation at the Transcriptional Level

The regulation of transcription initiation is the key step in eukaryotic gene regulation. The first level of control occurs at the transcriptional level which regulates whether a gene is transcribed and the rate at which transcripts are produced (Brown, 2002). This process involves the regulated assembly of multiprotein complexes on enhancers and promoter elements (Kornberg, 1999). These upstream factors play a central role in the recruitment and activation of the components of the basal transcription machinery, which is assembled in a sequential order or as an RNA polymerase II holoenzyme at the site of the initiation of transcription (Conaway, 1997; Buratowski 1994). Assembly of a preinitiation complex and several general transcription factors (GTFs) on promoter DNA causes activation of protein coding genes (Tang *et al.*, 1996; Roeder, 1996). General transcription factors are a complex made up of the TATA-binding protein (TBP) and at least 12 TBP-associated factors or TAFs (Brown, 2002).

Different environmental conditions alter gene expression and transcription in eukaryotes. For example plants have the ability to change their pattern of gene expression in response to environmental stimuli such as temperature, water availability, light, and different concentration of ions or wounding (Hazen *et al.*, 2003).

1.1.3 Gene Regulation at the Post-Transcriptional Level

Regulation at this level is mainly focused on the processing of mRNAs, the transcripts of protein coding genes which are translated into proteins at the latter stages of genome expression. Processing includes both 5' end capping and the addition of poly "A" tails at the 3' end of the mRNA. Both of these modifications synergistically enhance the translational efficiency of the mRNA (Day and Tuite, 1998). Another important process is splicing of mRNA which is the removal of the introns and joining the exons to make a transcript for translation and protein synthesis (Brown, 2002). Alternative splicing combines different transcript splice junctions. In this way multiple mRNA species and proteins can be created from a single gene resulting in the expansion of the potential informational content of eukaryotic genomes. The best example is in humans with only about 32,000 genes, far less than expected, but with vast amount of functional proteins, suggesting a major role for alternative splicing (Modrek and Lee, 2002). For example up to 20 different polypeptide variants arise from alternative splicing in three regions of the FN gene in human fibronectin (Gutman and Kornblihtt, 1987).

Other examples of alternative splicing are HPR and FCA in plants. Hydroxypyruvate reductase (HPR) is a leaf peroxisomal enzyme that functions in the glycolate pathway of photorespiration in plants. Two alternative spliced mRNAs of this gene (HPR1 and HPR2) were found in pumpkin. Splicing is light regulated and the translated proteins are localized in leaf peroxisomes and the cytosol respectively (Mano *et al.*, 1999). Flower Controlling gene in Arabidopsis, FCA, has four differentially-spliced transcripts (Macknight *et al.*, 2002).

It has also emerged that the mRNA processing such as capping, splicing, and polyadenylation not only influence one another but are linked to transcription (Proudfoot *et al.*, 2002). Cotranscriptionality does not imply that transcription and pre-mRNA splicing are coupled but in a long gene for example some introns could be spliced out while transcription is still occurring, whereas others could be processed well after transcription has been completed. It is not known which introns follow each pattern and if a particular intron always follows the same pattern of processing (Kornblihtt *et al.*, 2004).

1.1.4 Gene Regulation at the Translational and Post-Translational Levels

Translational regulation is one of several mechanisms that control gene expression in prokaryotes and eukaryotes. In plants, protein synthesis takes place in the cytoplasm, the chloroplast and the mitochondria (Cohen and Mayfield, 1997). Translational regulation of nuclear encoded genes has been shown to be influenced by several developmental and environmental factors including light, embryo development, wounding, heat shock, and oxygen deprivation (Cohen and Mayfield, 1997).

There is evidence for post-translational control of the gene encoding a chloroplast drought-induced 32 kDa stress protein (CDSP32) during leaf development in *Solanum tuberosum.* The conclusion comes from the data showing that although there is no change in transcript level during leaf development in well-watered wild type and CDSP32 over expressed plants, the amount of CDSP32 protein significantly decreases with leaf age in both plant types (Broin *et al.*, 2002).

1.2 Gene Regulation in Plants

Within a plant cell some genes are expressed constitutively (Albert *et al.*, 1992) while others respond to specific stimuli (Adam *et al.*, 1994). Both patterns depend on the interaction of transcription factors with cis-acting DNA elements and/or with other transcription factors required for gene expression (Guilfoyle, 1997).

According to molecular phylogenetic analysis, plants, animals and fungi all diverged from a common ancestor during a short period of time ~ 1.5 billion years ago (Riechmann *et al.*, 2000). It would be expected that most of the transcription factor families would either be shared by three lineages, if they were present in the common ancestor, or specific to each lineage, if they arose independently following divergence (Riechmann *et al.*, 2000). Members of lineage specific families represent 45% of the *Arabidopsis thaliana* transcription factors, 47% in *Caenorhabditis elegans,* and 32% in the *Saccharomyces cerevisiae* (Riechmann *et al.*, 2000).

1.3 Transcription Factors

Transcription factors are proteins that play important and diverse roles in gene expression, including chromatin remodeling and the recruitment /stabilization of the polII transcription-initiation complex (Singh K, 1998). Transcription factors can be divided into a number of functional classes based upon their interaction with DNA and other proteins. A major class of transcription factors is activators and repressors-proteins that bind to specific DNA sequences (enhancers and silencers) and give rise to gene specific regulation (activation or repression respectively). Most Myb transcription factors are transcription activators such as C1 in maize which activates transcription of genes

encoding enzymes involved in the biosynthesis of phenylpropanoids (Paz-Ares *et al.*, 1987). AtMYB4 is a member of the family of transcription factors containing a Myb domain but functions as a transcriptional repressor by repressing the transcription of the gene encoding the phenylpropanoid enzyme cinnamate-4 hydroxylase and thus accumulating sinapoylmalate which is a UV-protectant compound (Hemm *et al.*, 2001).

A second class of transcription factors is co-activators and co-repressors. These proteins mediate the transcriptional effects of specific activators/repressors, usually by remodeling chromatin. Members of this group of transcription factors are not able to bind DNA on their own, but they can function in a promoter-specific manner as a result of protein-protein interactions with specific activators and repressors. Examples of this group of co-factors that modify nucleosomes are histone acetylases and deacetylases (Lemon and Tjian, 2000).

A third class comprises the general transcription factors such as TATA binding protein (TBP) and other associated proteins which are important components of the pol II transcription–initiation complex (Conaway, 1997). A fourth class is architectural transcription factors that are also involved in remodeling DNA, by inducing bends that facilitate the binding of other proteins to promoters such as high mobility group (HMG) of nonhistone chromatin proteins (Reeves and Beckerbauer, 2001).

Two important features in transcription factors are DNA binding domains and activation domains. The activation domain of most transcription factors has acidic regions such as those found in GAL4 in yeast (Mitchell and Tjian, 1989) and Opaque2 in maize (Schmitz *et al.*, 1997), but an activation domain can also be proline-rich like the domain found it the murine HODX-4 (Rambaldi *et al.*, 1994) or a glutamine-rich domain

as is seen in Psr1 in *A.thaliana* (Wykoff *et al.*, 1999). Some transcription factors contain multiple activation domains. RF2a in rice is an example with an acidic domain and a proline rich and glutamine rich domains (Dai *et al.*, 2003). The acidic domain in RF2a is essential for the activation of gene expression although it is the glutamine domain that binds to the TBP (Dai *et al.*, 2003).

1.3.1 Transcription Factors in Plants

Regulated expression of genes is fundamental to most biological phenomena such as development, differentiation, cell growth, and response to environmental signals in plants as it is in other eukaryotes. Transcriptional regulation of gene expression is commonly utilized as the essential regulatory mechanism and it is largely mediated through sequence-specific DNA binding proteins that recognize cis-acting elements located in the promoter and enhancer regions of the corresponding genes (Meshi *et al.*, 1995). Typical plant transcription factors consist of a DNA-binding region, an oligomerization site, a transcription regulation domain and a nuclear localization signal (Liu *et al.*, 1999). The functional domains of plant transcription factors are usually inferred by comparison of the amino acid sequences deduced from cDNA clones with their animal counterparts as the regulatory proteins are related and plants regulate their genes in the same way (Liu *et al.*, 1999).

Data from the *Arabidopsis* genome project suggest that more than 5 % of its genes encode transcription factors (Riechmann and Ratcliffe, 2000). Comparative analysis of transcription factors among eukaryotes shows that those factors that are found in animals and yeast usually are present in plants although there is also an evolutionary generation of diversity in regulation of transcription (Riechmann *et al.*, 2000). Myb transcription

factors with 130 members in A*rabidopsis* (Riechmann *et al.*, 2000; Jiang *et al.*, 2004) and MADS family with about 100 members in *Arabidopsis* and 70 members in rice (Nam *et al.*, 2004) are the two transcription factor families that have been more substantially amplified in plants compared to animal and yeast. There are also many transcription factor families that are thought to be unique to plants, including the AP2/EREBP (APETALA2/Ethylene Responsive Element Binding Protein) (Riechmann *et al.*, 2000; Chen *et al*, 2003), NAC (Nitrogen Assimilation Control) (Aida *et al.*, 1997), and WRKY families. The conservation of transcription factors among species is usually determined by data base searches using programs such as BLAST. Recent searches revealed WRKY group-I–like sequences in two nonphotosynthetic eukaryotes (slime mold and protist) implying that these genes originated with eukaryotes. There is still no evidence of their presence in yeast and animals (Ulker and Somssich, 2004).

1.3.2 Classification of Transcription Factors in Plants

Most known transcription factors can be grouped into families according to the characteristics of the structural motifs and their DNA binding domain(s) (Meshi and Iwabuchi 1995; Riechmann *et al.*, 2000). Classification of transcription factors into a family also depends on their structural features-they can be subdivided according to the number and spacing of conserved residues in the most similar domain (Liu *et al.*, 1999).

Beside Myb transcription factors one of the largest groups of transcription factors in *A. thaliana* is the basic helix-loop-helix (bHLH) group of transcription factors which comprises 133 genes that are involved in a variety of functions in the plant (Heim *et al.*, 2003).

Some of the main classes of the transcription factors in plants are summarized in Table 1. Information about specific members of each class can be found in the references given in Liu *et al.*1999.

Table 1: Structural features of conserved domains that are used to classify plant transcription factors.

Information taken from Liu *et al.*,1999.

Domain type	Structure
Zinc finger	Finger motifs each maintained by cysteine and/or histidine residues organized around a zinc ion
bZIP	A basic region and a leucine rich zipper like motif
Myb-related	A basic region with one to three imperfect repeats each forming a helix-turn-helix
Trihelix	Basic, acidic and proline/glutamine-rich motif which forms a trihelix DNA-binding domain
Homeodomain (HD)	Approximately 60 amino acid residues producing either three or four α-helices and an N-terminal arm
b/HLH	A cluster of basic amino acid residues adjacent to a helix-loop-helix motif
MADS	Approximately 57 amino acid residues that comprise a long α-helix and 2 β-strands
AT-hook motif	A consensus core sequence R(GP)RGRP with the RGR region containing the minor groove of A/T-rich DNA
HMG-box	L-shaped domain consisting of three α-helices with an angle of about 80° between the arms
AP2/EREBP	A 68 amino acid region with a conserved domain that constitutes a putative amphiphatic α-helix
B3	A 120 amino acid conserved sequence at the C-termini of VP1 and ABI3
ARF	A 350 amino acid region similar to B3 in sequence

1.4 Myb Family of Transcription Factors

Transcription factors are classified in structural families according to the presence of specific DNA-recognition motifs. One such family consists of proteins containing the Myb-homologous DNA-binding domain (Myb-domain), originally identified in the v-myb oncogene found in the avian **my**elo**b**lastosis virus (Klempnauer *et al.*, 1982).

Proteins containing the Myb DNA binding domain have since been found in many eukaryotes including animals, plants, fungi and slime molds. Plants contain large number of myb genes. For example 130 myb genes from Arabidopsis and 85 genes from *Oryza sativa* (Jiang *et al.*, 2004) and almost 80 R2R3-myb genes from Maize (Rabinowicz *et al.*, 1999) were identified with distinct functions. There is also a report of approximately 200 myb genes in cotton and sorghum (Cedroni *et al.*, 2003). 206 members from the myb superfamily were identified in soybean (Tian *et al.*, 2004). From the high number of myb genes in these plants it is concluded that the myb gene is widely distributed in plant kingdom and controls diverse functions in them.

Members of the Myb family normally possess a conserved domain consisting of three related helix-turn–helix motifs of approximately 50 amino acids each. This sequence is required for DNA binding and is conserved among animals, plants and yeasts (Martin and Paz-Ares, 1997; Kranz *et al.*, 2000; Romero *et al.*, 1998; Lipsick, 1996). The first plant myb-like gene which showed similarity to the vertebrate proto-oncoprotein c-myb, especially in the DNA binding domain region, was the C1 gene isolated from *Zea mays*. The C1 protein regulates the synthesis of anthocyanin pigments in the aleurone of *Z. mays* by activating the transcription of several genes encoding the enzymes involved in the biosynthesis of this pigment (Paz-Ares *et al.*, 1987).

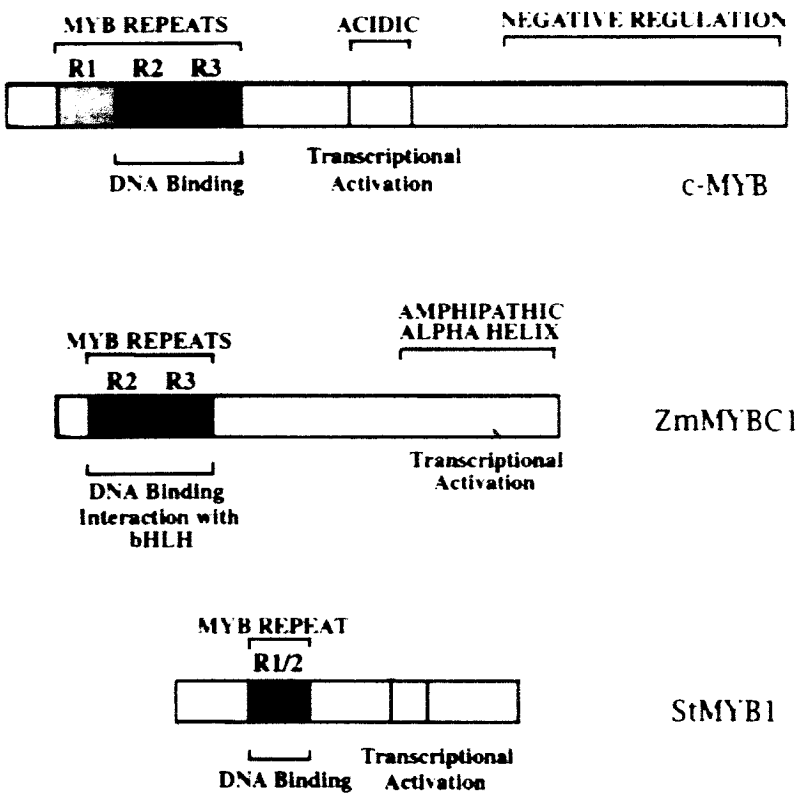
1.4.1 Myb Transcription Factor Structure in Plants

The DNA binding domain of Myb proteins from animals (c-myb) generally contains three repeats (R1, R2, and R3) (Ogata *et al.*, 1992; Gabrielsen *et al.*, 1991). While proteins containing three Myb repeats have also been found in plants (Braun and Grotewold, 1999; Kranz *et al.*, 2000), the Myb domain of plant proteins usually consists of two repeats (R2, R3) which are most similar to the second and third of the three repeats in animal Myb DNA binding domain (Williams and Grotewold, 1997). Each of these repeats encodes a helix-turn-helix structure with three regularly spaced tryptophan residues involved in DNA binding. The repeat most proximal to the N-terminus (R1) in c-myb does not affect DNA-binding specificity and is missing in oncogenic variants of c-myb such as v-myb and in the known plant R2R3-Myb proteins (Romero *et al.*, 1998). In some cases it can contain only one of these repeats or all three repeats (Martin and Paz-Ares, 1997; Jin and Martin 1999; Kranz *et al.*, 2000). Most Myb proteins have an activation domain C-terminal to the DNA binding domain (Weston, 1998). Due to the flexible structural determinants for activation domains this region of Myb proteins are not conserved (Jin and Martin, 1999).

Figure 1 shows the structural features of this family of transcription factors containing a DNA binding domain with one, two, or three repeats (R1, R2, R3) and an activation domain C-terminus to it.

Figure 1: Schematic showing functional domains of prototypic Myb proteins.

DNA binding domain at N-terminal and activation domain at the C-terminal of proteins from a three Myb repeat (c-MYB), two Myb repeat (ZmMYBC1), and one Myb repeat (StMYB1) protein are shown (Jin and Martin, 1999).


MYB REPEATS
ACIDIC
NEGATIVE REGULATION
R1
R2
R3
DNA Binding
Transcriptional Activation
c-MYB
AMPHIPATHIC ALPHA HELIX
MYB REPEATS
R2
R3
Transcriptional Activation
DNA Binding Interaction with bHLH
ZmMYBC1
MYB REPEAT
R1/2
DNA Binding
Transcriptional Activation
StMYB1

1.4.2 Single Myb-repeat Proteins

The DNA binding domain in most plants contains only two Myb repeats but proteins with three and even only one repeat have been identified (Martin and Paz-Ares 1997; Jin and Martin 1999; Kranz *et al.,* 2000; Braun and Grotewold, 1999). Almost all myb genes in *Arabidopsis* belong to the R2R3 group and only five R1R2R3 repeats exist in this plant (Riechmann *et al.*, 2000). Myb proteins with one repeat or sometimes a partial repeat are fairly divergent in plants. A few examples of single repeat Myb proteins are mentioned below.

Solanum tuberosum contains a novel myb gene, mybSt1, which encodes only one repeat of 61 amino acids (Baranowskij *et al.*, 1994). A series of deletions of N- and C-terminal regions of MybSt1 showed that the Myb-related motif is required for DNA binding. The MybSt1 protein sequence contains an acidic central region and a proline rich C-terminal domain. It has been shown that many regulatory proteins involved in transcriptional activation contain such structural motifs (Ptashne, 1988). *A. thaliana* AtmybL2 encodes a potential transcription factor of the Myb family with a Myb domain consisting of a single repeat in the amino-terminal half and a proline rich region involved in transactivation found in the C-terminal part of the protein (Kirik and Baumlein, 1996). Two other single domain Myb proteins from *Arabidopsis* are Circadian Clock Associated (CCA1) and Late Elongated Hypocotyl (LHY). They are believed to operate as oscillators similar to or part of a circadian clock mechanism governing flowering, leaf movements, photosynthetic gene expression and hypocotyl growth (Jin and Martin, 1999; Carre and Kim, 2002). CAPRICE (CPC), a gene with a role in root hair formation in *Arabidopsis*, encodes a single-domain Myb protein (Jin and Martin, 1999; Wada *et al.*,

1997). Three single repeat Myb transcription factors (OsMYBS1, OsMYBS2, and OsMYBS3) exist in rice that interact with the promoter of α-amylase gene and affects gibberellin's and sugar-regulated α-amylase gene expression (Lu *et al.*, 2002). ZmMybst1 from *Zea mays* is a single Myb-repeat protein which has a role in endosperm development (Mercy *et al.*, 2003). ZmMRP-1 is another myb-like gene that encodes a single-Myb domain protein in maize and is proposed to be involved in regulating transfer cell differentiation (Gomez *et al.*, 2002). Two more proteins from *Arabidopsis* were identified with a single Myb-like DNA binding domain at their N-terminus and a histone H1/H5-like DNA binding domain in the middle of the protein sequence. These proteins have affinity to telomeric DNA sequence (Schrumpfova *et al.*, 2004). Single Myb-domain proteins might bind DNA in a manner similar to homeodomain proteins and as dimers (either hetero- or homo-dimers), which may have an important role for their modes of action and biological functions (Jin and- Martin, 1999).

1.4.3 Myb Transcription Factor Function in Plants

myb genes have expanded and diversified functions and regulate many different aspects of metabolism and development in plants which are different from their animal counterparts (Jin and Martin, 1999; Martin and Paz-Ares, 1997). Myb protein functions in animals are mostly associated with the control of cell proliferation, prevention of apoptosis, and commitment to development (Graf, 1992; Lipsick, 1996). Most Myb proteins are presumed to be transcriptional activators with activation domains in the region C-terminal to the DNA-binding domain (Weston, 1998). Most members of the plant R2R3-Myb family with known functions have been implicated in regulation of the synthesis of different phenylpropanoids (Romero *et al.*, 1998), some regulate cellular

morphogenesis such as differentiation of hair cells in leaf and stems such as GL1 in *Arabidopsis* (Oppenheimer *et al.*, 1991) and development of the conical form of petal epidermal cells like phMYB1 from *Petunia hybrida* and MIXTA from *Antirrhinum majus* (Beverley *et al.*, 1998). Another role for plant Myb proteins is in the signal transduction pathways responding to plant growth regulators such as GAMYB from barley in response to giberellic acid (Gubler *et al.*, 1995) or AtMYB2 in *Arabidopsis* in response to abscisic acid (Urao *et al*, 1993). AtMYB30 in *A. thaliana* is a single copy gene maximally expressed during the hypersensitive response (Daniel *et al.*, 1999). NtMYB2 in tobacco (*Nicotiana tabacum*) is induced by wounding and it is involved in the expression of defense related genes and the activation of retrotransposons following stress (Sugimoto *et al.*, 2000). LjMYB101 from *Lotus japonicus* and GmMYB101 from *G. max* have roles in regulation of flavonoid biosynthesis in response to nitrate starvation (Miyake *et al.*, 2003). The expression of AtMYB102 transcription factor gene in *Arabidopsis* depends on and integrates signals derived from wounding and osmotic stress (Denekamp and Smeekens, 2003). Epidermal cell differentiation and root hair formation in *Arabidopsis* is determined by CAPRICE (CPC) gene which encodes a protein with a Myb-DNA binding domain (Wada *et al.*, 1997). AtMYB61 in Arabidopsis also regulates seed coat development (Panfield *et al.*, 2001). The Late Elongated Hypocotyl (LHY) and Circadian Clock Associated (CCA1) genes encode closely related Myb transcription factors which regulate circadian rhythms in *A. thaliana* (Carre and Kim 2002). The function of three repeat myb genes recorded in plants is more similar to their animal counterparts and involves controlling of the cell cycle (Stracke *et al.*, 2001).

Table 2 lists the biological functions of some R2R3-myb genes in different plants. These examples demonstrate the wide spread distribution of Myb transcription factors and the variety of their function, concluding the possibility of the existence of more genes with Myb domain in plants.

Table 2: List of R2R3 myb genes.

This table shows some of the plant R2R3 myb genes for which a function has been assigned (Jin and Martin, 1999).

R2R3 Myb proteins	Biological function	Species
	Phenylpropanoid metabolism	
ZmMYBC1	Anthocyanin	*Zea mays*
PhMYBAN2	Anthocyanin	*Petunia hybrida*
PhMYB3	Anthocyanin	*Petunia hybrida*
AmMYB305,340	Anthocyanin and flovonol	*Antirrhinum majus*
PsMYB26	Phenylpropanoid regulation	*Pisum sativum*
ZmMYBP	Phlobaphene	*Zea mays*
AmMYB308,330	Phenolic acid	*Antirrhinum majus*
	Development	
AtMYBGL1	Trichome development	*Arabidopsis thaliana*
AmMYBMIXTA	Conical cell development	*Antirrhinum majus*
PhMYB1	Conical cell development	*Petunia hybrida*
CotMYBA	Trichome development	*Gossypium hirsutum*
AmMYBPHAN	Dorsoventral determination &growth	*Antirrhinum majus*
ZmMYBRS2	PHAN-like, repress knox expression	*Zea mays*
AtMYB13	Shoot morphogenesis	*Arabidopsis thaliana*
AtMYB103	Expressed in developing anthers	*Arabidopsis thaliana*
	Signal transduction	
GAMYB	Gibberellin response	*Ilordeum vulgare*
AtMYB2	Dehydration and ABA regulation	*Arabidopsis thaliana*
ATR1	Tryptophan biosynthesis	*Arabidopsis thaliana*
Cpm5,Cpm7,Cpm10	Dehydration & ABA response	*Craterostigma plantagineum*
	Plant disease resistance	
NtMYB1	TMV,SA-inducible	*Nicotiana tabacum*
	Cell division	
AtCDC5	Cell cycle regulation	*Arabidopsis thaliana*

1.5 SANT Domain

Several different transcriptional regulators have been identified that contain more distantly related Myb repeats. These Myb-repeat containing proteins can be grouped into different families, such as the telobox family (Biluad *et al.*, 1996), the transcription terminator family (Reeder and Lang, 1997) and SANT domain family of transcriptional regulators (Asland *et al.*, 1996). SANT domain (Swi3, Ada2, N-Cor,TFIIB) is a putative DNA binding domain conserved in SWI3 (switching-defective protein3) and ADA2 (adaptor 2) transcriptional activation complexes, the transcriptional corepressor N-CoR (Nuclear receptor co-repressor) and the transcription factor TFIIB (Asland *et al.*, 1996). The SANT domain is a novel motif that was identified based on its sequence similarity to the DNA-binding domain of Myb related proteins containing helix-turn-helix structure (Asland *et al.*, 1996). Sequence alignments as well as secondary structure predictions, indicated that the SANT domain also consists of three α-helices containing aromatic residues (Asland *et al.*, 1996). The presence of the SANT domain in transcriptional co-factor initiation complexes and proteins with known functions suggest that the SANT domain is involved in transcriptional regulation (Asland *et al.*, 1996). It has been shown that SANT domain is crucial for the function of SANT-containing subunits of yeast chromatin remodeling complexes (Boyer *et al.*, 2002) and it also is a histone tail binding module (Boyer *et al.*, 2004).

1.6 Transcription Factors in Soybean

Different transcription factors have been identified in plants especially in *A. thaliana* (Riechmann, 2000). Through comparisons of sequences, similarities can be

recognized that allow the definition of different groups of transcription factors in other plants. In soybean many transcription factors or putative transcription factors have been identified which are involved in regulation of different aspects of plant growth and development or are responsible for the response to different stimuli and stress conditions. According to a recent search on soybean ESTs (Expressed Sequence Tag) and some BAC (Bacterial Artificial Chromosome) sequences from Genbank, and using the TFs from *Arabidopsis* protein database as query in a BLASTP search the total number of transcription factors in soybean has been estimated as 1,322 genes which is similar to 1,533 transcription factors in *Arabidopsis* and 1306 TFs in rice (Tian *et al.*, 2004). Some of the characterized transcription factors in soybean are discussed below.

A zinc finger protein from soybean, SCOF-1, is specifically induced by low temperature and abscisic acid (ABA). This protein functions as a positive regulator of COR gene expression which enhances cold tolerance of plants (Kim *et al.*, 2001). GmGT2 is a member of GT2 family of transcription factors from soybean (O'Grady, 2001). GmGT-2 message levels are down-regulated by light in a phytochrome-dependent manner. This result, when combined with previous data, implies the possible convergence of phytochrome and auxin signaling pathways (O'Grady, 2001).

GmMYB101 is a member of plant R2R3-Myb transcription factors isolated from a soybean cDNA library during nitrate starvation. Due to the homologus regions between the promoter of this gene and *GlnI* (Glutamine synthase) and upregulation of both genes as well as chalcone synthase (CHS) in response to nitrate starvation it is assumed that MYB101 has a possible role in the regulation of (iso)flavonoid biosynthesis in response to nitrate starvation (Miyake *et al.*, 2003). TFIIB, a member of the basal preinitiation

complex of transcription factors from soybean and *Arabidopsis* shows high homology and contains the same structural motifs and organization as seen in other eukaryotes and *Archaebacteria* (Baldwin and Gurley, 1996). Heat Shock transcription factors (HSF) are induced under thermal stress in soybean and result in the induction of DNA binding activity to the heat shock elements (HSEs) in promoters of heat shock genes leading to their transcription (Czarnecka-Verner *et al.*, 1995). Gmpzf is a protein isolated from soybean having a RING-finger domain, suggesting that the protein has a regulatory function (Schauser, 1995). SGBF-1 and SGBF-2 (soybean G-box binding factors) were isolated from soybean that interact with a G-box sequence of an auxin-responsive gene (Hong *et al.*, 1995). GBF transcription factors have a C-terminal basic/leucine zipper DNA binding domain and an N-terminal proline rich transcription activation domain (Schindler *et al.*, 1992).

1.7 Symbiosis Interaction and Nitrogen Fixation

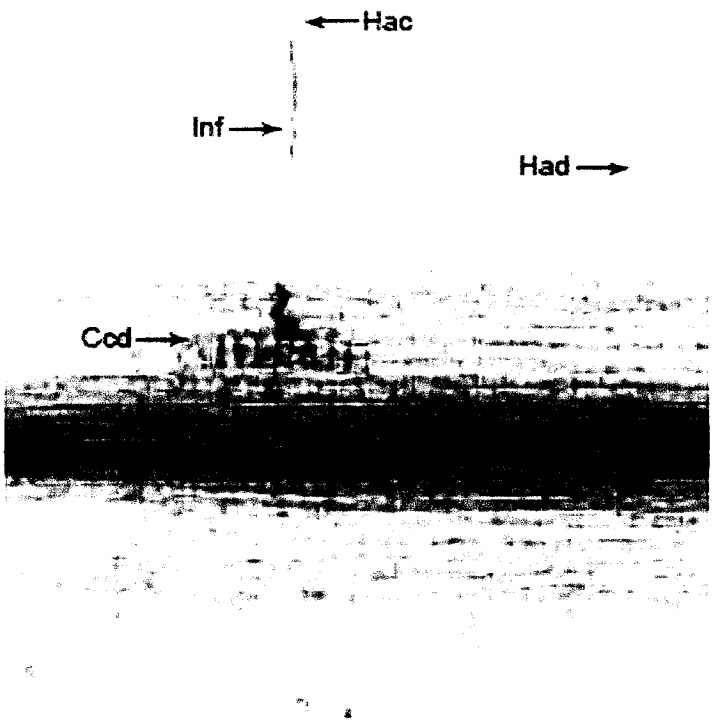
Under nitrogen-limiting conditions, bacteria from the family *Rhizobiaceae* are able to establish a symbiosis with leguminous plants that is capable of the reduction of atmospheric N_2 to ammonia, thus providing a source of fixed nitrogen for the plant to make its amino acids, proteins, and other essential nitrogenous compounds (Hirsch, 1992; Kuzma *et al.*, 1999; Crespi and Galvez, 2000). In these symbioses a new organ called the nodule is formed in which the infecting bacteria, now differentiated into bacteroids, make nitrogenase- the enzyme that catalyzes nitrogen fixation. The host plant supplies the carbon necessary for energy production (Hirsch, 1992; Gualtier and Bisseling, 2000; Schultze and Kondorosi, 1998).

The development of the nitrogen-fixing root nodule is the conclusion of a complex process facilitated by the exchange of information between the symbionts. Flavonoids released by a susceptible plant act as chemo attractants for a suitable Rhizobium (Long, 1989; Schultze and Kondorosi, 1998). The soybean root nodule is a product of the symbiosis between soybean and *Bradyrhizobium japonicum* (Kosslak *et al.*, 1987). The signal molecule in soybean is the isoflavone genistein which is secreted by soybean roots (Zhang *et al.*, 2002).

Bacteria attach to the root hairs and cause curling of the hair tip, trapping the bacteria which then are carried to the root cortex through an infection thread. A local lesion of the root hair cell wall is formed by hydrolysis at the point of adhesion and allows the entrance of the bacteria. (VanRhijn and Vanderleyden, 1995; Albrecht *et al.*, 1999). The infection thread is a plant derived structure originating from plasma membrane invagination accompanied by external deposition of cell wall material. As the tubular infection thread forms, root cortical cells are reactivated and form a nodule primordium. The infection thread grows toward this primordium, penetrates the cells and releases bacteria where they remain separated from the cytoplasm by a membrane called the peribacteroid membrane (Mylona *et al.*, 1995; Cullimore, 2001). Figure 2 shows the early stages of the symbiotic interaction between *Medicago sativa* roots and *Sinorhizobium meliloti* leading to nodule formation (Cullimore, 2001).

Figure 2: Early stage of nodule formation.

This figure shows nodule formation through symbiotic interaction between *Medicago sativa* roots and *Sinorhizobium meliloti* (adapted from Cullimore, 2001). Rhizobial inoculation causes deformation of root hairs (Had), root hair curling (Hac) and invagination of the root hair cell wall, leading to the initiation of infection threads (Inf) and cortical cell division (Ccd).


Hac
Inf
Had
Cod

1.8 Genes Implicated in Symbiotic Nitrogen Fixation

Nodule development involves the expression of nodule-specific plant genes called nodulin genes and bacterial nod genes. There are specific nod genes in bacteria that are not functionally or structurally conserved among Rhizobia, which are called host specific nodulation genes (*hsn*). These are necessary for the nodulation of a particular host plant and mutations in these genes can alter the host range (VanRhijn, 1995). Common nod genes like *nodABC* have been detected in all Rhizobia. These genes are responsible for the synthesis of the backbone of Nod factors that are modified further by *hsn* genes and cause expression of nodulin genes in the plant. n*odA, nodB* or *nodC* mutants are completely defective in nodule formation, confirming their vital role in nodule development (Nap and Bisseling, 1990). The bacterial gene *nodD* is the only nod gene expressed in both the free living and symbiotic state of Rhizobium. It acts as an activator for nod gene expression (Nap and Bisseling, 1990).

Nodulins are plant genes that express at different stages of nodule development due to the infection. The early nodulin genes in plants encode products that are expressed before the onset of nitrogen fixation. These proteins are involved in infection and nodule development and many of them are proline-rich like cell wall components (Nap and Bisseling, 1990). Late nodulin genes are expressed after complete development of the nodules and produce proteins that are involved in the interaction with the endosymbiont within the nodule for nitrogen fixation and in the metabolic specialization of the nodule (Nap and Bisseling, 1990). Very late nodulins are involved in plant senescence (Chan, 1995).

1.9 Nodulins

Plant genes that are specifically activated by the lipochitooligosaccharide signal molecules (Nod factors) from Rhizobia in legume hosts are referred to as nodulins which are important for the nodule growth and development and are nodule specific or nodule enhanced. Recent studies has revealed a number of homologues of nodulin genes in non-legumes such as *enod40* in tobacco (Sande *et al.*, 1996) and rice (Kouchi *et al.*, 1999), suggesting that nodulin genes have arisen as a result of the recruitment of pre-existing non-symbiotic genes which might have roles common to all plants.

Many nodulins in legumes have been identified by screening of cDNA libraries produced from the nodules in different legumes, e.g., *enod2.* This nodulin was originally identified as a cDNA clone from soybean nodule library and was then identified in other legumes like alfalfa and pea as well (Van de Wiel *et al.*, 1990). *Enod2* is an early nodulin and is expressed in the inner cortex of legume nodules. Since this region functions as a barrier to O_2 diffusion, it has been proposed that this protein functions in the control of nodule permeability to O_2. However the level of ENOD2 mRNA and protein do not differ among alfalfa nodules grown at different O_2 concentrations (Wycoff *et al.*, 1998).

Often nodulin function can be inferred from sequence information. Two early nodulin genes *enod5* and *enod12* were isolated from a pea nodule cDNA library by differential screening. Sequence analysis shows both nodulins are proline rich proteins. Because many proline rich proteins are hydroxylated and localized to the cell wall, it is likely that these proteins are cell wall components. The ENOD5 protein is also rich in the amino acids alanine, glycine, and serine as are arabinogalactan proteins. As

arabinogalactans are known to be components of the plasma membrane, thus the ENOD5 protein may also be part of the plasma membrane of the infection thread (Scheres *et al.*, 1990). By *in situ* hybridization with antisense RNA probes it was shown that the *enod5* gene is only expressed in cells containing a growing infection thread which shows the very early stages of development. The *enod12* gene is expressed not only in root hairs and cortical cells that contain an infection thread but in cells that are several layers in front of the growing infection thread and are undergoing the morphological changes that precede penetration by infection thread. This protein may be part of the new cortex cell wall that prepares for infection thread passage and may also be a component of the infection thread itself (Nap and Bisseling, 1990). *enod40*, also isolated by differential screening of a cDNA library, is highly expressed in nodule primordial cells. It encodes a small oligopeptide which stimulates cortical cell division and controls nodule initiation by changing the phytohormone balance (Cohn *et al.*, 1998). It was shown that ENOD40 isolated from tobacco (non-legume plant) can modulate the action of auxin and can be considered as a plant growth regulator that alters phytohormone responses and occurs in legumes as well as non-legumes (Sande *et al.*, 1996).

Medicago truncatula rip1 (Rhizobium induced peroxidase) expresses a peroxidase which was identified by subtractive hybridization PCR (Cook *et al.*, 1995). Since peroxidases can limit the growth of dividing carrot cells, by limiting cell expansion, it is possible that *rip1* facilitates the development or maintenance of nodule perimordia by restricting cell expansion (Peng *et al.*, 1996). Leghemoglobin is one of the late nodulins produced in the nodules with high affinity for oxygen. It binds O_2 and therefore protects nitrogenase from damage by free oxygen and controls the

concentration of free oxygen in nodules and its diffusion to the bacteriods (Mylona *et al.*, 1995). For most nodulins no apparent function can be inferred, e.g., the soybean gene GmNod53b that encodes a 53-kDa protein. The expression of this gene coincides with the onset of nitrogen fixation, so it is a late nodulin. No specific function for it has been identified. (Winzer *et al.*,1999).

1.9.1 Methods for Identifying Nodulins

Several standard methods have been employed for the identification of nodulins in a wide variety of legumes. Some commonly used approaches include

i. cDNA library screening: based on the differential screening of cDNA clones with radiolabeled probes from tissues at different stages of plant growth (Chan, 1995).

ii. Transposon mutagenesis: the basis of this method is to isolate nodule function mutations due to integration of a transposable element into the gene. The gene can be identified by a probe specific to the transposable element (Schauser *et al.*, 1999) and the mutant and wild type plant can be compared.

iii. Express Sequence Taq (EST) sequencing: sequences from the 5' or 3' ends of cDNA can be used to search databases and from the resulting matches, functions can be inferred (Gyorgyey *et al.*,2000).

iv. Differential display RT-PCR: mRNA of a cell is extracted, and by RT-PCR small cDNAs are made. These fragments are separated on denaturing polyacrylamide gels. The patterns of the amplified cDNA products of different mRNA samples can be compared side by side and differentially expressed cDNAs can be identified (Goormachtig *et al.*, 1995; Jimenez-Zurdo *et al.*, 2000).

v. Gene homology: a gene is used to identify the homologous gene from other species. *Enod40* from soybean was used to identify *OsENOD40*, the rice (*Orysa sativa*) homologue (Kouchi *et al.*, 1999).

1.10 Hypothesis and Objectives

During differential screening of a soybean (cv. Maple Arrow) cDNA library prepared from nodule RNA at the senescence stage, a novel gene named 7/2 was identified. Since the gene was first identified as being expressed in nodules it was hypothesized that 7/2 is a late nodulin involved in nitrogen fixation or nodule senescence. To investigate the expression of this gene and to infer its function, we set the following four objectives:

1) To determine the structure of the 7/2 gene by analysis of the sequences of the genomic and cDNA clones;

2) To use Bioinformatics tools to suggest probable functions for the 7/2 protein;

3) To determine the tissue-specific distribution of 7/2 mRNA by RT-PCR;

4) To monitor the time course of 7/2 mRNA expression during nodulation by RT-PCR.

CHAPTER TWO

MATERIALS AND METHODS

2.1 Plant Materials and Growth Conditions

Soybean seeds (Glycine max L. Merrill cv. Maple Arrow) were obtained from Dr. Elroy Cober, Eastern Cereal and Oilseed Research Center, Agriculture and Agri-Foods Canada, Ottawa.

Seeds were surface sterilized by twice washing with a 25% solution of commercial bleach for 5 minutes each. The seeds were then rinsed with distilled water and dispersed onto moistened Whatman 3MM filter paper that had been cut to fit within sterile, 24cm x 24cm culture dishes. The seeds were germinated in the dark at room temperature for 5 to 6 days. The filter paper was moistened daily. Seeds that did not germinate or had signs of fungal infection were discarded.

Healthy seedlings were planted in 6 inch pots containing Vermiculite (day 0). Each seedling was inoculated with 4 ml of a late log phase culture of *Bradyrhizobium japonicum* strain 61A76 grown in A1E HM medium (0.1% yeast extract, 3.9mM HEPES, 0.88mM Na_2HPO_4, 0.15%(w/v) arabinose, 1X salts, pH 6.8, Kuykendall and Weber, 1978). The plants were grown in a Conviron growth chamber with a 16 hour light/ 8 hour dark photoperiod. During the daylight hours the temperature was 25° C, whereas during the dark hours the temperature was reduced to 20° C. Up to 12 days post inoculation (dpi) with *B. japonicum*, the plants were watered daily with 1/8 strength nitrogen-free Hoagland's solution (Hoagland and Arnon, 1950), supplemented with minimal nitrogen (0.1 mM NH_4NO_3). After 12 dpi, the plants were watered once daily with a 1/4 strength

nitrogen-free Hoagland's solution, supplemented with 0.1 mM NH_4NO_3. From 28 dpi to harvest, plants were watered with 1/4 strength nitrogen-free Hoagland's solution without added nitrogen as by this stage symbiosis supplies sufficient nitrogen.

Plant tissues were collected at different time points by hand, frozen immediately in liquid nitrogen and then stored at -80°C for later use.

2.2 Nucleic Acid Isolation

2.2.1 Isolation of Plant RNA

Total plant RNA was extracted from frozen plant tissues collected at different stages of development including 24 dpi roots, 24 dpi stems, 24 dpi young leaves, 24dpi old leaves, 65dpi senescent leaves, 36 dpi flowers, 65 dpi pods, 7dpi cotyledons and 15, 20, 24, 40 dpi nodules. With the aid of a mortar and pestle, tissues were ground in liquid nitrogen into a fine powder. Total RNA was extracted by the guanidinum thiocyanate procedure described by Chomcyzynski and Sacchi (1987) and later modified by Chomcyzynski and Mackey (1995).

2.2.2 Isolation of Plant DNA

Soybean sprouts were ground to a fine powder under liquid nitrogen using a mortar and pestle. High molecular weight soybean DNA was extracted using the DNeasy Plant Mini Kit (Qiagen) as described in the kit's manual. A GeneQuant spectrophotometer (Pharmacia) was used to estimate the concentration and purity of the DNA.

2.2.3 Isolation of Plasmid DNA from Bacteria

Plasmid DNA was isolated from overnight cultures (12-16 hrs) of *Eschericia coli* cells grown in SOC broth (2%(w/v) tryptone, 0.5%(w/v) yeast extract, 10mM NaCl, 2.5 mM KCl, pH 7.0) or 2XYT broth (1.6%(w/v) tryptone, 1%(w/v) yeast extract, 0.5%(w/v) NaCl , 7mM KPi, pH 7.0) supplemented with the selective antibiotic. Depending on the purity required, plasmid DNA was isolated by the Easyprep Boiling Method (Berghammer and Auer, 1993) or, if purer DNA was needed, by the Promega "Wizard" miniprep method according to the manufacturer instructions. The concentration and purity of the DNA were measured by spectrophotometry.

2.2.4 Isolation of DNA Fragments from Gels

DNA fragments obtained by either PCR amplification or from restriction endonuclease digestion were separated by electrophoresis through 0.8 - 1.5 % (w/v) agarose gels. The fragment of interest was isolated from the gel with a scalpel and the DNA was isolated with the QIAQuick Gel Extraction kit (Qiagen) according to the manufacturer's instructions. The isolated DNA was used for cloning or sequencing or labeling probes.

2.3 General Molecular Methods

2.3.1 Restriction Digestion

Restriction digestions were normally done using 10 U of restriction enzyme per 1 μg DNA in the amount of One –Phor-All (OPA) buffer (Pharmacia) appropriate to the enzyme. Digestion at 37°C was for 1 hour to overnight as required. Restricted fragments

were then analyzed on agarose gels or used for ligation. Restriction enzymes were purchased from Invitrogen, e.g., BamHI, BglII, EcoRI, HindIII, PstI, or MBI e.g., PaeI.

2.3.2 Agarose Gel Electrophoresis of DNA

DNA fragments were separated by electrophoresis through horizontal agarose gels containing 0.4 μg/ml ethidium bromide in 1X TBE buffer (Sambrook *et al.*, 1989). The concentration of the gels varied from 0.8-1.5 % (w/v) depending on the size range of the fragments to be resolved. Electrophoresis was conducted under constant voltage, between 40 to100 V. Lambda DNA digested with HindIII and pPhiX174 DNA digested with HaeIII were used as size markers.

2.4 Cloning

2.4.1 Subcloning of DNA Fragments

In a typical experiment, 500 ng of restricted fragments were subcloned into 100ng of restricted vector using 0.5μl T4 DNA ligase (1 U/μl, from Invitrogen) and 1X ligase buffer (Invitrogen). The vectors used were pUC8, pUC18, pUC19 (Amersham Biotech) and pGEM-4Z (Promega). Recombinant plasmids were detected by PCR amplification of the insert using M13 forward and reverse primers that flank the multiple cloning site of the vector.

2.4.2 TA Cloning

Products of PCR amplification of genomic DNA or cDNA using specific primers were ligated into pGEM-T Easy vector (Promega). Ligation mixtures were composed of

3 µl PCR product, 5 µl of 2X rapid ligation buffer, 1µl of pGEM-T Easy vector and 1µl (3U) of T4 DNA ligase. The mixture was left at room temperature over night.

2.4.3 Transformation with Plasmid DNA

Plasmids or ligation products were transformed into *E. coli* strain DH5αmerF' (Sambrook *et al.* 1989) by the method of Hanahan (1985). Briefly cells were made competent using Standard Transformation Buffer (STB: 100mM KCl, 45mM $MnCl_2.4H_2O$, 10mM $CaCl_2.2\ H_2O$, 3mM Hexamine $CoCl_3$ and 10mM potassium-morpholinoethane sufonate, pH 6.2) and DnD Solution (1M DTT, 90% (v/v) DMSO, 10mM KAc, pH 7.5) with intervals of suspension and cooling on ice. Following addition of DNA, the cells were subjected to a heat shock at 37°C for 5 min and quick cooling on ice. SOC medium was added to the cells which were shaken at 37°C, for 30 min. to allow for expression of antibiotic resistance. The cells were then spread on selective LB plates containing ampicillin (100mg/L) and Xgal (2% w/v). Transformation competence of the cells was determined by a control transformation with 10 ng of pUC19 plasmid.

E. coli competent cells from Invitrogen were also used as recommended by the manufacturer. Aliquots of 50 µl of transformed cells were stored at – 80°C until used.

2.5 Polymerase Chain Reaction (PCR) Amplification

2.5.1 Oligonucleotide Primers for PCR

Primers used in this study are listed in Table 3. The primers were synthesized by IDT (Integrated Device Technology), Genosys, and Invitrogen. The position of each primer on the final gene map is indicated in the table and Figure 3 shows their positions on the map.

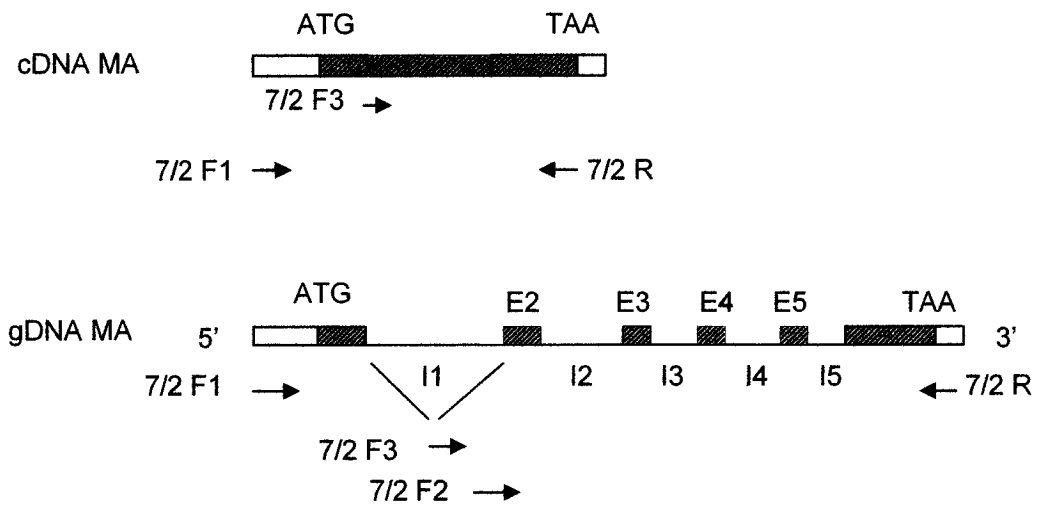
Table 3: Primers used for the amplification of gDNA and cDNA in this study.

The position of each forward primer and reverse primer on the gene as well as the sequences are indicated in this table. The forward primer 7/2 F 3 spans intron 1 and is complementary to exon 1 and exon 2 sequences as indicated in the brackets.

Primer	Sequence	Position on the gene
7/2 F 1	5' CAATCAGAATCAGAGTGTGTTGG 3' (23mer)	1478-1500
7/2 F 2	5' TTCAGAGGCGACTCCCAAGTC 3' (21mer)	2434-2454
7/2 F 3	5' TGGCCCTGATAAGGCGACT 3' (19mer)	(1887-1897) (2439-2446)
7/2 R	5' GAGATAACATGCTGTCCTGATACGC 3' (25mer)	4179-4203

Figure 3: Position of the primers on the gDNA and cDNA sequences in this study.

The schematics show the gDNA and cDNA from cv. Maple Arrow and the position of the primers on each. The number of exons and introns is indicated on top and bottom of the gDNA respectively. The hatched boxes indicate coding region.


ATG
TAA
cDNA MA
7/2 F3
7/2 F1
7/2 R
ATG
E2
E3
E4
E5
TAA
gDNA MA
5'
3'
7/2 F1
I1
I2
I3
I4
I5
7/2 R
7/2 F3
7/2 F2

2.5.2 DNA PCR

PCR amplification of plasmid DNA or gDNA was performed using a Perkin Elmer GeneAmp PCR thermocycler. A typical reaction of 20µl or 50µl contained 1X Taq reaction buffer (Invitrogen), 2mM $MgCl_2$ (Invitrogen), 200 µM of each dXTP (Invitrogen), 0.2 µM of each forward and reverse primer and 0.5 U Taq DNA polymerase (Invitrogen). The profile consisted of a denaturation step of 5 min at 94°C followed by 30 cycles of denaturation at 94°C for 30 seconds, primer annealing (temperature was in accordance with the Td of the specific primers used) for 30 seconds and extension at 72°C for 1 to 3 minutes depending on the expected length of the product. After an additional step of polymerization at 72°C for 10 min. the reaction was held at 4°C.

2.5.3 RT- PCR

The concentration of total RNA samples from different tissues was determined by spectrophotometry and normalization was carried out by northern analysis using 18 S probe to verify the correct concentrations of each sample by comparing the intensity of each band using quantity one software.

RNA samples were prepared for conversion to cDNA following treatment with 1U of RNase free DNaseI per 2.5 µg RNA for 10 minutes at 37°C (Invitrogen) to degrade contaminating DNA. The reverse transcriptase reaction was carried out according to the Invitrogen manual. Briefly, 1 µl of 0.5µg/µl random hexamer primer was added to 2.5µg DNase treated RNA in a total volume of 13 µl of DEPC-treated water and incubated at 70°C for 10 minutes in a thermocycler to denature RNA. After incubation in ice for 5 minutes, the following reagents were then added to give the final concentrations of 1X

RT buffer, 2.5 mM $MgCl_2$, 0.5 mM of each dXTP, 10 mM DTT. The mixture was incubated at 42°C for 5 minutes and then 1 µl of Superscript II RT (Invitrogen) was added. The reaction was incubated at 42°C for 50 minutes and then terminated by incubation at 70°C for 10 minutes. Resulting cDNA preparations were checked by doing PCR with 18 S primers as performed by C.Webb. 2 µl of each cDNA preparation was used for subsequent PCR amplification.

2.6 Sequencing Reaction

Plasmid DNAs were sequenced by the dideoxy chain termination method (Sanger *et al.*, 1977) using the PCR based Amersham Thermo Sequenase kit. The primers for sequencing were Lycor M13 IRD 800 (forward) and Lycor M13 IRD 700 (reverse). Samples were incubated in the thermocycler at 92 ° C for 2 min and then 30 cycles of 92° C for 30 sec, 50°C for 15 sec and 70°C for 30 sec.

Sequence reactions were processed and the sequences were assembled by Canadian Molecular Research Services.

2.7 Southern Blot Analysis

2.7.1 Transfer of Nucleic acid DNA to Biotrans

Restricted DNA was first separated on agarose gels. The DNA in the gels was denatured by gentle agitation in 1.5 M NaCl and 0.5 M NaOH for 30 minutes, twice. The denaturation buffer was then replaced by neutralization buffer containing 3 M sodium acetate (pH 5.2). The DNA was transferred to BIOTRANS membrane by the method of Southern (1975). Following the transfer, the DNA was immobilized onto the membrane by crosslinking for 5 minutes (Church and Gilbert, 1984) with an ultraviolet light.

2.7.2 Preparation of Radioactively Labeled Probe

Combinations of forward and reverse primers, e.g., 7/2 F1-7/2 R were used to prepare fragments via PCR. The 40-50 ng of PCR product or restricted DNA was labeled using the random primer method (Feinberg and Vogelstein, 1983) with *E. coli* DNA polymerase I, Klenow fragment (Amersham), and 50 μCi of [α- ^{32} P] – dCTP (Amersham). The reaction mixture was incubated for 45 min at room temperature and unincorporated nucleotides were removed by passing the labeled fragments through a Sephadex G-50 spin-column as indicated in Sambrook *et al.*, (1989). Prior to use, the labeled fragment was boiled for 5 minutes, fast cooled on ice and added to the hybridization solution.

2.7.3 Hybridization of DNA Membranes

Hybridization of radiolabeled probes to DNA was performed using 4 ml of hybridization solution (6X SSC, 5 mM EDTA, 50 mM sodium phosphate, 5X Denhardt's solution, 0.2 mg/ml herring sperm, 0.2% SDS, pH 7.0) per 100 cm^2 of membrane for at least one hour at 65°C. The probe was then added to the solution and left at 65°C overnight in a rotating TekStar Hybridization oven. The blots were then washed 2 times with 2X SSC, 0.1% SDS and then 0.1X SSC, 0.1% SDS for 20 min each at 65 °C. The membrane was exposed to BioRad Imaging Screen-K for an appropriate time. The screens were scanned in a Molecular Imager FX and the data was collected and analyzed by Quantity One software.

2.8 Molecular Biology Data Bases, Analytical Tools and Software Programs

The open reading frame (ORF) of 7/2 cDNA sequence was determined with the aid of the NCBI ORF finder at http://www.ncbi.nlm.nih.gov/gorf/gorf.html. Translation of cDNA ORF (from nt 418 to 1215 of Figure 5) into a protein sequence of 265 amino acids was performed using the DNA-MAN Program (Lynnon BioSoft version 4). The mass of the predicted peptide was calculated at http://bioinformatics.org/sms/prot_mw.html.

PROSITE was used to look for motifs on protein sequence (http://www.expasy.ch/tools/scnpsit1.html).

Sequence alignments were performed using DNA-MAN program and CLASTALW (http://www.ebi.ac.uk/clustalw/index.html).

Basic Local Alignment Search Tool (BLAST) from NCBI site was performed on both protein and DNA sequences for a sequence similarity search and to identify homology with known genes. To look for matching ESTs, searches through the soybean EST project site at Center for Computational Genomics and Bioinformatics (CCGB) at http://soybean.ccgb.umn.edu/ and also TIGR (The Institute for Genomic Research) site at http://tigrblast.tigr.org/tgi/ were performed.

Promoter search was performed at PLACE data bases (Plant Cis-acting Regulatory DNA Elements http://www.dna.affrc.go.jp/PLACE/).

To look for NLS on protein sequence the following sites were searched using PROSITE, at http://cubic.bioc.columbia.edu/predictNLS and PSORT (Prediction of Protein Sorting

Signals and Localization Sites in Amino Acid Sequences) at http://psort.ims.u-tokyo.ac.jp/form.html.

CHAPTER THREE

RESULTS

A soybean (*Glycine max* L. Merrill) nodule cDNA library had been constructed by a previous M.Sc student, C. Chan, using RNA isolated from cultivar Maple Arrow grown for 40 days post inoculation. Differential screening of this library lead to the isolation of a novel cDNA clone named 7/2. Using 7/2 cDNA as a probe, Southern blot analysis showed that 7/2 is a single copy gene.

3.1 Approaches Used to Obtain cDNA and Genomic Sequences

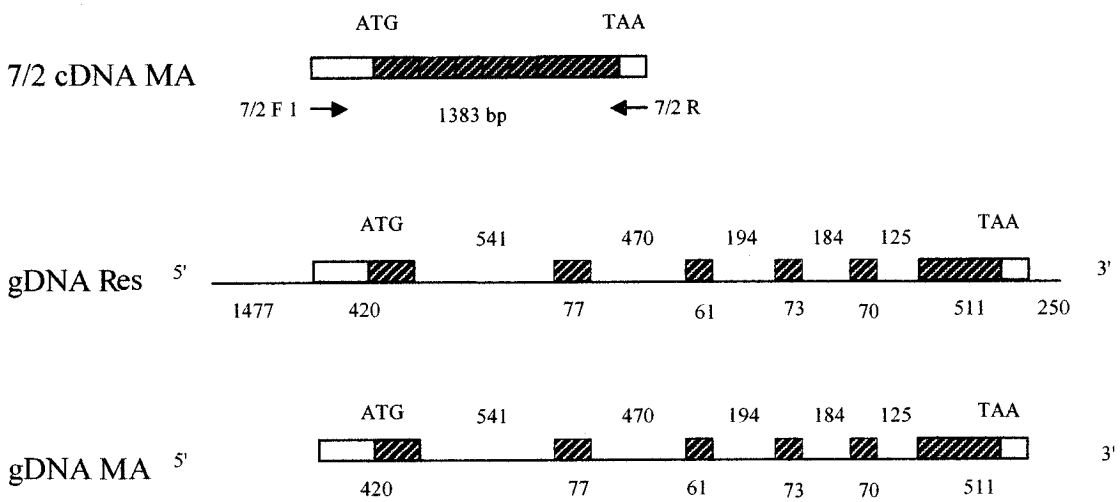
The computational full length 7/2 cDNA sequence was obtained by C. Webb who used 5' and 3' RACE to extend the original cDNA sequence obtained by C. Chan. Based upon this sequence two primers that flanked the longest potential open reading frame, 7/2 F1 and 7/2 R (Table 3 and Figure 3 in Materials and Methods) were designed by Dr. Johnson and were used to obtain a new cDNA clone from cv. Maple Arrow. This fragment was cloned into pGEM-T Easy and subclones in pUC8 were isolated following restriction digestion with EcoRI and HindIII. Each subclone was sequenced a minimum of 2 times and the cDNA sequences obtained were compared to the sequences of the genomic DNA (gDNA) contigs.

Using the cDNA as a probe, the 7/2 genomic sequence was isolated by D. Johnson from a commercial genomic library (Clonetech) constructed in the vector λEMBL3 from DNA isolated from a different cultivar, cv. Resnik. Several subclones of this gDNA sequence were isolated by traditional approaches using restriction enzymes. A gDNA sequence of 7/2 was also obtained from cv. Maple Arrow by PCR amplification of

gDNA using the same primers indicated above used to amplify the full length cDNA. This PCR based gDNA gave rise to a 2726 bp fragment which was cloned into pGEM-T Easy and then subcloned for sequencing. Figure 4 shows a schematic representation of the 7/2 cDNA and gDNA structures. Both gDNAs from cvs. Resnik and Maple Arrow have the same structure of exon and intron when aligned with the sizes indicated for each. The cv. Resnik gDNA sequence is 1477 bp and 250 bp longer respectively at 5' and 3' ends compared to cv. Maple Arrow which are indicated by lines in the figure.

Figure 4: The cDNA and gDNA structure of 7/2.

Gray boxes indicate non-coding regions and hatched boxes are coding exons. Numbers above the lines and below the five introns and six exons give the sizes in bp. The gDNA from cv. Resnik is longer at 5' and 3' ends compared to gDNA from cv. Maple Arrow as is indicated by the lines (not to scale). The arrows show the position of the primers on cDNA sequence. Res: cv. Resnik and MA: cv. Maple Arrow.


7/2 cDNA MA
ATG
TAA
7/2 F 1
1383 bp
7/2 R
gDNA Res
5'
3'
ATG
541
470
194
184
125
TAA
1477
420
77
61
73
70
511
250
gDNA MA
5'
3'
ATG
541
470
194
184
125
TAA
420
77
61
73
70
511

3.2 Analysis of the 7/2 cDNA and Genomic Sequences

The 7/2 cDNA codes for a protein of 265 amino acids with a predicted molecular weight of ~30 kDa (Discussed in Section 3.3). The alignment of the cDNA contig (cv. Maple Arrow) with the two gDNA contigs (cvs Maple Arrow and Resnik) was performed using DNAMAN software. A schematic of the gene structure is shown in Figure 4 and with the final alignment of edited genomic and cDNA sequences given in Figure 5. gDNA cv. Resnik has 1477 bp more at its 5' end compared to gDNA cv. Maple Arrow and 1345 bp more compared to cDNA sequence. Also cDNA cv. Maple Arrow is 133 bp longer at the 5' end than gDNA cv. Maple Arrow. The alignments in Figure 5 start from nucleotide 1 on cDNA, as a result the 5' end sequence of gDNA cv. Resnik prior to nucleotide 1345 is not shown. The primers used in the experiment are in italics and are shown by arrows in Figure 5.

The 7/2 genomic DNA contains 6 small exons and 5 introns (Figure 4). The sizes of the introns range from 125-541 bp and the intron sizes are conserved between the two cultivars. All 5 intron sequences begin with GT and end with AG as expected for plant type II introns (Brown and Simpson, 1998).

The initial alignments revealed several mismatches between the two cultivars and between the cDNA and genomic sequences (see Table 1 in the Appendix). These may represent true cultivar to cultivar polymorphisms or errors introduced during reading and/or editing of the sequence (sequence calls) or errors introduced by the Superscript II reverse transcriptase (RT) or Taq polymerase during the generation of the RT-PCR or PCR products used in cloning. The error rate for Taq polymerase has been measured as 1.9×10^{-5} (http://www.invitrogen.co.jp/literatures/bta0101e.pdf) and 3.4×10^{-5} for

Superscript II (http://www.invitrogen.co.jp/focus/251019.pdf). Most mismatches were located within the intron sequences but 2 of them were in coding region and could potentially result in an altered protein.

To resolve these mismatches two strategies were adopted. To resolve any problems with base "calls", repeated sequencing of clones was done. To resolve problems with mis-incorporation, many more clones arising from independent RT-PCR reactions were isolated and sequenced. Amplification of gDNA with thermostable polymerases that have lower error rates was also attempted, e.g., High Expand Fidelity PCR System with an error rate of 8.5×10^{-6} errors/bp (http://www.roche-applied-science.com/pack-insert/3553426a.pdf) and Platinum *pfx* DNA polymerase with an error rate that is 26 fold less than *Taq* polymerase (http://www.invitrogen.co.jp/literatures/bta0101e.pdf), but these failed to give PCR products in our hands and this approach was abandoned.

cDNA from the nodules was amplified using the primers 7/2 F 3 and 7/2 R (Table 3 and Figure 3 in Materials and Methods) and Taq polymerase. This process would normally give rise only to cDNA sequence but as described in Section 3.4.2 several clones representing partially spliced 7/2 mRNAs were isolated and these clones contained intron sequences that could be used to resolve mismatches. Nine individual clones were obtained and the sequences were compared to those of the previously obtained gDNA contigs and cDNA contig. We attempted to resolve the mismatches between two genomic sequences by amplification of the gDNA from cv. Maple Arrow using 7/2 F 2 and 7/2 R primers (Table 3 and Figure 3 in Materials and Methods) to isolate new gDNA clones. PCR amplification of the gDNA with these primers gave rise to a band of 1773 bp. The first exon and intron are missing in this gDNA fragment compared to the other contigs.

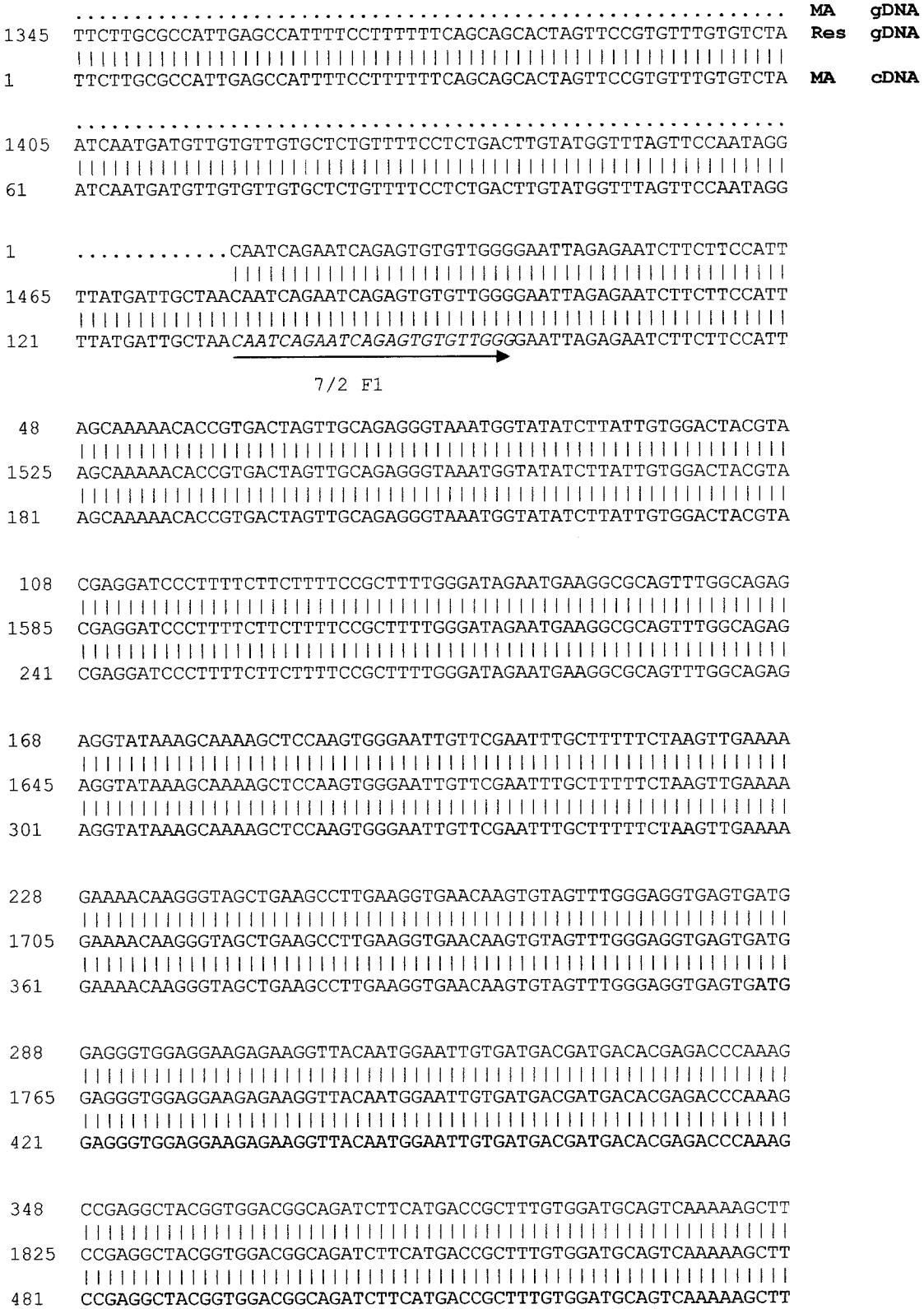
Fragments from four individual PCR reactions were separately ligated into pGEM-T Easy. After transformation, two of them gave rise to plasmids with appropriate insert size which were sequenced and used for further analysis.

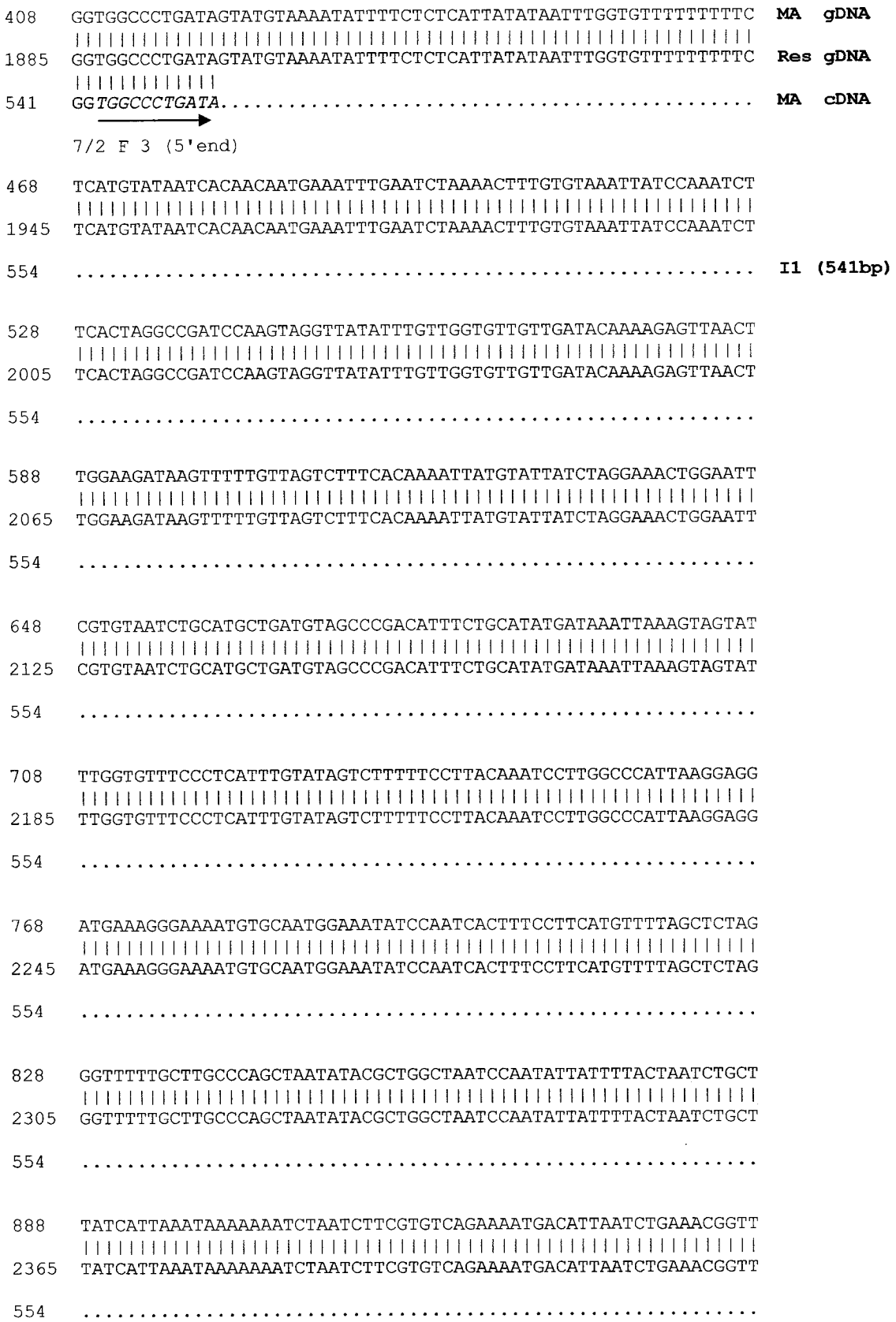
These new cDNA sequences along with the sequence of the new genomic clone from cv Maple Arrow (1773 bp) obtained by PCR with primers 7/2 F2 and 7/2 R resolved the mismatches between gDNA and cDNA. The mismatches between the two genomic DNAs were also resolved according to the new gDNA sequence and the sequences obtained from RT-PCR reactions containing introns due to partial splicing (Section 3.4.2)

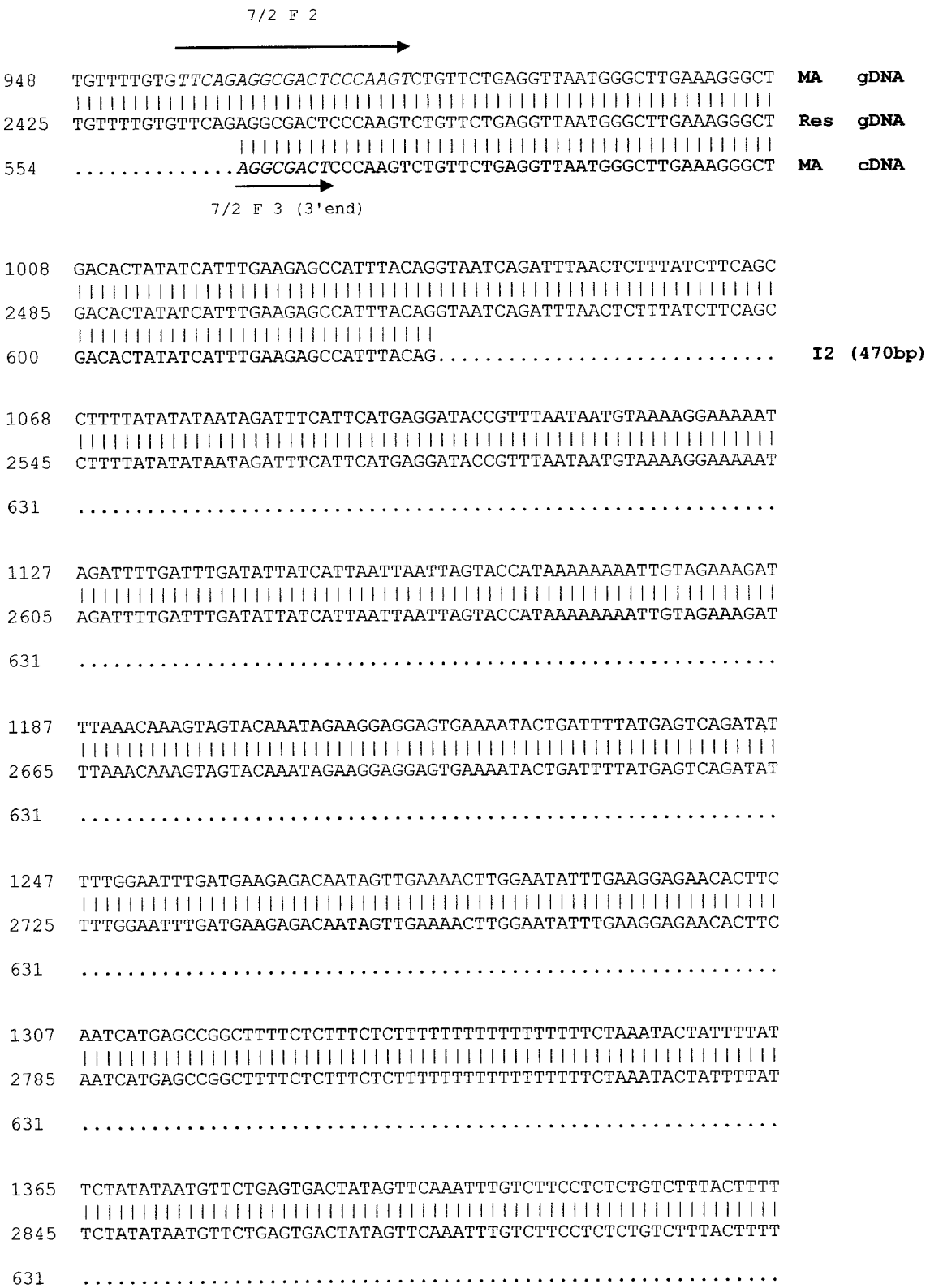
As a result of this strategy, all the mismatches between the sequences were resolved. It appears that the mismatches were generated during PCR amplification. Table 1 in Appendix summarizes information about the initial sequence differences between the contigs and their resolution, including the position of each nucleotide that differed between the contigs and the number of times each clone was sequenced.

Figure 5: Alignment of cDNA and gDNA sequences.

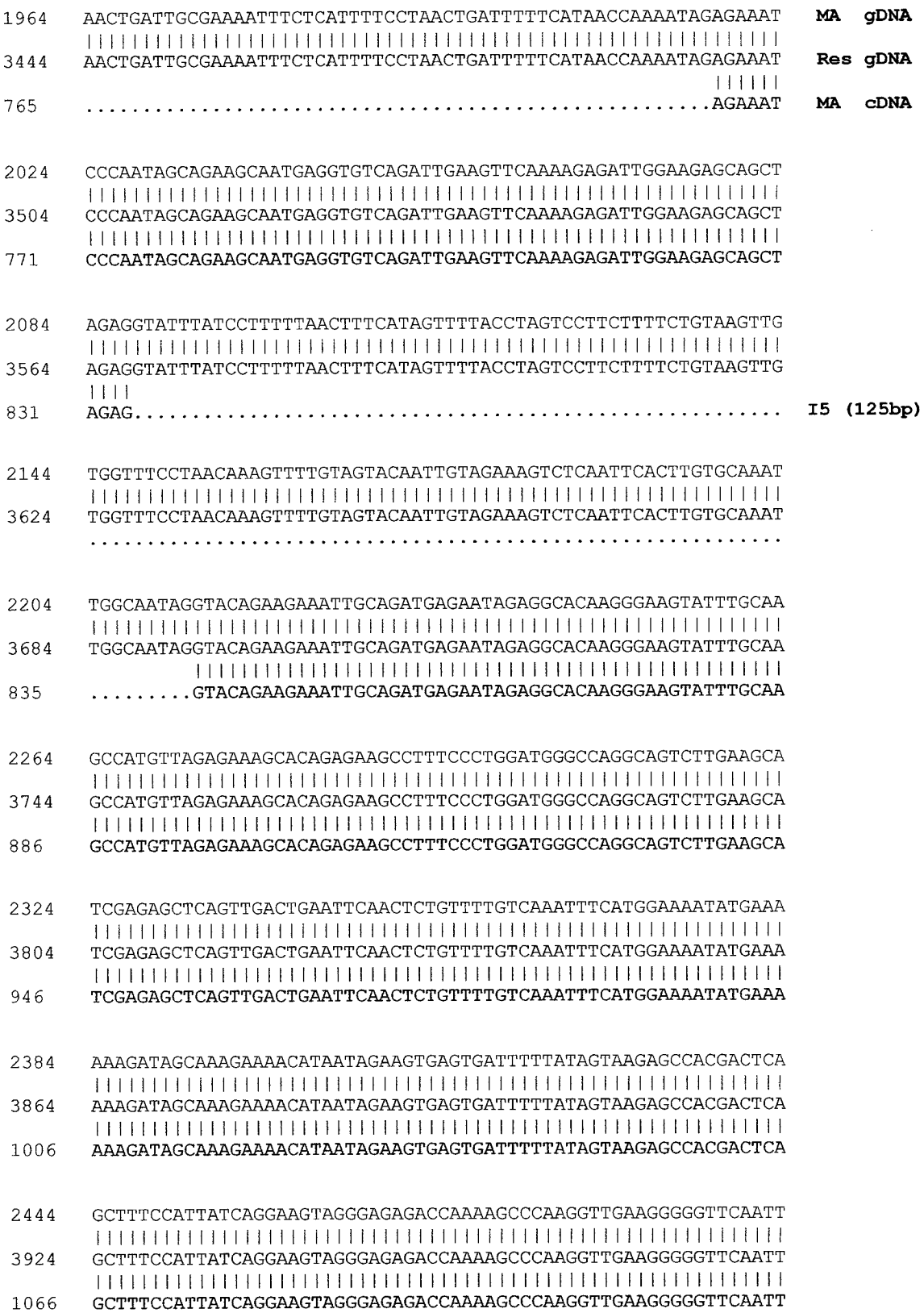
This figure shows the alignment of cDNA from cv. Maple Arrow with gDNA sequences from cvs Resnik and Maple Arrow. gDNA cv. Resnik (Res) has 1477 bp more at its 5' end compared to gDNA cv. Maple Arrow (MA) and 1345 bp more compared to cDNA sequence. Also cDNA cv. Maple Arrow is 133 bp longer than gDNA cv. Maple Arrow. The alignments demonstrated here start from nucleotide 1 on the cDNA sequence. As a result the 5' end of gDNA cv. Resnik is not shown and its sequence starts from nucleotide 1345. The ORF in the cDNA is marked by yellow. The number of the introns is written at the beginning of each in the right side of the sequence along with its size. The primer sites are in Italics and shown by arrows while the name of each primer is written underneath.

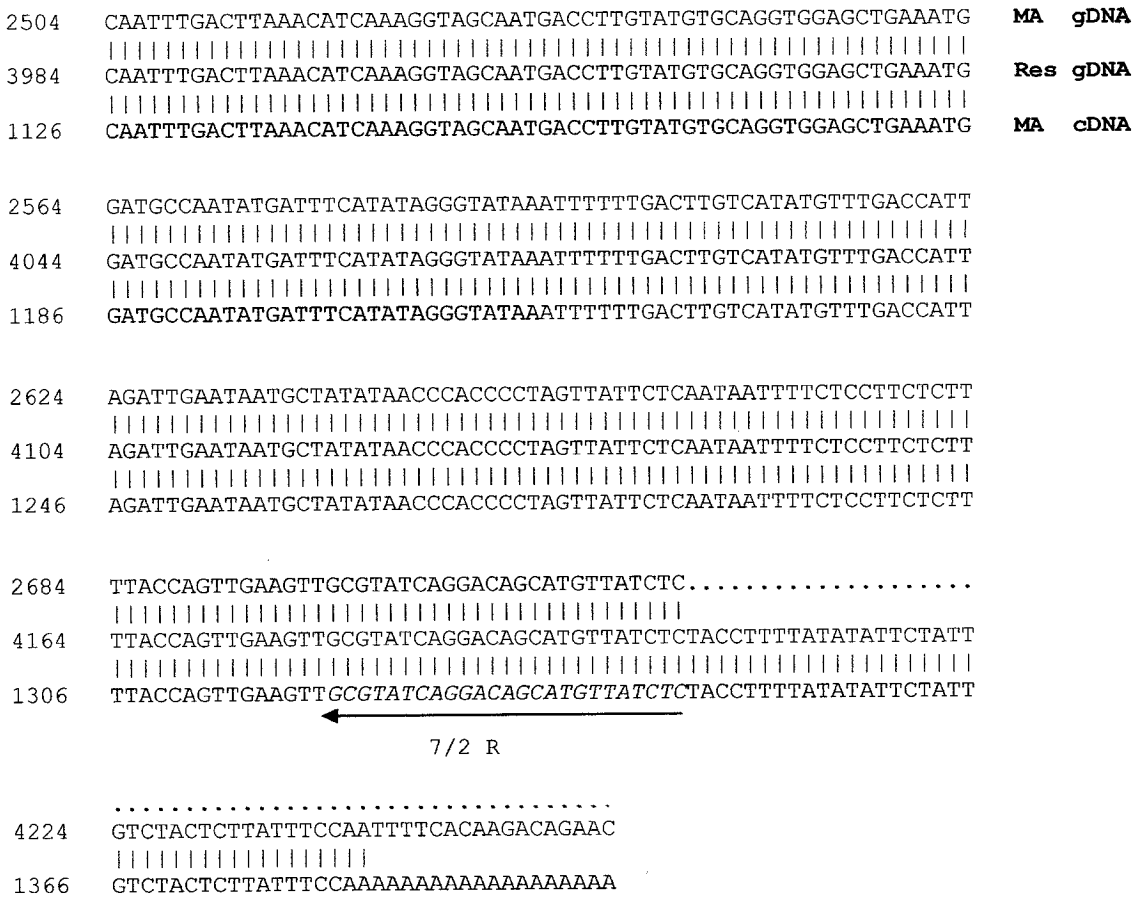

MA gDNA
1345 TTCTTGCGCCATTGAGCCATTTTCCTTTTTTCAGCAGCACTAGTTCCGTGTTTGTGTCTA
Res gDNA
1 TTCTTGCGCCATTGAGCCATTTTCCTTTTTTCAGCAGCACTAGTTCCGTGTTTGTGTCTA
MA cDNA
1405 ATCAATGATGTTGTGTTGTGCTCTGTTTTCCTCTGACTTGTATGGTTTAGTTCCAATAGG
61 ATCAATGATGTTGTGTTGTGCTCTGTTTTCCTCTGACTTGTATGGTTTAGTTCCAATAGG
1 .....CAATCAGAATCAGAGTGTGTTGGGGAATTAGAGAATCTTCTTCCATT
1465 TTATGATTGCTAACAATCAGAATCAGAGTGTGTTGGGGAATTAGAGAATCTTCTTCCATT
121 TTATGATTGCTAACAATCAGAATCAGAGTGTGTTGGGGAATTAGAGAATCTTCTTCCATT
7/2 F1
48 AGCAAAAACACCGTGACTAGTTGCAGAGGGTAAATGGTATATCTTATTGTGGACTACGTA
1525 AGCAAAAACACCGTGACTAGTTGCAGAGGGTAAATGGTATATCTTATTGTGGACTACGTA
181 AGCAAAAACACCGTGACTAGTTGCAGAGGGTAAATGGTATATCTTATTGTGGACTACGTA
108 CGAGGATCCCTTTTCTTCTTTTCCGCTTTTGGGATAGAATGAAGGCGCAGTTTGGCAGAG
1585 CGAGGATCCCTTTTCTTCTTTTCCGCTTTTGGGATAGAATGAAGGCGCAGTTTGGCAGAG
241 CGAGGATCCCTTTTCTTCTTTTCCGCTTTTGGGATAGAATGAAGGCGCAGTTTGGCAGAG
168 AGGTATAAAGCAAAAGCTCCAAGTGGGAATTGTTCGAATTTGCTTTTTCTAAGTTGAAAA
1645 AGGTATAAAGCAAAAGCTCCAAGTGGGAATTGTTCGAATTTGCTTTTTCTAAGTTGAAAA
301 AGGTATAAAGCAAAAGCTCCAAGTGGGAATTGTTCGAATTTGCTTTTTCTAAGTTGAAAA
228 GAAAACAAGGGTAGCTGAAGCCTTGAAGGTGAACAAGTGTAGTTTGGGAGGTGAGTGATG
1705 GAAAACAAGGGTAGCTGAAGCCTTGAAGGTGAACAAGTGTAGTTTGGGAGGTGAGTGATG
361 GAAAACAAGGGTAGCTGAAGCCTTGAAGGTGAACAAGTGTAGTTTGGGAGGTGAGTGATG
288 GAGGGTGGAGGAAGAGAAGGTTACAATGGAATTGTGATGACGATGACACGAGACCCAAAG
1765 GAGGGTGGAGGAAGAGAAGGTTACAATGGAATTGTGATGACGATGACACGAGACCCAAAG
421 GAGGGTGGAGGAAGAGAAGGTTACAATGGAATTGTGATGACGATGACACGAGACCCAAAG
348 CCGAGGCTACGGTGGACGGCAGATCTTCATGACCGCTTTGTGGATGCAGTCAAAAAGCTT
1825 CCGAGGCTACGGTGGACGGCAGATCTTCATGACCGCTTTGTGGATGCAGTCAAAAAGCTT
481 CCGAGGCTACGGTGGACGGCAGATCTTCATGACCGCTTTGTGGATGCAGTCAAAAAGCTT


408 GGTGGCCCTGATAGTATGTAAAATATTTTCTCTCATTATATAATTTGGTGTTTTTTTTTC MA gDNA
1885 GGTGGCCCTGATAGTATGTAAAATATTTTCTCTCATTATATAATTTGGTGTTTTTTTTTC Res gDNA
541 GGTGGCCCTGATA..... MA cDNA
7/2 F 3 (5'end)
468 TCATGTATAATCACAACAATGAAATTTGAATCTAAAACTTTGTGTAAATTATCCAAATCT
1945 TCATGTATAATCACAACAATGAAATTTGAATCTAAAACTTTGTGTAAATTATCCAAATCT
554 ..... I1 (541bp)
528 TCACTAGGCCGATCCAAGTAGGTTATATTTGTTGGTGTTGTTGATACAAAAGAGTTAACT
2005 TCACTAGGCCGATCCAAGTAGGTTATATTTGTTGGTGTTGTTGATACAAAAGAGTTAACT
554 .....
588 TGGAAGATAAGTTTTTGTTAGTCTTTCACAAAATTATGTATTATCTAGGAAACTGGAATT
2065 TGGAAGATAAGTTTTTGTTAGTCTTTCACAAAATTATGTATTATCTAGGAAACTGGAATT
554 .....
648 CGTGTAATCTGCATGCTGATGTAGCCCGACATTTCTGCATATGATAAATTAAAGTAGTAT
2125 CGTGTAATCTGCATGCTGATGTAGCCCGACATTTCTGCATATGATAAATTAAAGTAGTAT
554 .....
708 TTGGTGTTTCCCTCATTTGTATAGTCTTTTTCCTTACAAATCCTTGGCCCATTAAGGAGG
2185 TTGGTGTTTCCCTCATTTGTATAGTCTTTTTCCTTACAAATCCTTGGCCCATTAAGGAGG
554 .....
768 ATGAAAGGGAAAATGTGCAATGGAAATATCCAATCACTTTCCTTCATGTTTTAGCTCTAG
2245 ATGAAAGGGAAAATGTGCAATGGAAATATCCAATCACTTTCCTTCATGTTTTAGCTCTAG
554 .....
828 GGTTTTTGCTTGCCCAGCTAATATACGCTGGCTAATCCAATATTATTTTACTAATCTGCT
2305 GGTTTTTGCTTGCCCAGCTAATATACGCTGGCTAATCCAATATTATTTTACTAATCTGCT
554 .....
888 TATCATTAAATAAAAAAATCTAATCTTCGTGTCAGAAAATGACATTAATCTGAAACGGTT
2365 TATCATTAAATAAAAAAATCTAATCTTCGTGTCAGAAAATGACATTAATCTGAAACGGTT
554 .....


7/2 F 2
948 TGTTTTGTGTTCAGAGGCGACTCCCAAGTCTGTTCTGAGGTTAATGGGCTTGAAAGGGCT MA gDNA
2425 TGTTTTGTGTTCAGAGGCGACTCCCAAGTCTGTTCTGAGGTTAATGGGCTTGAAAGGGCT Res gDNA
554 .....AGGCGACTCCCAAGTCTGTTCTGAGGTTAATGGGCTTGAAAGGGCT MA cDNA
7/2 F 3 (3'end)
1008 GACACTATATCATTTGAAGAGCCATTTACAGGTAATCAGATTTAACTCTTTATCTTCAGC
2485 GACACTATATCATTTGAAGAGCCATTTACAGGTAATCAGATTTAACTCTTTATCTTCAGC
600 GACACTATATCATTTGAAGAGCCATTTACAG..... I2 (470bp)
1068 CTTTTATATATAATAGATTTCATTCATGAGGATACCGTTTAATAATGTAAAAGGAAAAAT
2545 CTTTTATATATAATAGATTTCATTCATGAGGATACCGTTTAATAATGTAAAAGGAAAAAT
631 .....
1127 AGATTTTGATTTGATATTATCATTAATTAATTAGTACCATAAAAAAAATTGTAGAAAGAT
2605 AGATTTTGATTTGATATTATCATTAATTAATTAGTACCATAAAAAAAATTGTAGAAAGAT
631 .....
1187 TTAAACAAAGTAGTACAAATAGAAGGAGGAGTGAAAATACTGATTTTATGAGTCAGATAT
2665 TTAAACAAAGTAGTACAAATAGAAGGAGGAGTGAAAATACTGATTTTATGAGTCAGATAT
631 .....
1247 TTTGGAATTTGATGAAGAGACAATAGTTGAAAACTTGGAATATTTGAAGGAGAACACTTC
2725 TTTGGAATTTGATGAAGAGACAATAGTTGAAAACTTGGAATATTTGAAGGAGAACACTTC
631 .....
1307 AATCATGAGCCGGCTTTTCTCTTTCTCTTTTTTTTTTTTTTTTCTAAATACTATTTTAT
2785 AATCATGAGCCGGCTTTTCTCTTTCTCTTTTTTTTTTTTTTTTCTAAATACTATTTTAT
631 .....
1365 TCTATATAATGTTCTGAGTGACTATAGTTCAAATTTGTCTTCCTCTCTGTCTTTACTTTT
2845 TCTATATAATGTTCTGAGTGACTATAGTTCAAATTTGTCTTCCTCTCTGTCTTTACTTTT
631 .....

```
1425  CTTGCTTTATCCTAACCATCATCAAAGAAAATACCATTTAGAAGAATAATCTGGAATTTC  MA  gDNA
      ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
2905  CTTGCTTTATCCTAACCATCATCAAAGAAAATACCATTTAGAAGAATAATCTGGAATTTC  Res gDNA

631   ..........................................................  MA  cDNA


1485  TCGTTATTTGTTAATTTGCAGAAGTATAGACTTGGACAGCAAGCTCGGAAACAAAATGAG
      ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
2965  TCGTTATTTGTTAATTTGCAGAAGTATAGACTTGGACAGCAAGCTCGGAAACAAAATGAG
                              ||||||||||||||||||||||||||||||||||||
631   ....................AAGTATAGACTTGGACAGCAAGCTCGGAAACAAAATGAG


1545  GATATGCACAAAGAAAATAATAGTGAGTCCATTGCAAGTTTTAACAAACACTGGGCATCC
      ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
3025  GATATGCACAAAGAAAATAATAGTGAGTCCATTGCAAGTTTTAACAAACACTGGGCATCC
      ||||||||||||||||||||
670   GATATGCACAAAGAAAATAATA.....................................  I3 (194bp)


1605  CATGCATTTCTTTTACTTATTATCTAGAATAGTTTGTCCAACTTCTCCCACAGAGATTTT
      ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
3085  CATGCATTTCTTTTACTTATTATCTAGAATAGTTTGTCCAACTTCTCCCACAGAGATTTT

692   ..........................................................


1665  GTCCTTCCATTTATTTACTTAAATAGAGAACATAAAATCGGGAAATGTTTGGTAATAAT
      ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
3145  GTCCTTCCATTTATTTACTTAAATAGAGAACATAAAATCGGGAAATGTTTGGTAATAAT

692   ..........................................................


1724  GTAATAATATTCTTTCCATGTTTCTGTTTCTTTTTAGGATGTTCGTATGTAAATTTTAGC
      ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
3204  GTAATAATATTCTTTCCATGTTTCTGTTTCTTTTTAGGATGTTCGTATGTAAATTTTAGC
                                          ||||||||||||||||||||||||
692   ..................................GATGTTCGTATGTAAATTTTAGC


1784  AATCGTTCTTCAGCACCTAACACCAGTTACAGAGGTGATGATGAAGGGGGGTATGTTTTA
      ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
3264  AATCGTTCTTCAGCACCTAACACCAGTTACAGAGGTGATGATGAAGGGGGGTATGTTTTA
      ||||||||||||||||||||||||||||||||||||||||||||||||||
715   AATCGTTCTTCAGCACCTAACACCAGTTACAGAGGTGATGATGAAGGGGG.........  I4 (184bp)


1844  GATTATATGCATATAGTTTCCATGAGCCAAACTCTCATCTAAATTTCGTATCCACGTGAT
      ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
3324  GATTATATGCATATAGTTTCCATGAGCCAAACTCTCATCTAAATTTCGTATCCACGTGAT

765   ..........................................................


1904  TAGCATTTATAGTTTTGTGTACTTCAAAAAAACAGAAAACAAAATAAGTAGTGTAAATAC
      ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
3384  TAGCATTTATAGTTTTGTGTACTTCAAAAAAACAGAAAACAAAATAAGTAGTGTAAATAC

765   ..........................................................
```


1964 AACTGATTGCGAAAATTTCTCATTTTCCTAACTGATTTTTCATAACCAAAATAGAGAAAT MA gDNA
3444 AACTGATTGCGAAAATTTCTCATTTTCCTAACTGATTTTTCATAACCAAAATAGAGAAAT Res gDNA
765 .....AGAAAT MA cDNA
2024 CCCAATAGCAGAAGCAATGAGGTGTCAGATTGAAGTTCAAAAGAGATTGGAAGAGCAGCT
3504 CCCAATAGCAGAAGCAATGAGGTGTCAGATTGAAGTTCAAAAGAGATTGGAAGAGCAGCT
771 CCCAATAGCAGAAGCAATGAGGTGTCAGATTGAAGTTCAAAAGAGATTGGAAGAGCAGCT
2084 AGAGGTATTTATCCTTTTTAACTTTCATAGTTTTACCTAGTCCTTCTTTTCTGTAAGTTG
3564 AGAGGTATTTATCCTTTTTAACTTTCATAGTTTTACCTAGTCCTTCTTTTCTGTAAGTTG
831 AGAG..... I5 (125bp)
2144 TGGTTTCCTAACAAAGTTTTGTAGTACAATTGTAGAAAGTCTCAATTCACTTGTGCAAAT
3624 TGGTTTCCTAACAAAGTTTTGTAGTACAATTGTAGAAAGTCTCAATTCACTTGTGCAAAT
2204 TGGCAATAGGTACAGAAGAAATTGCAGATGAGAATAGAGGCACAAGGGAAGTATTTGCAA
3684 TGGCAATAGGTACAGAAGAAATTGCAGATGAGAATAGAGGCACAAGGGAAGTATTTGCAA
835 .....GTACAGAAGAAATTGCAGATGAGAATAGAGGCACAAGGGAAGTATTTGCAA
2264 GCCATGTTAGAGAAAGCACAGAGAAGCCTTTCCCTGGATGGGCCAGGCAGTCTTGAAGCA
3744 GCCATGTTAGAGAAAGCACAGAGAAGCCTTTCCCTGGATGGGCCAGGCAGTCTTGAAGCA
886 GCCATGTTAGAGAAAGCACAGAGAAGCCTTTCCCTGGATGGGCCAGGCAGTCTTGAAGCA
2324 TCGAGAGCTCAGTTGACTGAATTCAACTCTGTTTTGTCAAATTTCATGGAAAATATGAAA
3804 TCGAGAGCTCAGTTGACTGAATTCAACTCTGTTTTGTCAAATTTCATGGAAAATATGAAA
946 TCGAGAGCTCAGTTGACTGAATTCAACTCTGTTTTGTCAAATTTCATGGAAAATATGAAA
2384 AAAGATAGCAAAGAAAACATAATAGAAGTGAGTGATTTTTATAGTAAGAGCCACGACTCA
3864 AAAGATAGCAAAGAAAACATAATAGAAGTGAGTGATTTTTATAGTAAGAGCCACGACTCA
1006 AAAGATAGCAAAGAAAACATAATAGAAGTGAGTGATTTTTATAGTAAGAGCCACGACTCA
2444 GCTTTCCATTATCAGGAAGTAGGGAGAGACCAAAAGCCCAAGGTTGAAGGGGGTTCAATT
3924 GCTTTCCATTATCAGGAAGTAGGGAGAGACCAAAAGCCCAAGGTTGAAGGGGGTTCAATT
1066 GCTTTCCATTATCAGGAAGTAGGGAGAGACCAAAAGCCCAAGGTTGAAGGGGGTTCAATT


2504 CAATTTGACTTAAACATCAAAGGTAGCAATGACCTTGTATGTGCAGGTGGAGCTGAAATG MA gDNA
3984 CAATTTGACTTAAACATCAAAGGTAGCAATGACCTTGTATGTGCAGGTGGAGCTGAAATG Res gDNA
1126 CAATTTGACTTAAACATCAAAGGTAGCAATGACCTTGTATGTGCAGGTGGAGCTGAAATG MA cDNA
2564 GATGCCAATATGATTTCATATAGGGTATAAATTTTTTGACTTGTCATATGTTTGACCATT
4044 GATGCCAATATGATTTCATATAGGGTATAAATTTTTTGACTTGTCATATGTTTGACCATT
1186 GATGCCAATATGATTTCATATAGGGTATAAATTTTTTGACTTGTCATATGTTTGACCATT
2624 AGATTGAATAATGCTATATAACCCACCCCTAGTTATTCTCAATAATTTTCTCCTTCTCTT
4104 AGATTGAATAATGCTATATAACCCACCCCTAGTTATTCTCAATAATTTTCTCCTTCTCTT
1246 AGATTGAATAATGCTATATAACCCACCCCTAGTTATTCTCAATAATTTTCTCCTTCTCTT
2684 TTACCAGTTGAAGTTGCGTATCAGGACAGCATGTTATCTC.....
4164 TTACCAGTTGAAGTTGCGTATCAGGACAGCATGTTATCTCTACCTTTTATATATTCTATT
1306 TTACCAGTTGAAGTTGCGTATCAGGACAGCATGTTATCTCTACCTTTTATATATTCTATT
7/2 R
4224 GTCTACTCTTATTTCCAATTTTCACAAGACAGAAC
1366 GTCTACTCTTATTTCCAAAAAAAAAAAAAAAAAA

3.3 7/2 Conceptual Protein

Although the 7/2 protein has not been identified experimentally, the longest Open Reading Frame (ORF) of 7/2 cDNA translates to a protein with 265 amino acids and a calculated molecular weight of 29.98 kDa. This protein contains 34% hydrophobic amino acids, 35% uncharged amino acids, 14% acidic amino acids, and 17% basic amino acids. Figure 6 shows the amino acid sequence of the 7/2 conceptual protein.

A search using the full length 7/2 protein sequence to find similar proteins in the SwissProt database at NCBI (BLASTP, December 2004) suggests that 7/2 belongs to the family of Myb-related proteins which contain a single Myb DNA binding domain. The DNA binding domain comprises a conserved region of 52 aa (amino acids 23 to 74 in Figure 7).

Figure 6: Sequence of the 7/2 conceptual protein.

7/2 conceptual protein has 265 amino acids and a calculated molecular weight of 29.98 kDa.

```
1     MEGGGREGYNGIVMTMTRDPKPRLRWTADLHDRFVDAVKKLGGPDKATPK
51    SVLRLMGLKGLTLYHLKSHLQKYRLGQQARKQNEDMHKENNRCSYVNFSN
101   RSSAPNTSYRGDDEGGEIPIAEAMRCQIEVQKRLEEQLEVQKKLQMRIEA
151   QGKYLQAMLEKAQRSLSLDGPGSLEASRAQLTEFNSVLSNFMENMKKDSK
201   ENIIEVSDFYSKSHDSAFHYQEVGRDQKPKVEGGSIQFDLNIKGSNDLVC
251   AGGAEMDANMISYRV
```

Figure 7: Presence of single repeat Myb-DNA binding domain in the 7/2 protein.

This figure shows the result of BLASTP search recognizing a Myb-DNA binding domain on 7/2 protein. The alignment underneath shows the conserved Myb-containing region of 7/2 protein versus the consensus Myb domain as defined by NCBI. Identical amino acids are in red and chemically related amino acids are in blue.


```
7/2 protein: 23 RLRWTADLHDRFVDAVKKLGGPDKATPKSVLRLMGLKGLTLYHLKSHLQKYR 74
consensus  :  1 RGPWTPEEDELLIEAVAKHGNGN---WSKIAKKLP--GRTDKQCKNRWNNYL 47
```

Searches of the protein data base identified many genes that were similar because of matches within the Myb domain (data not shown, see Discussion). Thus in order to search for proteins with a high degree of similarity to 7/2, we removed the amino terminal end of 7/2 (amino acids 1-74) containing the Myb domain and re-queried the data base with the region containing amino acids 75 to 265.

Several strong matches with conceptual proteins were recovered but the best two matches (match values $2 \times e^{-33}$ and $6 \times e^{-27}$ respectively) are a putative calcium dependent protein kinase substrate protein (CDPKS, Accession no.AAP45171) and a putative phosphate starvation response regulator (PSRR, Accession no.AAP45156), both from *Solanum bulbocastanum*. The alignment of these proteins with 7/2 is shown in Figure 8. All three proteins are about the same size and all contain a single Myb DNA binding domain and another conserved region, C-terminal to the DNA binding domain, which may be a putative activation domain. Outside of these two conserved domains, the three proteins show little similarity, perhaps allowing for specificity of function.

Figure 8: Alignment of 7/2 protein with CDPKS and PSRR.

CLUSTALW alignment of 7/2 protein with CDPKS (Calcium Dependent Protein Kinase Substrate protein, AAP45171) and PSRR (Phosphate Starvation Response Regulator, AAP45156) from *Solanum bulbocastanum.* The DNA binding domain of 7/2 protein is in red. The potential activation domains are in magenta. Positions that are identical in three sequences are indicated with a star (*), fully conserved "strong group" are indicated with double dots (:) and fully conserved "weak groups" are indicated by a single dot (.). These two types of amino acid groups are defined by the CLUSTALW program.

```
CDPKS   MDRMYSGGGDMGYGYE-NGVVM--TRDPKPRLRWTADLHDRFVDAVTKLGGPDKATPKSV 57
7/2     ME----GGGREGY----NGIVMTMTRDPKPRLRWTADLHDRFVDAVKKLGGPDKATPKSV 52
PSRR    MERAGYGVGVGGAGAVGAGVVL--SRDPKPRLRWTPDLHERFVEAVTKLGGPDKATPKSV 58
        *:    * *  *      *:*:  :**********.***:***:**.*************

CDPKS   LRLMGLKGLTLYHLKSHLQKYRLGQQTKKQNAAEQNRENIGESFRQFSLHSSGPSITSSS 117
7/2     LRLMGLKGLTLYHLKSHLQKYRLGQQARKQNED-MHKENNRCSYVNFSNRSSAP-NTSYR 110
PSRR    LRLMGMKGLTLYHLKSHLQKYRLGKQNKKDTGLEASRG--AFAAHGISFASAAPPTIPSA 116
        *****:******************:* :*:.     :     :   :*  *:.*   .

CDPKS   MDGMQGEAPISEALRCQIEVQKRLHEQLEVQQKLQMRIEAQGKYLQAILDKAQKSLSTDM 177
7/2     GDDEGGEIPIAEAMRCQIEVQKRLEEQLEVQKKLQMRIEAQGKYLQAMLEKAQRSLS--L 168
PSRR    ENNNAGETPLADALRYQIEVQRKLHEQLEVQKKLQMRIEAQGKYLQTILEKAQNNLSYDA 176
         :.  ** *:::*:* *****::*.******:***************::*:***..**

CDPKS   NSPSAVDETRAQLTDFNIALSNLMDYMHGHN-GDETSAGERTQDDTNKDLQRSTYLTEGE 236
7/2     DGPGSLEASRAQLTEFNSVLSNFMENMK----KDSKENIIEVSDFYSKSHDSAFHYQEVG 224
PSRR    TGTANLEATRTQLTDFNLALSGFMNNVSQVCEQNNGELAKAISEDNLRTTNLGFQLYHGI 236
         ... :: :*:***:** .**.:*: :      :. .     .:   :  : .    .

CDPKS   QKKI-MNIKLEETSVSFDLNSRSS-YDFIG---MSSAALEAKHFSNGRLEI 282
7/2     RD---QKPKVEGGSIQFDLNIKGS-NDLVC---AGGAEMDANMISYRV--- 265
PSRR    QDSDDVKCSQDEGLLLLDLNIKGGGYDHLSSNAMRGGESGLKISQHRR--- 284
        :.    : . :   : :*** :..  * :      ..    :  .
```

As there was no further information available about the properties of these two conceptual proteins in their respective GENBANK accessions or in ENTREZ, we looked at other sequences identified by BLAST searches as being similar to 7/2 for which there was some experimental data. Although not the best match (with E value of $3 \times e^{-8}$), CSP1 (Accession no. AAF32350) from the ice plant, *Mesembryanthemum crystallinum*, has been characterized experimentally. With 470 amino acids, the protein is larger than 7/2 mainly due to an amino terminal extension of 235 amino acids (Figure 9). CSP1 is a Calcium Dependent Protein Kinase Substrate (CDPKS) and is a member of the pseudo-response regulator-like proteins that have a conserved basic helix-loop-helix DNA binding domain and a C-terminal activation domain (Patharkar and Cushman, 2000). Subclass II response regulators contain a putative Myb DNA-binding domain (Lohrmann *et al.*, 1999). The DNA binding domain is at position 257-309 and the activation domain at position 343-386 on the protein sequence (Figure 9). The C-terminal domain in CSP1 has acidic residues and a glutamine-rich stretch which may confer transcription activating properties to this protein (Mitchell and Tjian, 1989; Courey and Tjian, 1988). BLASTP search identified the DNA binding domain of CSP1 protein as Myb-like domain. Besides the similarity between the DNA binding domain on 7/2 and CSP1, the 7/2 protein sequence also has a glutamine and glutamic acid rich stretch from amino acids 120-163. The CSP1 protein sequence contains a potential nuclear localization signal (NLS) at its C-terminus, from amino acids 460-464. It also contains numerous serine/threonine residues that are potential phosphorylation sites for CDPKs. CSP1 can be phosphorylated *in vitro* by the *M. crystallinum* CDPK, McCDPK1 which was induced by stress (Patharkar and Cushman, 2000) although the site of phosphorylation has not been

mapped. Because of these properties, the DNA binding domain, the activation domain, the potential phosphorylation sites and the nuclear localization signal, it is assumed that CSP1 is a transcriptional activator. The gene(s) that are regulated by CSP1 has not yet been identified (Patharkar and Cushman, 2000).

In order to better visualize the relationship between them, the CSP1 and 7/2 protein sequences were aligned using CLUSTALW (Figure 9).

Figure 9: Alignment of 7/2 protein sequence with CSP1.

Similarity between 7/2 and CSP1 amino acids is indicated by a (*), (:) and (.) as defined in Figure 8 according to CLUSTALW program. The DNA binding domain is in red and the putative activation domain (rich in acidic residues and glutamine) is in magenta. The underlined amino acids within the DNA binding domain are identical to amino acid sequences found in the Myb DNA binding domains of Transfactor (accession no. BAA75684), Psr1 (AAD55941), ARR1 (BAA74528), ARR2 (BAA74527) and ARLP1 (CAA06431).

```
CSP1    MNMRPALPMQTSGGNCFNELKVSGPYSSQLPVLPNPLEDFPKSPDPFAVSSSREMIPNPL 60
7/2     ------------------------------------------------------------

CSP1    QIQANPMVSHFGSSHNSSTYASGFPTDLHFPSFSPRERQSQNSPLGGGVAFPPSQNTSPD 120
7/2     ------------------------------------------------------------

CSP1    VQSAGFINYQKEDDDNSWSTGHLQDLLDFPEGIPVSNGQVGTSTEVMSNDNHVKRIGWRE 180
7/2     ------------------------------------------------------------

CSP1    LTEDLYADSIEPNWNDFLADSNVADQQPKVTQPSSDVRVHQPLIQQQLSLPPREVSAAAN 240
7/2     -----------------------------------------------------MEGGG 5
                                                                 ...

CSP1    QTSAAANQTSAAHSNRPRMRWTPELHEAFVDAVNQLGGSERATPKGVLRHMNVEGLTIYH 300
7/2     REGYNGIVMTMTRDPKPRLRWTADLHDRFVDAVKKLGGPDKATPKSVLRLMGLKGLTLYH 65
        : .  .   : ::. :**:***.:**: *****::***.::****.*** *.::***:**

CSP1    VKSHLQKYRTARVRPESSE----------GNSERRASSVDPVSSVDLKTSVTITEALR 348
7/2     LKSHLQKYRLGQQARKQNEDMHKENNRCSYVNFSNRSSAPNTSYRGDDEGGEIPIAEAMR 125
        :******** .:   :..* .  *:* ::.  .  . *  .:.*:**:*

CSP1    MQMEVQKQLHEQLEIQRKLQLQIEEQGKYLLQMLENQN----------KVEKEKLN 394
7/2     CQIEVQKRLEEQLEVQKKLQMRIEAQGKYLQAMLEKAQRSLSLDGPGSLEASRAQLTEFN 185
         *:****:*.****:*:***::** *****  ***: :          :.:  ::*

CSP1    PDGSSAHNDKSEGSQPEPSREGAVISISSQGPGESSHGSKGKQKAPEADTTGDHHLEDGG 454
7/2     SVLSNFMENMKKDSKENIIEVSDFYSKSHDSAFHYQEVGRDQKPKVEGGSIQFDLNIKGS 245
        .  *.  :: .:.*: :  . . . * * :.. . .. .:.::   *..:   .  .*.

CSP1    SNPPPMKRARTDDSFL---- 470
7/2     NDLVCAGGAEMDANMISYRV 265
        .:      *. * .::
```

Another protein which has been characterized, and is 74% identical with the 7/2 protein throughout its DNA binding domain, is Psr1 (AAD55941). Psr1 is a nuclear-localized protein in *Chlamydomonas reinhardtii* which is involved in the regulation of phosphorus metabolism. It activates genes in response to P starvation, contains a single Myb DNA binding at its N-terminus and its C-terminus possesses a glutamine rich sequence. These are characteristics of transcription factors (Wykoff *et al.*, 1999). Removal of DNA binding and putative activation domain from 7/2 protein didn't pick any significant similarity in the database.

To help to identify similar motifs on the 7/2 protein, and thus infer its function, a search at PROSITE (http://www.expasy.ch/tools/scnpsit1.html) was performed. The identified motifs on 7/2 protein are presented in Table 4. Sites with similar characteristics were found in the CSP1 protein, however, these were not located at the equivalent position in the alignment with the 7/2 sequence. The only motif in the CSP1 protein that aligned with 7/2 was a potential protein kinase c phosphorylation site, TPK, found from amino acids 48-50 on 7/2 protein and amino acids 283-285 on CSP1 (see Figure 9). According to Table 4, there are a few potential phosphorylation sites in the 7/2 protein. As the alignment in Figure 9 demonstrates, the 7/2 protein also contains a glutamine rich region which is C-terminal to the DNA binding domain and may function as a putative activation domain. No NLS was identified on 7/2 protein using PROSITE, at http://cubic.bioc.columbia.edu/predictNLS and PSORT (Prediction of Protein Sorting Signals and Localization Sites in Amino Acid Sequences http://psort.ims.u-tokyo.ac.jp/form.html). Although no significant NLS was identified in the 7/2 protein,

based on the other characteristics of 7/2 protein such as the presence of a DNA binding domain and putative activation domain, it is probable that 7/2 is a transcription factor.

Table 4: The motifs on the 7/2 protein identified by PROSITE.

The position of the motif in the sequence, the sequence motif and the identified character are listed in this table.

# of matches	Position	aa sequence	Characteristic
3	97-100 100-103 106-109	NFSN NRSS NTSY	N-glycosylation site
1	196-199	KKDS	cAMP &cGMP dependent protein kinase phosphorylation
4	48-50 99-101 108-110 262-264	TPK SNR SYR SYR	protein kinase c phosphorylation site
3	4-9 11-16 172-177	GGREGY GIVMTM GSLEAS	N-myristoylation site

3.4 Expression of 7/2 in Soybean Plants

3.4.1 Tissue Distribution

Preliminary experiments using Northern transfer analysis demonstrated that 7/2 encodes an mRNA of approximately 1400 bases (a minimum size calculated from the cDNA sequence is 1383 bases excluding poly "A" tail, data not shown). To further investigate the temporal and spatial patterns of 7/2 expression, the amount of 7/2-specific message was examined by RT-PCR. Total RNA from seedlings, 7 dpi cotyledons, 24 dpi roots, 24 dpi stems, 24 dpi "younger" leaves (not fully expanded leaves at the top of the plant), 24 dpi "older" leaves (dark green, fully expanded leaves), 65 dpi senescent leaves (brownish/yellowish leaves showing the signs of aging), 36 dpi flowers, 65 dpi pods, 24 dpi nodules, and 40 dpi nodules were extracted and RT-PCR was performed as described in Materials and Methods.

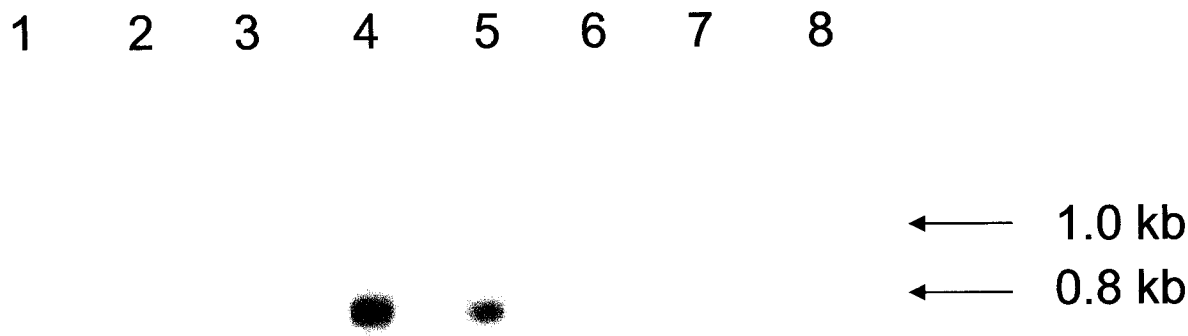
The primer pair 7/2 F 3 and R, predicted to give a product of 803 bp, was used. Since total mRNA preparations can be contaminated with genomic DNA, even following treatment with DNase I, the forward primer was designed to span first intron and consists of 11 bp from the 3' end of exon 1 and 8 bp from the 5' end of exon 2 (Table 3 and Figure 3 in Materials and Methods). In addition, any amplification of genomic DNA would include the sequences from introns 2,3,4, and 5 and the predicted size of the PCR product would be 1,765 bp, which is larger than the 803bp expected for the cDNA product.

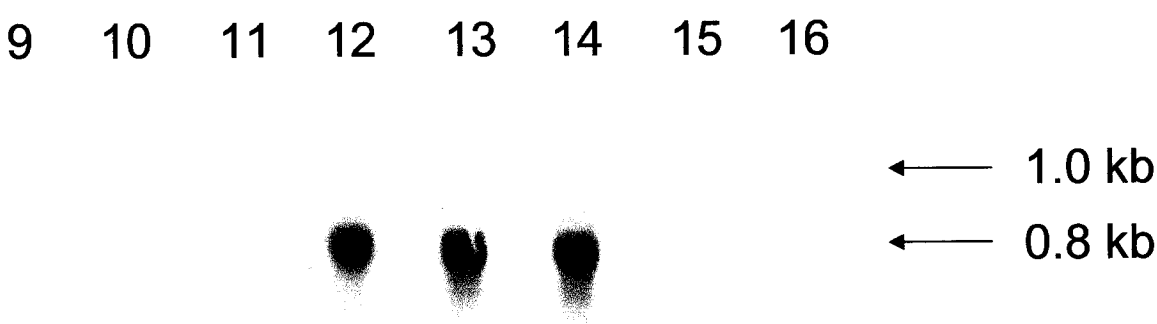
Tissues used in these experiments were collected from two individual sets of plants (sets A and B) and total RNA was extracted. Following conversion to cDNA by reverse transcriptase, PCR was performed twice on each cDNA sample. As controls a

PCR reaction with no DNA added and a reaction with the cloned 7/2 gene were included. Controls containing 1pg, 10pg and 100 pg of the 7/2 cDNA clone were also included. Following gel electrophoresis, the products were processed for Southern hybridization. Figure 10 presents the best result in that no contamination in controls was seen after 2 hours of exposure to BioRad Imaging Screen-K. The other three replicates, presented in the Appendix, show evidence of contamination with the presence of a band with the expected cDNA size (0.8 kb) and a band with the genomic size (~1.7 kb) on gDNA control. To confirm that the PCR product was indeed the expected sequence, we sequenced several clones arising from cloning of the 803bp band into pGEM-T Easy. All 4 sequences obtained were the expected cDNA sequence.

Figure 10: The mRNA expression of 7/2 in various tissues (Set A1).

RT-PCR samples were separated on a 1.5 % agarose gel and processed for Southern hybridization as described in the Materials and Methods. The probe was full length 7/2 cDNA. The sizes are indicated by arrows. Lane 1, seedlings; lane 2, 7 dpi cotyledons; lane 3, 24 dpi roots; lane 4, 24 dpi nodules; lane 5, 40 dpi nodules; lane 6, 24 dpi stems; lane 7, 24 dpi younger leaves; lane 8, 24 dpi older leaves; lane 9, 65 dpi senescent leaves, lane 10, 36 dpi flowers; lane 11, 65 dpi pods; lane 12, 1 pg cDNA clone 133c; lane 13, 10 pg cDNA clone 133c ; lane 14, 100 pg cDNA clone 133c; lane 15, 100 pg gDNA clone 133g; lane 16, negative control.


1 2 3 4 5 6 7 8
1.0 kb
0.8 kb


9 10 11 12 13 14 15 16
1.0 kb
0.8 kb

Based upon the detection of a signal at the expected size of 803 bp (Figure 10, experimental set A1), 7/2 is expressed in almost all tissues. The expression is highest in nodules and it declines as the nodules age from 24 days post inoculation with *Rhizobium japonicum* to 40 dpi (compare lanes 4, and 5). Expression is lower in roots (lane 3), stems (lane 6), older leaves (lane 8), and flowers (lane 10), and even lower in seedlings (lane 1), cotyledons (lane 2), pods (lane 11) and senescent leaves (lane 9). In this experiment no signal in younger leaves (lane 7) was detected using conditions that could easily detect 1pg of the 7/2 cDNA control plasmid. There is no evidence of contamination (lane 15, genomic DNA clone; lane 16, blank). Thus the 7/2 mRNA is expressed in a wide variety of tissues at different times and is not nodule-specific. But it appears to be upregulated in nodules. This conclusion must be tempered because it is based upon on a single replicate. Other replicates showed the same relative expression patterns (see Appendix) but the controls showed evidence of contamination (see Discussion).

In addition to the expected 803 bp band for 7/2 RT-PCR product there is a second band at about 1.0 kb in those tissues in which the expression of 7/2 is higher. The origin of this product is investigated in section 3.4.2.

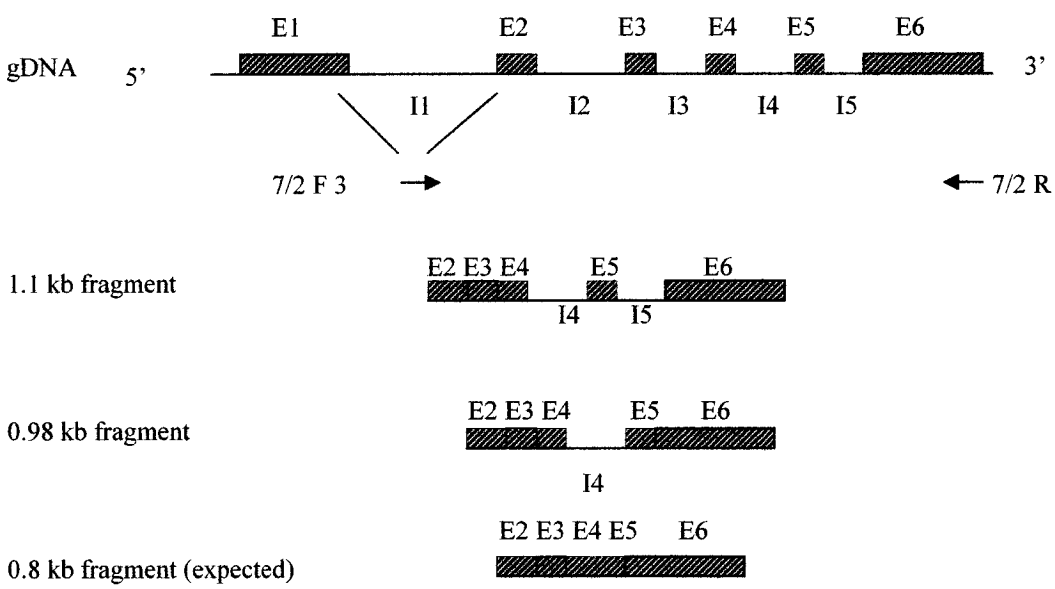
3.4.2 Partial Splicing

Although there is no band in the negative control (Figure 10, lane 16), many samples contain a band of ~1.0 kb in addition to the expected one of 803bp. This higher molecular weight band is present in many tissues and is especially prominent in those tissues where the expression of the 7/2 gene is highest.

To determine the origin of this fragment, the ~1.0 kb band arising from the nodule samples was gel purified and cloned in pGEM-T Easy. Several clones were purified and sequenced. Alignments of these sequences to the full length cDNA and gDNA sequences indicated that the size increase is due to the presence of introns in the PCR product. Of the 9 clones sequenced, 2 of them contained intron 4 and 3 of them contained both intron 4 and intron 5. Figure 11 is a schematic representing all the possible spliced mRNAs found in this experiment. The sequence and alignment results suggest that the ~1.0 kb band is due to the presence of splicing intermediates. The presence of sequences representing two types of splicing intermediate suggest that the agarose gel did not resolve the two molecular species.

Figure 11: Schematic showing the structures of the 7/2 gene and the RT-PCR products.

E refers to Exon and, I refers to Intron (sequences defined in Figure 5). The positions of the forward (7/2 F3) and reverse (7/2 R) primers on gDNA sequence are shown. The intron/exon composition of the RT-PCR products is also indicated.


E1
E2
E3
E4
E5
E6
gDNA
5'
3'
I1
I2
I3
I4
I5
7/2 F 3
7/2 R
1.1 kb fragment
E2 E3 E4
E5
E6
I4
I5
0.98 kb fragment
E2 E3 E4
E5
E6
I4
0.8 kb fragment (expected)
E2 E3 E4 E5
E6

CHAPTER FOUR

DISCUSSION

Two genomic DNAs and a cDNA sequence of 7/2 were obtained and compared during these set of experiments. The comparison of the two gDNA sequences from cv. Resnik and Maple Arrow showed identical nucleotides between these two cultivars and alignment of these gDNAs with the cDNA revealed the structure of the gene as having 6 exons and 5 introns. The conceptual 7/2 protein is similar to Myb-like transcription factors. The 7/2 message is expressed in most soybean tissues and is most highly expressed in nodules.

4.1 Analysis of 7/2 DNA Sequence

The sequence of 7/2 was assembled from the genomic sequences of cultivars Resnik and Maple Arrow, and also the cDNA sequence from cv. Maple Arrow. They were aligned and compared to find the gene structure and exon/intron boundaries. The cv. Resnik gDNA sequence is 4454 bp while the Maple Arrow sequence is 2726 bp. The Resnik clone was isolated from a λ library while the Maple Arrow clone was generated by PCR amplification of total soybean gDNA using the 7/2 F1 and 7/2 R primers designed according to the full length cDNA. The full length cDNA of 1383 bp, consisting of 417 bp of 5'UTR, and ORF of 798 bp (including stop codon), and a 3'UTR of 168 bp (excluding the poly "A" tail) was assembled *in silico* and the true 5' end of the 7/2 mRNA has not been mapped. The alignment of the gDNAs with cDNA revealed that 7/2 gene contains 6 small exons and 5 introns (Figure 4).

Initial sequence alignments identified nine nucleotide differences between the gDNA sequences of the 2 cultivars, mostly located in introns. Just two of them were in the coding region and could potentially change the protein sequence (Table 1 in Appendix). These mismatches could have been due either to true single nucleotide polymorphisms between the two cultivars or errors introduced during reading/editing of the sequence (sequence calls) or errors introduced by the Superscript II reverse transcriptase (RT) or Taq polymerase during the generation of the RT-PCR or PCR products used in cloning. In order to distinguish between these possibilities we tried several approaches. As described in Results, attempts to use thermostable polymerases with lower error rates such as Platinum *pfx* DNA polymerase and High Expand Fidelity PCR System were unsuccessful. Repeated sequencing of existing clones and cloning more independent clones, generated by RT-PCR or gDNA PCR, resolved the problem of mismatches. This indicated that the mismatches were due to errors introduced during PCR reaction. As can be seen from the final alignment (Figure 5), both genomic sequences of 7/2, from cv. Resnik and cv. Maple Arrow, are identical throughout the studied region and the cDNA from cv. Maple Arrow is identical to the exon sequences. The 100% identity of 7/2 gene in two cultivars with different genealogy and no recent common ancestor (Maple Arrow = Harosoy/PI 438477; Resnik = A31274/Williams 82) is surprising, and may indicate that there is selective pressure on the gene. A more complete ancestry is presented in Figure 4 in Appendix.

Table 1 in the Appendix shows the results of the alignment of the gDNA contigs with cDNA and the initial mismatches between them and their resolution, including the

position of each nucleotide that differed between the contigs and the number of times each clone was sequenced respectively.

To identify similar genes to 7/2 a BLASTN search for ESTs in the soybean database through the Center for Computational Genomics and Bioinformatics (CCGB) was performed. An EST contig with identity of 406 out of 407 nt was identified. This EST contig is made of two sequences from cDNA libraries of somatic embryo and seedlings of *Glycine max* cvs Jack and Clark respectively. The search performed at The Institute for Genomic Research (TIGR) identified the same soybean contig. The same search also identified several EST matches in other species which the highest match for each one are: *A. thaliana* (TC269957, transfactor like protein, 71 % identity), *Oryza sativa* (TC253755, unknown, 63 % identity), *Medicago truncatula* (TC81051, unknown, 62% identity) and *Zea mays* (TC262586, Phosphate starvation response regulator-like protein, 70 % identity). This analysis shows that there are related genes with some degree of similarity to 7/2 in other plants than soybean.

4.2 7/2 Conceptual Protein

The Myb DNA binding domain is a conserved region found in many transcription factors in eukaryotes and usually consists of two repeats designated R2, and R3 in plants. Each repeat is about 50 amino acids long and contains three helices forming a helix-turn –helix DNA binding motif (Martin and Paz-Ares, 1997). Many transcription factors have been defined that contain a Myb DNA binding domain with only one repeat such as mybSt1 in *Solanum tuberosum* (Baranowskij *et al.*, 1994), Circadian Clock Associated (CCA1), Late Elongated Hypocotyl (LHY) or CAPRICE (CPC) from *Arabidopsis* (Jin

and Martin, 1999; Wada *et al.*, 1997; Carre and Kim, 2002) all of which were discussed in the Introduction.

The longest Open Reading Frame (ORF) of 7/2 cDNA translates to a protein with 265 amino acids and a calculated molecular weight of 29.98 kDa. According to the BLASTP search there is a conserved region at the N-terminus from amino acids 23-74 similar to the Myb DNA binding domains (Figure 7). The number of conserved amino acids in 7/2 protein (52) in the Myb domain shows that this protein contains a single Myb repeat. Following removal of amino acids 1 to 74 from 7/2 protein which eliminates the conserved Myb domain, another conserved region was identified, by BLASTP search, from amino acids 120-163. This region of the protein contains acidic residues and is rich in glutamine which is characteristic of activation domains found within transcription factors (Mitchell and Tjian, 1989; Courey and Tjian, 1988; Wykoff *et al.*, 1999).

It is sometimes possible to infer function by the identification of proteins that are similar to the protein of interest and that have a known function. Amino acid sequences similar to the 7/2 myb motif were found in a number of proteins in databases mostly in unknown or hypothetical proteins or as Myb-related transcription factors with unknown function. Similarity was usually restricted to the myb domain.

One of the most similar protein to 7/2 is a Myb family transcription factor from *A. thaliana* (Accession no. NP_199371.1) with 95% identity over 63 amino acids at DNA binding domain and 74% identity over 46 amino acids in the putative activation domain. Another protein similar to 7/2 is transfactor from *Nicotiana tabacum* (Accession no. BAA75684) with a Myb-like DNA-binding domain. Amino acids 21-74 of 7/2 have 68% identity and 96% identity plus similarity with amino acids 72-125 of the Transfactor and

amino acids 120-161 of 7/2 have 71% identity and 98 % identity plus similarity with amino acids 159-200 to this protein. In order to search for proteins with a high degree of similarity to 7/2 the myb domain was removed and re-queried to the data base with the region containing amino acids 75 to 265. High level of similarity between 7/2 protein at positions 18-76 and 116-163 with a putative calcium dependent protein kinase substrate protein (CDPKS, Accession no.AAP45171) and a putative phosphate starvation response regulator (PSRR, Accession no.AAP45156), both from *Solanum bulbocastanum* was observed (Figure 8). These proteins are about the same size as 7/2 and contain a Myb DNA-binding domain. Experimental evidence for the precise function of all these proteins is not available.

CSP1 (Accession no. AAF32350) from the ice plant, *Mesembryanthemum crystallinum*, is a transcription factor which has been experimentally characterized with lower level of similarity to 7/2 protein compared to previous proteins (Figure 9). Experiments on CSP1 have shown that it is a transcription factor belonging to pseudo-response regulators and is a substrate for McCDPK1 (Patharkar and Cushman, 2000). Typical bacterial two component response regulators have an N-terminal phosphate receiving domain, a basic helix-loop-helix DNA binding domain and a C-terminal transcriptional activation domain (Lohrmann *et al.*, 1999). CSP1 protein is similar to plant pseudo-response regulator transcription factors which lack the conserved D-D-K in their phosphate receiving domain compared to their bacterial counterparts (Makino *et al.*, 2000). Alignment of CSP1 with ARLP1 which is a member of plant pseudo-response regulators shows differences in the N-terminus region but conservation of the DNA binding domain (Patharkar and Cushman, 2000). Regardless of these differences there are

many serine/threonine residues in CSP1 that are potential phosphorylation sites for CDPKs and the results from *in vitro* phosphorylation experiment proved that CSP1 is a substrate for McCDPK1 (Patharkar and Cushman, 2000). The precise amino acid(s) that is phosphorylated is unknown.

CSP1 has a SV40-like nuclear localization signal at its C-terminus and by fusion to GFP it was shown that it co-localizes in the nucleus with McCDPK1 (Patharkar and Cushman, 2000). ARLP1 from *Arabidopsis* is a homologue of CSP1 which has two SV40-like NLS. In a subcellular localization experiment ARLP1 polypeptide with two, one or no NLS motifs was fused to GFP and expressed in parsley protoplasts. These experiments showed that either SV40-like NLS acting alone is sufficient to direct the protein to the nucleus (Lohrmann *et al.*, 1999). To find an NLS in 7/2 protein, its sequence was manually compared to known NLS sequences such as the SV40-like NLS or the Bipartite NLS (LaCasee and Lefebvre, 1995). In addition, programs such as PROSITE, at http://cubic.bioc.columbia.edu/predictNLS, or PSORT, at http://psort.ims.u-tokyo.ac.jp/form.html, were used. PSORT but not PROSITE could detect a putative NLS in ARLP1 and CSP1. Neither detected an NLS on 7/2. Thus the issue of an NLS in 7/2 remains unresolved.

As a potential transcription factor, 7/2 needs to get into the nucleus to function. If it lacks an NLS, it still might be translocated via association with another protein(s). Although there is no NLS on the Psr1 protein, which is a transcription factor, immunocytochemical studies revealed that the protein is localized to the nucleus (Wykoff *et al.*, 1999). APETALA-3 (AP3) and PISTILLATA (PI) are transcription factors responsible for the initiation of petal and pistil formation in *Arabidopsis* (Goto and

Meyerowitz, 1994). They are translocated into the nucleus as heterodimers and it has been proposed that such an association is sufficient to allow the translocation of proteins lacking an NLS into the nucleus via association with a protein containing an NLS (Schwechheimer and Bevan, 1998). Using the program PSORT, a NLS was detected in AP3 and not in PI.

The subcellular localization of the 7/2 protein can experimentally be determined by fusion of the protein to GFP and introduction of the construct into a suitable host cell (Patharkar and Cushman, 2000; Lohrmann *et al.*, 1999; Makino *et al.*, 2000).

PROSITE identified several motifs on 7/2 protein which can be potential sites for posttranslational modification such as glycosylation, myristoylation, and phosphorylation sites. Glycosylation (Ohnishi *et al.*, 2001) and myristoylation (Maurer-Stroh *et al.*, 2004) are important to modulate the activity, folding and interaction of proteins with other proteins. One of the most important regulatory mechanisms in all organisms is signal transduction mediated by phosphorylation/dephosphorylation by protein kinases/protein dephosphorylases (Hardie, 1999). According to previous studies the function of transcription factors such as DNA binding, trans-activation and sub-cellular localization are usually regulated by phosphorylation and dephosphorylation (Hunter, 1995; Hunter and Karin, 1992; Meshi *et al.*, 1998). In the case of c-Myb proteins it has been shown that phosphorylation can modulate transcriptional activity and DNA binding (Vorbrueggen *et al.*, 1996). Members of the plant Myb family also contain several serine or threonine residues which can serve as potential sites for kinases and affect the activity of protein (Martin and Paz-Arez, 1997). The DNA binding affinity of AmMYB340 but not AmMYB305 from *Antirrhinum majus* was reduced due to phosphorylation (Maoyano *et*

al., 1996). 7/2 protein contains several protein kinase phosphorylation sites, consistent with its role as a potential regulatory protein.

Beside CSP1 another Myb protein with known function is Psr1 from *Chlamydomonas* which is responsible for phosphorus metabolism and has a glutamine rich activation domain characteristic of transcriptional activators (Wykoff *et al.*, 1999). Psr1 and 7/2 are 74% identical in the DNA binding domain and 43% in the putative activation domain.

In c-Myb proteins with three repeats of R1-R2 and R3 in the DNA binding domain each repeat contains three conserved tryptophan residue that are separated from each other by 18-19 amino acids (Anton and Frampton, 1988). Each repeat forms a helix-turn-helix structure which the first repeat (R1) is not involved in DNA recognition and binding. NMR spectroscopy has revealed that conserved tryptophans (trp,W) in the Myb repeat 3 contribute to the formation of a hydrophobic DNA-binding structure and the last helix in R3 is the recognition helix which makes specific contact within the major groove of the DNA double helix (Ogata *et al.*, 1992; Ogata *et al.*, 1994). The presence of conserved tryptophan residues and formation of the helix structure in each repeat is applicable to all known Myb proteins in plants and is conserved among all eukaryotes (Romero *et al.*, 1998; Kranz *et al.* 2000). ZmMybst1 from maize (a homologue of Mybst1 from *Solanum tuberosum*) has a single Myb DNA binding domain from amino acids 87-140 and there are two conserved trp (W) in this protein sequence at positions 94 and 114 and the third trp (W) is replaced with alanine (A) at position 131 (Mercy *et al.*, 2003). According to Mercy *et al.* trp (W) can be substituted with other hydrophobic amino acids as it was observed with a few other Myb domain containing proteins (Mercy

et al., 2003). Alignment of 7/2 with the DNA binding domain of ZmMybst1 shows that 7/2 contains only one conserved trp (W), at position 26.

In the case of the MybSt1 protein which encodes a DNA binding domain with a single repeat of 61 amino acids containing a conserved motif of SHAQKYF at the C-terminus of the binding domain, it was shown that this protein is able to bind to GGATA-containing sequences on DNA in a sequence specific manner (Baranowskij *et al.*, 1994). A few other Myb proteins are conserved in the SHAQKYF motif such as CCA1 (NP-850460), LHY (XP_480189), ZmMRP-1(CAC86577), ZmMybst1 (AAO47339) and LeMYB1 (CAB65169).

The SHLQKYR motif at the C-terminus of the DNA binding domain of 7/2 is identical to amino acid sequences found in the Myb DNA binding domains of Transfactor (Accession no. BAA75684), Psr1 (AAD55941), ARR1 (BAA74528), ARR2 (BAA74527), ARLP1 (CAA06431), CDPKS (AAP45171), PSRR (AAP45156), and CSP1 (AAF32350).

As the last α helix of Myb domains has been shown to be involved in sequence specificity of DNA binding these results illustrate that the SHLQKYR region might resemble the DNA contact portion of 7/2 protein and contributes to the recognition of different DNA sequence elements and differences within the Myb domains are responsible for the different DNA binding specificities of each protein and distinct physiological function.

Alignment of 7/2 Myb DNA binding domain with known single Myb domain proteins such as MybSt1 (Accession no. AAB32591), CCA1 (NP-850460), LHY (XP_480189) and ZmMRP-1 (CAC86577) showed less than 30% identity. Due to the

similarity of the SANT domain with Myb DNA binding domain in regard of the three repeats and helix formation (Asland *et al.*, 1996) BLASTP search picks the conserved region of DNA binding domain in 7/2 and CSP1 as well as all these proteins as a region similar to the SANT domain beside the Myb domain. According to all the information obtained from 7/2 protein sequence such as the presence of DNA binding domain, putative activation domain and phosphorylation sites, and also the information obtained from known transcription factors it is assumed that the 7/2 gene codes for a transcription factor and 7/2 protein is a regulatory protein. The low level of similarity between 7/2 protein and other Myb containing proteins might be due to the specificity in their DNA binding and functions.

Myb transcription factors can also bind to other regulatory proteins. Myb C1 interacts with R protein in maize to regulate anthocyanin pigment biosynthesis (Goff *et al.*, 1992). This observation may provide a route to infer their function, if hypothetical binding partners can be isolated and identified.

4.3 Expression of 7/2

Infection of a legume such as soybean with a specific bacterium such as *Bradyrhizobium japonicum* leads to the expression of many genes implicated in nodule development and biochemistry, i.e., the formation, function or ageing of the nodules (Nap and Bisseling, 1990). Differential screening of a nodule cDNA library constructed from soybean cv. Maple Arrow grown for 40 days post inoculation led to the isolation of 7/2. Initial experiments on the expression of 7/2 in nodules suggested that it is a late nodulin (data not shown) (Nap and Bisseling, 1990). To further investigate the expression pattern of 7/2, the presence of 7/2 mRNA was examined in different tissues of soybean and it

determined that the expression is not restricted to nodules (Section 3.4.1), although the expression is higher in nodules than any other tissue tested. If 7/2 is a transcription factor, its expression in different tissues suggests that it may regulate the expression of other genes not involved in nodulation or alternatively it may regulate the same genes in all tissues but these genes are more highly expressed in nodules. The gene enod40, found in a variety of legumes and non-legumes is an example of a gene that is expressed in many tissues but more strongly expressed in nodules (Rohrig *et al.*, 2004). It functions to stimulate cortical cell division in legumes (Cohn *et al.*, 1998) and it is proposed that it alters sucrose utilization in nodules (Rohrig *et al.*, 2004). Its presence in non-legumes such as tobacco (Sande *et al.*, 1996) and rice (Kouchi *et al.*, 1999) indicates a more general role in plants as a growth regulator that alters phytohormone responses.

The expression pattern of 7/2 also shows a decrease during nodule development (from 15 dpi to 40 dpi, see Figures 1-3 in Appendix) which is an indication of more functionality in earlier stages of nodule development and formation.

The expression and regulation of functional genes governs plant development, cellular differentiation, and response to environmental stimuli. Many forms of regulation are transcriptional and require both cis-acting DNA sequences and trans-acting regulatory proteins. The DNA sequence 5' to the gene contains core promoter elements that determine the basal transcription activity of the gene and usually direct the positioning of the transcription initiation start site (Brown, 2002). TATA boxes and initiator elements that have been identified in plant promoters are the sequences recognized by RNA polymerase II and sequence-specific binding of transcription factor IID (Grace *et al.*, 2004). In eukaryotes the TATA box is usually located ~ 30 bp upstream from the

transcription initiation site with a consensus sequence of TATA(T/A)A(T/A) (Zhu *et al.*, 1995), yet eukaryotic RNA polymerase II promoters can stretch hundreds of base pairs upstream from the transcription start site (Brown, 2002).

To search for the promoter on 7/2 DNA the 5' end of cv. Resnik gDNA with 1761 bp upstream from the translation start site "ATG" was queried in PLACE data bases (Plant Cis-acting Regulatory DNA Elements) at http://www.dna.affrc.go.jp/PLACE/. According to this computational analysis several potential TATA and CAAT boxes were recognized on this genomic sequence. Figure 12 shows the 5' UTR of 7/2 as far as translation initiation (ATG). Although the 5' end of the message has not been mapped, the longest cDNA sequence identified by 5' RACE experiments would put transcription start site near position 1345 on this figure (shown by arrow). There are five TATA-like boxes upstream from this region, indicated in boxes that are potential promoter sites due to their proximity to the putative site for initiation of transcription. There are others that are less likely because they are too far away from this site. These putative promoters are underlined in the figure.

There have been reports of the presence of multiple TATA boxes in the promoter region of functional genes. As an example, two TATA boxes are present in the soybean tubulinB1 promoter at positions -122 to -117 and -35 to -30. These two TATA boxes function additively to direct transcription in seedlings as shown by site directed mutagenesis. They are also differentially sensitive to light conditions and so might provide a mechanism for titrating gene activity in response to altered environmental conditions (Doyle and Han, 2001). CAAT boxes also have been identified as promoter elements upstream of the transcription start site in eukaryotes including plants and

CAAT-binding factors, CBF, play important role for interacting with this element and its activity for gene expression (Fickett and Hatzigeorgiou, 1997).

The PLACE program also identified a very large number of sequences that potentially could bind transcription factors that may regulate 7/2 expression including a core motif that binds MybSt1 or a core binding site of rice WRKY71, amongst others. The significance and functionality of these is unknown.

These data show that the region in 5' to the 7/2 gene contains possible promoter elements required for initiation of 7/2 transcription and provides a possible experimental route to directing the roles of these elements via mutagenesis and transformation.

Figure 12: 5' region of the 7/2 genomic DNA.

This region includes the 5' upstream region, showing the presence of putative promoter elements as designated by the PLACE program. Putative TATA boxes near the putative initiation of transcription are indicated in boxes while others that are more distant are underlined. CAAT boxes are underlined. The start of the conceptual ORF (ATG) is in red. The start of the most 5' cDNA sequence as determined by 5'RACE is indicated by arrow.

1 TATCAGCTCTTCACGGACTAWCAATCCTTAAACCAAGCATGCTACTATCTTGCATAGTTG

61 GACACTATTTTTAGTAGTTAAAATTTCATCGTGATGATCCAACTAAAAATAAGTTAACAC

121 TACGTACACTTGGAGATTAAAATATGAAATTATAAAACGAAATTATAATTTTAACACTTT

181 TATAATATTATTATATATAATATTTCATTACAATTTTTTTTTTAAATTGCATATTTTATT

241 ACAATTTCAATTATTGATTTTAAAATAAATGTTTGCAATTTCGAGTTTTAACATATTCCT

301 CACTTTACTGATAAAAAAACTTATTCCTCACTTTAAAAGGATTAAAGTTCAATAAAAAAA

361 ATTGCATAAAATGGTTCCTAAAATATTAGAATGGATATATGTTTAAATAATTAGTTGGGA

421 ACAGCTTCTATGTTATTGGACCACCTTATACAAAATTCTCTGCCTAAATTTAACATAGTA

481 GTTGGGATAGAACAGAACAGAACATTATTTTGTTCTTTGTCACACATTTGTTTTTATAAA

541 AATCATTATGTACCATTTTTTTTTATAATTTCATTCATGTTTTATTTTATTTATAAACAT

601 GTTAAGTATAAAATAAAATTTATATTTCTTTATTCACTTTCTGTTAAAAGGGGGTCAATG

661 ACAATGGGCCATCGAGAGATTTTACTAAATAGTTTTTTTTTAAAAAAAATAAAAATTATC

721 AAAGATAGATAGAATTTATCTTCTTATAGTCCATTCCTCAGCCGGAGAAGGAATATTCTT

781 CCGTCAGCATCATTAACGGACAAACCGAAACGTGTTCATTTGGTGGGCCCCACTTCTCTC

841 TCCCTCCAATCGAAATCCATGTGACCCTCTAACTCTAACGTGACCTTCCCTTTTGGACTC

901 TCATCATTCATTTCTCTTAGATCCGAATCTTCCACGCTGGCATCCACGTGTCCCCCATTC

961 CATAACACACACCGTTCGATCCTCCCTAACAAACTCACTTATTCCGATCCAACGGACACA

1021 ACGAGCACCTTAAGTCTTTGAATTTCATCTCCCCGATTAAAAAACTTGCTATCCACGATG

1081 TTACTTCGAATTCGAATTCGTCGCCGAGATTTCCTCTAACATGTTCCTCAAACTAACTAT

1141 TCTAGCTGCTAATTTCTACCATGCTCATATTAATATTAATATCATATCATATCATCTACA

1201 TGCATAATGCATTGCACCTAAACATGCCACATACTTAATTATTTATCATTAGATACTAAA

1261 TAATTGAACCACTGTTACAAAATAGCATTATATAAGTCCATGAAACGCCGTATATAATGC

1321 TTTTGTTATTATTATTAGCGTTGCTTCTTGCGCCATTGAGCCATTTTCCTTTTTTCAGCA

1381 GCACTAGTTCCGTGTTTGTGTCTAATCAATGATGTTGTGTTGTGCTCTGTTTTCCTCTGA

1441 CTTGTATGGTTTAGTTCCAATAGGTTATGATTGCTAACAATCAGAATCAGAGTGTGTTGG

1501 GGAATTAGAGAATCTTCTTCCATTAGCAAAAACACCGTGACTAGTTGCAGAGGGTAAATG

1561 GTATATCTTATTGTGGACTACGTACGAGGATCCCTTTTCTTCTTTTCCGCTTTTGGGATA

1621 GAATGAAGGCGCAGTTTGGCAGAGAGGTATAAAGCAAAAGCTCCAAGTGGGAATTGTTCG

1681 AATTTGCTTTTTCTAAGTTGAAAAGAAAACAAGGGTAGCTGAAGCCTTGAAGGTGAACAA

1741 GTGTAGTTTGGGAGGTGAGTGATG

4.3.1 Tissue Distribution

Attempts to do RT-PCR in a semi-quantitative reaction by using 18 S rRNA as an internal control failed. Non-quantitative experiments designed to measure the expression of 7/2 in different soybean tissues showed that 7/2 is expressed in a wide variety of tissues (Figure 10).

The expression is highest in nodules and it declines as the nodules age from 24 days post inoculation with *Rhizobium japonicum* to 40 dpi (compare lanes 4, and 5). Expression is lower in roots (lane 3), stems (lane 6), older leaves (lane 8), and flowers (lane 10), and even lower in seedlings (lane 1), cotyledons (lane 2), pods (lane 11) and senescent leaves (lane 9). In this experiment no signal in younger leaves (lane 7) was detected. There is no evidence of contamination (lane 15, genomic DNA clone; lane 16, blank). Although RT-PCR is not quantitative in this experiment, the amount of product can be estimated by comparison to dilutions of the standard cDNA clone (from 1 to 100pg, depending on the replicate). The RNA used in RT-PCR reactions was normalized by reference to 18 S rRNA as described in Materials and Methods. Given these caveats, the results suggest that there are between 1-10 pg of 7/2 message in nodules and lower than 1 pg in other tissues.

The controls that were used in this experiment were a PCR reaction with no DNA added and a reaction with the 100 pg of cloned 7/2 gene. The primer 7/2 F3 that was used for amplification is complementary to the 3' end of exon 1 and the 5' end of exon 2, and spans the first intron. This primer was designed to prevent the amplification of any contaminating gDNA in the RNA samples (Table 3 and Figure 3 in Materials and Methods). Figure 10 shows the tissue distribution of set A1 plants which is the best

result. In this replicate there is no sign of contamination in lanes 15 and 16 representing the gDNA standard and a sample with no DNA respectively. But there is an extra band in some of the tissues at about 1.0 kb. This represents partially spliced products, discussed in Section 4.4.

In the other three replicates shown in the Appendix, sets A2, B1 and B2, there is no band in the PCR sample with no DNA added (lane 17) but 2 bands in gDNA control (lane 16). The size of the first band is the same as the expected size of 0.8 kb for cDNA and the second band is ~1.7 kb which represents the genomic size. Sequence alignment of the primer 7/2 F3 with the exon I/intron I and also intron I/exon II junction, shows that this primer can only bind to the exon sequences as designed and there is no significant binding to the introns (data not shown). The calculated Td for this part of the primer is about 26°C. The annealing temperature used for PCR amplification was 64°C which is much higher and should eliminate the possibility of this priming. Although the sterilized filter tips were used and the pipetters were washed, and fresh aliquots of the gDNA clone were used in different PCR reactions, contamination was repeatedly observed. It is unlikely due to a problem with gel loading as a separate sterilized filter tip for each sample was used and an empty lane was left between samples to eliminate spill over. The possibility that the gDNA clone stock was contaminated is not eliminated although this should be checked by remaking the plasmid.

In addition to the contamination problem the expression of 7/2 in the tissues varies between replicates. Two more nodule samples, 15 dpi nodules and 20 dpi nodules were added in replicates of set A2, B1 and B2 to look for 7/2 expression at different time points of nodulation . As it is clear in these replicates and specially in set B1 and B2, the

expression of 7/2 decreases as the nodules grow and comparisons of 15 dpi to 40 dpi samples show that the message is more highly expressed in the earlier stages of nodule development although it remains high in 40 dpi compared to non-nodule tissues. The differences in other tissues of the replicates may be due to the differences in the growth condition or even during the PCR amplification of the message. In general according to 1 pg and 10 pg cDNA controls 7/2 expression in nodules is approximately within this range and is lower than 1 pg in other tissues such as stems, roots, seedlings, flowers, younger leaves, and is very low in pods, cotyledons, older leaves and senescent leaves. There is a slight difference between the replicates. For example there is no band for younger leaves in set A1 or no band for stems and older leaves in set B1 which as was mentioned could be due to the PCR amplification. The low amount of 7/2 transcript is in accordance with the previous studies indicating that transcription factors expression is lower than other genes with a different level of expression between cells of different tissues and organs (Czechowski *et al.*, 2004).

The EST contig described in Section 4.1 suggests that 7/2 is found in seedlings and somatic embryos. We also found it in seedlings but did not test somatic embryos. The present results suggest that 7/2 is not nodule specific as it is expressed in other tissues. But the higher level of expression in nodules might indicate more functionality in nodules. To achieve reproducibility and eliminate contamination this experiment needs to be repeated to obtain convincing results.

4.4 Detection and Analysis of Partially Spliced 7/2 mRNA

While carrying out the RT-PCR experiments to determine the pattern of expression of 7/2 mRNA, a higher molecular weight band of about 1.0 kb was unexpectedly observed in addition to a band of 803 bp expected with the primers 7/2 F3 and 7/2 R. When this amplimer was cloned, and clones were sequenced, it was observed that the increase in size was due to the incomplete splicing of the message. Several clones had retained intron 4 and some of the sequences had retained both introns 4 and 5 (Figure 11). The addition of intron 4 adds 184 nucleotides and results in a fragment of 987 bp. The fragment with both introns 4 and 5 has 309 more nucleotides and the total size will be 1112 bp. The resolution in the agarose gels used to detect the RT-PCR products was insufficient to resolve the two molecular species into two bands.

Thus partially spliced transcripts of 7/2 message have been detected. This extra band is only observed in nodules and in a few other tissues such as younger leaves, older leaves, and to a lesser extent, flowers and seedlings (Figure 10 in Results and other replicates in the Appendix). The expression of the fully-spliced product appears to be higher in these tissues as well, suggesting that the detection of the partially spliced products is a function of the mRNA concentration. When more mRNA is detected the level of detectable intermediate is higher. None of the clones contained introns 2 or 3. We do not know whether this indicates lack of sampling or whether these introns are removed earlier in splicing. The forward primer was designed to span the first intron and consists of 11 bp from the 3' end of exon 1 and 8 bp from the 5' end of exon 2. Therefore we do not know the splicing status of intron 1. Several models have been put forward to explain the order of intron removal, e.g., transcript shortening from the 5' or 3' ends and exon

skipping. These data do not directly address this question but suggest that introns 2,3, and 5 are removed before intron 4.

Alternative splicing is a major source of protein diversity from the genome and occurs in more than 35% of human mRNAs (Graveley, 2001). 7-10% of Arabidopsis genes are alternatively spliced (Ner-Gaon *et al.*, 2004). In the case of 7/2, splicing intermediates have been documented but we have no evidence whether they represent simply intermediates in a process that removes all the introns or whether they represent alternative forms or a combination of both ideas. Translation of these two transcripts result in the same truncated protein due to the presence of an in-frame stop codon in intron 4. This hypothetical 132 aa protein would contain the first 116 aa of the Myb protein plus 16 aa encoded by the intron before the stop codon. This putative truncated protein contains putative Myb DNA binding domain. Database similarity searches such as BLAST P retrieve proteins with Myb–like DNA binding domains as recovered previously (Results, section 3.3). Searches with the Myb domain removed do not identify accessions with any significant similarity, thus there is no evidence that the partially spliced form (s) is functional.

4.5 Conclusions and Future Work

A novel cDNA clone (7/2) from a 40 dpi soybean nodule cDNA library cv. Maple Arrow was analyzed during these set of experiments. The full length cDNA contains an ORF of 798 bp encoding a novel protein with molecular weight of 29.98 kDa and 265 amino acids which contains an N-terminal single repeat Myb DNA binding domain with SHLQKYR motif as a recognition motif at the C-terminus of its DNA binding domain and a putative activation domain characteristic of transcription activators. The alignment

of 7/2 gDNA sequence from cv. Resnik and Maple Arrow showed conservation of nucleotides and exon/intron boundaries between the cultivars. The 7/2 genomic sequence contains 5 introns and is a single copy gene which its message is expressed almost in all tissues but with higher extent in nodules. Two partially spliced mRNAs were detected during RT-PCR experiment which encode a truncated protein with no evidence of functionality.

Several lines of investigation are necessary to advance this work:

1) The ~1.7 kb fragment obtained on the genomic DNA control in the tissue distribution experiment needs to be cloned and sequenced in order to determine its origin.

2) To understand more about the expression of 7/2 and its protein sequence experimentally, complementary DNA sequence of 7/2 can be fused in frame to Glutathion S-transferase (GST) in an expression vector such as pGEX at appropriate restriction site and expressed in *E.coli* as was shown for other plant proteins (Serra *et al.*, 1993). The fusion protein can be purified to make an antibody for further detection.

3) Primer extension can be performed on soybean RNA to confirm the transcription start site which was predicted by 5' RACE and identify promoter regions more precisely (Baldwin and Gurley, 1996). Once this has been mapped promoter deletions could be constructed.

4) To investigate the subcellular localization of the 7/2 protein, a GFP fusion could be constructed and introduced into tissue culture cells by transformation. The expressed protein could be located by fluorescence microscopy (von Arnim *et al.*, 1998).

5) To determine whether the 7/2 interacts with other proteins as would be inferred for a transcription factor, the yeast two hybrid system can be employed. The putative

activation domain of 7/2 cDNA could be used as a bait to identify such proteins (Chien *et al.*, 1991).

6) Quantitative RT-PCR can be performed for measuring the amount of 7/2 mRNA in each tissue or at any stage of development (Riedy *et al.*, 1995).

REFRENCES

Adam E, Szell M, Szekeres M, Schaefer E, and Nagy F. (1994). The developmental and tissue-specific expression of tobacco phytochrome-A genes. Plant J. 6: 283-293.

Aida M, IshidaT, Fukaki H, Fijisawa H, Tasaka M. (1997). Genes involved in organ separation in Arabidopsis: an analysis of the cup shaped cotyledons mutants. Plant Cell 9: 841-857.

Albert HA, Martin T, and Sun SSM. (1992). Structure and expression of a sugarcane gene encoding a housekeeping phosphoenolpyruvate carboxylase. Plant Mol. Biol. 20: 663-671.

Albrecht C, Geurts R, and Bisseling T. (1999).Legume nodulation and mycorrhizae formation; two extremes in host specificity meet. EMBO J. 18(2):281-288

Anderson JD, Lowary PT, and Widom J. (2001). Effects of histone acetylation on the equilibrium accessibility of nucleosomal DNA target sites. J.Mol.Biol. 307: 977-985.

Anton IA, and Frampton J. (1988). Tryptophans in myb proteins. Nature. 336(6201):719.

Asland R, Stewart AF, and Gibson T. (1996). The SANT domain: a putative DNA binding domain in the SW1-SNF and ADA complexes, the transcriptional co-repressor N-CoR abd TFIIIB. Trends. Biochem. Sci. 21(3): 87-88.

Baldwin DA, and Gurley WB. (1996). Isolation and characterization of cDNAs encoding transcription factor IIB from Arabidopsis and soybean. Plant J. 1996 10(3):561-568.

Baranowskij N, Frohberg C, Prat S, and Willmitzer L. (1994). A novel DNA binding protein with homology to Myb oncoproteins containing only one repeat can function as transcriptional activator. EMBO J. 13(22): 5383-5392.

Berghammer H and Auer B. (1993) Isolation of plasmid DNA using the EasyPrep oiling Method. Biotechniques. 14: 524-525.

Berger SL. (2002). Histone modification in transcriptional regulation. Curr Opin Genet Dev.12: 142-148.

Beverley JG, Perez-Rodriguez M, and Martin C. (1998). Development of several epidermal cell types can be specified by the same MYB-related plant transcription factor. Development. 125(17): 3497-3508.

Bilaud T, Koering CE, Binet-Brasselet E, Ancelin K, Pollice A, Gasser SM, and Gilson E. (1996). The telobox, a Myb-related telomeric DNA binding motif found in proteins from yeast, plants and human. Nucleic Acids Res. 24(7): 1294-1303.

Boyer LA, Langer MR, Crowley KA, Tan S, Denu JM, and Peterson CL. (2002). Essential role for the SANT domain in the functioning of multiple chromatin remodeling enzymes. Mol Cell. 10: 935-942.

Boyer LA, Latek RL, and Peterson CL. (2004). The SANT domain: a unique histone-tail-binding module. Nature. 5:1-6.

Braun EL, and Grotewold E. (1999). Newly discovered plant c-myb-like genes rewrite the evolution of the plant myb gene family. Plant Phys. 121: 21-24.

Broin M, Cuine S, Eymery F, and Rey P. (2002). The plastidic 2-cysteine peroxiredoxin is a target for a thioredoxin involved in the protection of the photosynthetic apparatus against oxidative damage. Plant Cell. 14(6):1417-32.

Brown TA. (2002). Genomes, 2nd edition. John Wiley and Sons, INC., Publication.

Brown JWS and Simpson CG. (1998). Splicing site selection in plant pre-mRNA splicing. Annu.Rev.Plant Physiol.Plant Mol.Biol. 49: 77-95.

Buratowski S. (1994). The basics of basal transcription by RNA polymerase II. Cell. 77: 1-3.

Carre IA, and Kim JY. (2002). MYB transcription factors in the Arabidopsis circadian clock. J Experimental Bot 53(374): 1551-1557.

Cedroni ML, Cronn RC, Adams KL, Wilkins TA and Wendel JF. (2003) Evolution and expression of Myb genes in diploid and polyploid cotton. Plant Mol. Biol. 51, 313–325.

Chan CCY. (1995).The isolation and characterization of a senescence associated nodulin cDNA. University of Ottawa.

Chen JQ, Dong Y, Wang YJ, Liu Q, Zhang JS, and Chen SY (2003). An Ap2/EREBP-type transcription factor gene from rice is cold inducible and encodes a nuclear localized protein. Theor Appl Genet. 107(6); 972-979.

Cheung P, Allis CD, and Corsi PS. (2000). Signaling to chromatin through histone modifications. Cell 103: 263-271.

Chien C, Bartle PL, Sternglanz R, and Fields S. (1991). The Two-Hybrid System: A Method to Identify and Clone Genes for Proteins that Interact with a Protein of Interest. PNAS. 88: 9578-9582.

Chomczynski P, and Sacchi N. (1987) Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. Anal Biochem. 62: 156-159.

Chomczynski P, and Mackey K. (1995) Short technical reports. Modification of the TRI reagent procedure for isolation of RNA from polysaccharide- and proteoglycan-rich sources. Biotechniques. 19(6):942-945.

Church G and Gilbert W.(1984) Genomic sequencing. PNSA.81: 1991-1995.

Cohn J, Bradley DR and Stacey G.(1998). Legume nodule organogenesis. Trends in Plant Science: 3(3):105-110.

Cohen A, and Mayfield SP. (1997). Translational regulation of gene expression in plants. Curr Opin Biotechnol. 8: 189-194.

Conaway R, and Conaway J. (1997). General transcription factors for RNA polymerase II. Prog Nucleic Acid Res Mol Biol. 56:327-46.

Cook D, Dreyer D, Bonnet D, Howell M, Nony E and Vandenbosch K.(1995). Transient induction of a peroxidase gene in *Medicago truncatula* precedes infection by *Rhizobium meliloti.* The Plant Cell 7: 43-55.

Courey A, and Tjian R (1988). Analysis of Sp1 in vivo reveals multiple transcriptional domains including a novel glutamine rich activation motif. Cell, 55: 887-898.

Crespi M, and Galvez S. (2000). Molecular mechanisms of root nodule development. J. Plant Growth Regul. 19: 155-166.

Cullimore JV, Ranjeva R and Bono J. (2001).Perception of lipo-chitooligosaccharidic Nod factor in legumes. Trends Plant Sci. 6(1): 24-30.

Czarnecka-Verner E, Yuan CX, Fox PC, and Gurley WB. (1995).Isolation and characterization of six heat shock transcription factor cDNA clones from soybean. Plant Mol Biol. 29(1):37-51.

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

Dai S, Petruccelli S, Ordiz MI, Zhang Z, Chen S, and Beachy RN. (2003). Functional analysis of RF2a, a rice transcription factor. J Biol Chem. 278(38); 36396-36402.

Daniel X, Lacomme C, Morel J B, and Roby D. (1999). A novel myb oncogene homologue in *Arabidopsis thaliana* related to hypersensitive cell death. Plant J. 20(1): 57-66.

Day DA, and Tuite MF. (1998). Post-transcritional gene regulatory mechanisms in eukaryotes: an overview. J Endocrinol. 157: 361-371.

Denekamp M, and Smeekens SC. (2003). Integration of wounding and osmotic stress signals determines the expression of the AtMYB102 transcription factor gene. Plant Physiol. 132: 1415-1423.

Doyle MC, and Han IS. (2001). The roles of two TATA boxes and 3'-flanking region of soybean beta-tubulin gene (tubB1) in light-sensitive expression. Mol Cells.12(2):197-203.

Feinberg AP, and Vogelstein B. (1983) A technique for radiolabeling DNA restriction endonuclease fragments to high specific activity. Anal Biochem.132: 6-13.

Fickett JW, and Hatzigeorgiou AG. (1997). Eukaryotic promoter recognition. Genome Research. 7(9): 861-878.

Fischle W, Wang Y, and Allis CD. (2003). Histones and chromatin cross-talk. Curr Opin Cell Biol. 15: 172-183.

Gabrielsen OS, Sentenac A, and Fromageot P. (1991). Specific DNA binding by c-MYB: evidence for a double helix-turn-helix related motif. Science. 253: 1140-1143.

Goff SA, Cone KC, and Chandler VL. (1992). Functional analysis of the transcriptional activator encoded by the maize B gene: evidence for a direct functional interaction between two classes of regulatory proteins. Genes Dev. 6(5): 864-75.

Gomez E, Royo J, Guo Y, Thompson R, and Hueros G.(2002). Establishment of cereal endosperm expression domain: identification and properties of a Maize transfer cell-specific transcription factor, ZmMRP-1. Plant Cell. 14: 599-610.

Goormachtig S, Valerio-Lepiniec M, Szczyglowski K, Montagu MV,Holsters M and Bruijn FJ. (1995). Use of differential display to identify novel *Sesbania rostrata* genes enhanced by *Azorhizobium caulinodans* infection. Mol Plant Microbe Interact. 8(6):816-824.

Goto K and Meyerowitz EM. (1994). Function and regulation of the Arabidopsis floral homeotic gene PISTILLATA. Genes Dev. 8: 1548-1560.

Grace ML, Chandrasekharan MB, Hall TC, and Crowe AJ. (2004). Sequence and spacing of TATA box elements are critical for accurate initiation from the beta-phaseolin promoter. J Biol Chem. 279(9):8102-8110.

Graf T. (1992). Myb: a transcriptional activator linking proliferation and differentiation in hematopoietic cells. Curr Opin Genet Dev. 2: 249-255.

Graveley BR. (2001). Alternative splicing: increasing diversity in the proteomic world. Trends Genet. 17(2):100-107.

Grunstein M. (1997). Histone acetylation in chromatin structure and transcription. Nature 389: 349-352.

Guilfoyle TJ. (1997). The structure of plant gene promoters. Genet. Engin.19: 15-47.

Gualtier G, and Bisseling T. (2000). The evolution of nodulation. Plant Mol Biol. 42: 181-194.

Gubler F, Roger K, Roberts JK, and Jacobsen JV. (1995). Gibberellin-Regulated Expression of a myb gene in Barley Aleuron Cells: Evidence for myb Transactivation of a High-pl α-Amylase Gene Promoter. The Plant Cell. 7: 1879-1891.

Gutman A, and Kornblihtt AR. (1987). Identification of a third rigon of cell specific alternative splicing in human fibronectin mRNA. PNAS. 84: 7179-7182.

Gyorgyer J, Vaubert D, Jimenez-Zurdo JI, Charon C, Troussard L, Kondorosi A and Kondorosi E. (2000). Analysis of *Medicago truncatula* nodule expressed sequence tags. Mol Plant Microbe Interact. 13:62-71.

Hanahan D.(1985). Techniques for the transformation of E.coli. In: DNA Cloning. A Practical Approach. Vol. I. ed. D. Glover. IRL Press. Oxford. 109-135.

Hardie DG. (1999). Plant protein serine/threonine kinases: classification and functions. Annu Rev Plant Physiol Plant Mol Biol. 50: 97–131.

Hazen SP, Wu Y, Kreps JA. (2003) Gene expression profiling of plant responses to abiotic stress. Funct Integr Genomics. 3:105–111.

Hemm MR, Herrmann KM, and Chapple C. (2001). AtMYB4 a transcription factor general in the battle against UV. Trends Plant Sci. 6(4); 135-136.

Heim MA, Jakoby M, Weber M, Martin C, Weisshaar B, and Bailey PC. (2003). The basic Helix-Loop-Helix transcription factor family in plants: A genomic-wide study of protein structure and functional diversity. Mol Biol Evol. 20(5): 735-747.

Hirsch AM. (1992). Developmental biology of legume nodulation. New Phytol. 122:211-237.

Hoagland DR, and Arnon DI. (1950). The water culture for growing plants without soil. Calif. Agr.Exp. Stat.Cire.347, Univ of California Berkeley Press, CA.

Hong JC, Cheong YH, Nagao RT, Bahk JD, Key JL, and Cho MJ. (1995). Isolation of two soybean G-box binding factors which interact with a G-box sequence of an auxin-responsive gene. Plant J. 8(2):199-211.

Hunter T, and Karin M. (1992). The regulation of transcription by phosphorylation. Cell. 70: 375–387.

Hunter T. (1995). Protein kinases and phosphatases: the Yin and Yang of protein phosphorylation and signaling. Cell. 80: 225–236.

Jiang C, Gu X, and Peterson T. (2004). Identification of conserved gene structures and carboxy-terminal motifs in the Myb gene family of Arabidopsis and Oryza sativa L.ssp.indica. Genome Biol. 5(7): R46.

Jin H, and Martin C. (1999). Multifunctionality and diversity within the plant MYB-gene family. Plant Mol Biol. 41: 577-585.

Jimenez-Zurdo JI, Frugier F, Crespi M and Kondorosi A. (2000). Expression profiles of 22 novel molecular markers for organogenetic pathways acting in alfalfa nodule development. Mol Plant Microbe Interact. 13(1):96-106.

Johnson TK, Schweppe RE, Septer J, and Lewis RE. (1999). Phosphorylation of B-Myb regulates its transactivation potential and DNA binding. J Biol Chem.274(51):36741-9.

Kim JC, Lee SH, Cheong YH, Yoo CM, Lee SI, Chun HJ, Yun DJ, Hong JC, Lee SY, Lim CO, and Cho MJ.(2001). A novel cold-inducible zinc finger protein from soybean, SCOF-1, enhances cold tolerance in transgenic plants. Plant J. 25(3):247-259.

Kirik V, and Baumlein H. (1996). A novel leaf-specific myb-related protein with a single binding repeat. Gene. 183:109-113.

Klempnauer KH, Gonda TJ, and Bishop JM.(1982). Nucleotide sequence of the retroviral leukemia gene v-myb and its cellular progenitor c-myb: the architecture of a transduced oncogene. Cell. 31: 453 -463

Kornberg RD. (1999). Eukaryotic transcriptional control. Trends Cell Biol. 9(12): M46-49.

Kornblihtt AR, de la Mata M, Fededa JP, Munoz MJ, and Nogues G. (2004). Multiple links between transcription and splicing. RNA. 10(10):1489-1498.

Kosslak RM, Bookland R, Barkel J, Paaren H, and Applebaum ER. (1987). Induction of *Bradyrhizobium japonicum* common nod genes by isoflavones isolated from *Glycin max.* Proc Natl Acad Sci. 84: 7428-7432.

Kouchi H, Takane K, So RB, Ladha JK, and Reddy PM. (1999). Rice ENOD40: isolation and expression analysis in rice and transgenic soybean root nodules. Plant J. 18(2): 121-129.

Kranz H, Scholz K, and Weisshaar B. (2000). c-myb oncogene like genes encoding three MYB repeats occur in all major plant lineages. Plant J. 21(2): 231-235.

Kuykendall LD, and Weber DF. (1978). Genetically marked Rhizobium identifiable as inoculum strain in nodules of soybean plants grown in fields populated with Rhizobium japonicum. Appl. Environ. Micro. 36(6): 915-919.

Kuzma M, Winter H, Storer P, Oresnik I, Atkins CA, and Layzell DB. (1999). The site of oxygen limitation in soybean nodules. Plant Physiology. 119: 399-408.

LaCasse EC, and Lefebvre YA.(1995). Nuclear localization signals overlap DNA- or RNA-binding domains in nucleic acid-binding proteins. Nucleic Acids Res. 23(10):1647-1656.

Lemon B, and Tjian R.(2000). Orchestrated response: a symphony of transcription factors for gene control. Genes Dev. 14(20):2551-69.

Lipsick JS.(1996). One million year of MYB. Oncogene. 13: 223-235.

Liu L, White M J and MacRae TH. (1999). Transcription factors and their genes in higher plants, functional domains, evolution and regulation. Eur J Biochem. 262: 247-257.

Lohrmann J, Buchholz G, Keitel C, Sweere U, Kircher S, Baurle I, Kudla J, Schafer E, and Harter K. (1999). Differential expression and nuclear localization of response regulator-like proteins from Arabidopsis thaliana. Plant Biol. 1: 495-505.

Long S.R. (1989). Rhizobium-legume nodulation: life together in the underground. Cell 56: 203-214.

Lu CA, Ho TH, Ho SL, and Yu SM. (2002). Three novel MYB proteins with one DNA binding repeat mediate sugar hormone regulation of α-Amylase gene expression. Plant Cell. 14: 1963-1980.

Lusser A, Kolle D, and Loidi P (2001). Histone acetylation: lessons from the plant kingdom. Trends Plant Sci. 6(2): 59-65.

Macknight R, Duroux M, Laurie R, Dijkwel P, Simpson G, and Dean C. (2002). Functional significance of the alternative transcript processing of the arabidopsis floral promoter FCA. Plant Cell. 14: 877-888.

Makino S, Kiba T, Imamura A, Hanaki N, Nakamura A, Suzuki T, Taniguchi M, Ueguchi C, Sugiyama T, and Mizuno T. (2000). Genes encoding pseudo-response regulators: insight into His-to-Asp phosphorelay and circadian rhythm in Arabidopsis thaliana. Plant Cell Physiol. 41(6):791-803.

Mano S, Hayashi M, and Nishimura M. (1999). Light regulates alternative splicing of hydroxypyrovate reductase in pumpkin. Plant J. 17(3): 309-320.

Martin C, and Paz-Ares J.(1997). MYB transcription factors in plants. Trends Genet. 13(2): 67-73.

Maurer-Stroh S, Gouda M, Novatchkova M, Schleiffer A, Schneider G, Sirota FL, Wildpaner M, Hayashi N, and Eisenhaber F. (2004). MYRbase: analysis of genome wide Glycine myristoylation enlarges the functional spectrum of eukaryotic myristoylated proteins. Genome Biol. 5(3): R21.

Mercy IS, Meeley RB, Nichols SE, and Olsen OA. (2003). Zea mays ZmMybst1 cDNA, encodes a single Myb-repeat protein with the VASHAQKYF motif. J Exp Bot. 54(384): 1117-1119.

Meshi A, and Iwabuchi M. (1995). Plant transcription factors. Plant cell physiol. 36(8):1405-1420.

Meshi T, Moda I, Minami M, Okanami M, and Iwabuchi M. (1998). Conserved Ser residues in the basic region of the bZIP-type transcription factor HBP-1a(17): importance in DNA binding and possible targets for phosphorylation. Plant Mol Biol. 1998, 36(1):125-136.

Mitchell PJ and Tjian R. (1989). Transcriptional regulation in mammalian cells by sequence specific DNA binding proteins. Science, 245: 371-378.

Miyake K, Ito T, Senda M, Ishikawa R, Harada T, Niizeki M, and Akada S. (2003). Isolation of a subfamily of genes for R2R3-MYB transcription factors showing up regulated expression under nitrogen nutrient-limited conditions. Plant Mol Biol. 53: 237-245.

Modrek B, Lee C. (2002). A genomic view of alternative splicing. Nat genet.30: 13-19.

Moyano E, Martinez-Gracia JF, and Martin C. (1996). Apparent redundancy in myb gene function provides gearing for the control of flavonoid biosynthesis in antirrhinum flowers. Plant Cell. 8(9).1519-32.

Mylona P, Pawlowski K, and Bisseling T. (1995). Symbiotic Nitrogen Fixation. Plant Cell. 7:869-885.

Nam J, Kim J, Lee S, An G, Ma H, and Nei M.(2004). Type I MADS-box genes have experienced faster birth-and-death evolution than type II MADS-box genes in angiosperms. PNAS.101(7):1910-1915.

Nap JP, and Blsseling T.(1990).Developmental biology of a plant prokaryot symbiosis:The legume root nodule.Science.250:948-954.

Ner-Gaon H, Halachmi R, Savaldi-Goldstein S, Rubin E, Ophir R, and Fluhr R. (2004). Intron retention is a major phenomenon in alternative splicing in Arabidopsis. Plant J. 39(6):877-885.

Ogata K, Hojo H, Aimoto S, Nakai T, Nakamura H, Sarai A, Ishii S, and Nishimura Y. (1992). Solution structure of a DNA-binding unit of Myb: a helix-turn-helix-related motif with conserved tryptophans forming a hydrophobic core. PNAS. 89(14):6428-32.

Ogata K, Morikawa S, Nakamura H,Sekikawa SA, Imoue T, Kanai H, Sarai A, Ishii S, and Nishimura Y.(1994). Solution Structure of a Specific DNA Complex of the Myb DNA-Binding Domain with Cooperative Recognition Helices. Cell. 79: 639-648.

O'Grady K, Goekjian VH, Naim CJ, Nagao RT, and Key JL. (2001). The transcript abundance of GmGT-2, a new member of the GT-2 family of transcription factors from soybean, is down-regulated by light in a phytochrome-dependent manner. Plant Mol Biol.47(3):367-78.

Ohnishi T, Muroi A, and Tanamoto K. (2001). N-linked glycosylation at Asn(26) and Asn(114) of human MD-2 are required for toll like receptor 4-mediated activation of NF-Kappa B by lipopolysaccharide. J Immunol. 167(6): 3354-9.

Oppenheimer DG, Herman PL, Sivakumaran S, Esch J, and Marks MD. (1991). A myb gene required for leaf trichome differentiation in arabidopsis is expressed in stipules. Cell. 67: 483-493.

Panfield S, Meissner RC, Shoue DA, Carpita NC, and Bevan MW. (2001). MYB61 is required for mucilage deposition and extrusion in the Arabidopsis seed coat. Plant Cell. 13: 2777-2791.

Patharkar OR, and Cushman JC (2000). A stress-induced calcium-dependent protein kinase from Mesembryanthemum crystallinum phosphorylates a two-component pseudo-response regulator. The Plant J, 24(5): 679-691.

Paz-Ares J, Ghosal D, Wienand U, Peterson P, and Saedler H. (1987). The regulatory c1 locus of Zea mays encodes a protein with homology to myb oncogene products and with structural similarities to transcriptional activators. EMBO. 6:3553-3558.

Peng HM , Dreyer DA , VandenBosch KA and Cook D. (1996). Gene structure and differential regulation of the Rhizobium induced peroxidase gene rip1. Plant Physiol. 112: 1437-46.

Proudfoot NJ, Furger A, and Dye MJ. (2002). Integrating mRNA processing with transcription. Cell. 108: 501-512.

Ptashne M. (1988). How Eukaryotic transcription activators works. Nature. 335: 683-689.

Rabinowicz PD, Braun EL, wolfe AD, Bowen B, and Grotewold E. (1999). Maize R2R3 Myb genes: sequence analysis reveals amplification in the higher plants. Genetics. 153:427-444.

Rambaldi I, Kovacs EN, and Featherstone MS. (1994). A proline rich transcription activation domain in Murine HOXD-4. Nucleic Acids Res. 22(3): 376-382.
Reeder RH, and Lang WH. (1997). Terminating transcription in eukaryotes: lessons learned from RNA polymeraseI. TIBS. 22: 473-477.

Reeves R, and Beckerbauer L. (2001). HMGI/Y proteins: flexible regulators of transcription and chromatin structure. Biochim Biophys Acta. 1519(1-2):13-29.

Riechmann JL, and Meyerowitz EM (1998). The AP2/EREBP family of plant transcription factors. Biol Chem. 379(6):633-46.

Riechmann JL, and Ratcliffe O. (2000). A genomic perspective on plant transcription factors. Curr Opin Plant Biol.3: 423-434

Riechmann JL, Heard L, Martin C, Reuber L, Jiang CZ, Keddie J, Adam L, Pineda O, Ratcliffe OJ, Samaha RR, Creelman R, Pilgrim M, Broun P, Zhang JZ, Ghandehari D, Sherman BK, and Yu GL. (2000). Arabidopsis transcription factors: Genome wide comparative analysis among eukaryotes. Science. 290: 2105-2110.

Riedy MC, Timm EA, and Stewart CC. (1995). Quantitative RT-PCR for measuring gene expression. Biothecniques. 18(1): 70-4, 76.

Ro hrig H, Schmidt J, Miklashevichs E, Schell J, and John M (2002). Soybean ENOD40 encodes two peptides that bind to sucrose synthase. PNAS. 99: 915–1920.

Ro hrig H, John M, and Schmidt J. (2004). Modification of soybean sucrose synthase by S-thiolation with ENOD40 peptide A. Biochem Biophys Res Commun. 325(3):864-70.

Roeder RG. (1996). The role of general initiation factors in transcription by RNA polymerase II. TIBS. 21: 327-335.

Romero I, Fuetes A, Benito MJ, Malpica JM, Leyva A, and Paz-Ares J. (1998). More than 80 R2R3-MYB regulatory genes in the genome of Arabidopsis thaliana. Plant J. 14(3): 273-284.

Sambrook F, Fritsch EF, and Maniatis T. (1989). Molecular Cloning: A Laboratory Manual. 2nd ed. Cold spring Harbor Laboratory Press. Cold Spring Harbour, N.Y.

Sanger FS, Nicklen S and Coulson AR. (1977). DNA sequencing with chain-terminating inhibitors. PNAS. 74:5463-5467.

Schauser L, Christensen L, Borg S, and Poulsen C (1995). PZF, a cDNA isolated from Lotus japonicus and soybean root nodule libraries, encodes a new plant member of the RING-finger family of zinc-binding proteins. Plant Physiol. 107(4):1457-8.

Schauser L, Roussis A,Stiller J, and Stougaard J. (1999). A plant regulator controlling development of symbiotic root nodules. Nature. 402:191-195

Scheres B, van Engelen F, van der Knaap E, van de Wiel C, van Kammen A, and Bisseling T. (1990). Sequential induction of nodulin gene expression in the developing pea nodule. Plant Cell. 2(8): 687-700.

Schindler U, Terzaghi W, Beckmann H, Kadesch T, and Cashmore AR (1992). DNA binding site preferences and transcriptional activation properties of the Arabidopsis transcription factor GBF1. EMBO.11(4):1275-89.

Schmitz D, Lohmer S, Salamini F, and Thompson RD. (1997). The activation domain of the maize transcription factor Opaque-2 resides in a single acidic region. Nucleic Acids Res. 25(4): 756-763.

Schultze M and Kondorosi A.(1998). Regulation of symbiotic root nodule development. Annu Rev Genet. 32:33-57.

Schrumpfova P, Kuchar M, Mikova G, Skrisovska L, Kubicarova T, and Fajkus J.(2004). Characterization of two arabidopsis thaliana myb-like proteins showing affinity to telomeric DNA sequence. Genome. 47(2): 316-324.

Schwechheimer C and Bevan M. (1998). The regulation of transcription factor activity in plants. Trends Plant Sci. 3:378-383.

Serra EC, Carrillo N, Krapp AR, and Ceccarelli EA. (1993). One-step purification of plant ferredoxin-NADP+ oxidoreductase expressed in Escherichia coli as fusion with glutathione S-transferase. Protein Expr Purif. 4(6):539-546.

Singh KB. (1998). Transcriptional regulation in plants: The importance of combinatorial control. Plant Physiol. 118: 1111-1120.

Southern EM. (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis. J. Mol. Biol. 98: 503-517.

Stracke R, Weber M, and Weisshaar B. (2001). The *R2R3-MYB* gene family in *Arabidopsis thaliana.* Curr Opin Plant Biol. 4: 447-456.

Sugimoto K, Takeda S, and Hirochika H. (2000). MYB_related transcription factor NtMYB2 induced by wounding and elicitors is a regulator of the Tobacco Retrotransposon Ttl1 and defense-related genes. Plant Cell. 12: 2511-2527

Tang H, Sun X, Reinberg D, and Ebright RH. (1996). Protein-protein interaction in eukaryotic transcription initiation: structure of the preinitiation complex. PNAS. 93: 1119-1124.

Tian AG, Wang J, Cui P, Han YJ, Xu H, Cong LJ, Huang XG, Wang XL, Jiao YZ, Wang BJ, Wang YJ, Zhang JS, and Chen SY. (2004)._Characterization of soybean genomic features by analysis of its expressed sequence tags. Theor Appl Genet.108(5):903-13.

Turner BM. (2000). Histone acetylation and an epigenetic code. Bioassays. 22: 836-845.

Ulker B, and Somssich IE. (2004). WRKY transcription factors: from DNA binding towards biological function. Curr Opin Plant Biol. 7: 491-498.

Urao T, Yamaguchi-Shinozaki K, Urao S, and Shinozaki K. (1993). An Arabidopsis myb Homolog is Induced by Dehydration Stress and Its Gene Product Binds to the Conserved MYB Recognition Sewunce. the Plant Cell, 5:1529-1539.

Van de Wiel C, Narris JH , Bochenek B, Dickstein R, Bisseling T and Hirsch M(1990). Nodulin gene expression and enod2 localization in effective nitrogen fixing and ineffective bacteria free nodules of alfalfa. Plant cell. 2:1009-1017.

Van de Sande K, Pawlowski K, Czaja I, Wieneke U, Schell J, Schmidt J, Walden R, Matvienko M, Wellink J, Kammen A, Franssen H, and Bisseling T. (1996). Modification of phytohormone response by peptide encoded by enod40 of legumes and a nonlegume. Science. 273:370-373.

VanRhijn P and Vanderleyden J. (1995).The Rhizobium plant symbiosis. Microbiol Rev.59(1):124-142.

von Arnim AG, Deng XW, and Stacey MG. (1998). Cloning vectors for the expression of green fluorescent protein fusion proteins in transgenic plants. Gene.221(1):35-43.

Vorbrueggen G, Lovric J, and Moelling K.(1996). Functional analysis of phosphorylation at serine 532 of human c-Myb by MAP kinase. Biol Chem. 377(11):721-30.

Wada T, Tachibana T, Shimura Y, and Okada K.(1997). Epidermal cell differentiation in Arabidopsis determined by a Myb homolog, CPC. Science. 277: 113-116.

Weston K. (1998). MYB proteins in life, death and differentiation. Curr Opin Genet Dev. 8(1): 76-81.

Williams CE, and Grotewold E. (1997). Differences between plant and animal MYB domains are fundamental for DNA binding activity, and chimeric MYB domains have novel DNA binding specificities. J Biol Chem. 272: 563-571.

Winzer T, Bairl A, Linder M, Linder D, Werner D, and Muller P. (1999).Anovel 53-Kda nodulin of symbiosome membrane of soybean nodules, controlled by Bradyrhizobium japonicum. MPMI. 12(3):218-226.

Wykoff DD, Grossman AR., Weeks DP, Usuda H, and Shimogawara K. (1999). Psr1, a nuclear localized protein that regulates phosphorus metabolism in Chlamydomonas. PNAS. 96(26):15336-15341.

Wycoff KL, Hunt S, Gonzales MB, Van den Bosch KA, Layzell DB, and Hirsch AM. (1998). Effects of oxygen on nodule physiology and expression of nodulins in alfalfa. Plant Physiol. 117: 385-395.

Zhang H, Daoust F, Chares TC, Driscoll BT, Prithiviraj B, and Smith DL. (2002). Bradyrhizobium japonicum mutants allowing improved nodulation and nitrogen fixation of field-grown soybean in a short season area. J Agricultural Science.138: 293-300

Zhu Q, Dabi T, and Lamb C (1995). TATA BOX and Initiator Functions in the Accurate Transcription of a Plant Minimal Promoter in Vitro. The Plant Cell, 7:1681-1689.

APPENDIX

Figure 1 : The mRNA expression of 7/2 in various tissues (Set A2).

RT-PCR samples were separated on a 1.5 % agarose and process for Southern hybridization as described in the Materials and Methods. The probe was full length cDNA. Lane 1, seedlings; lane 2, 7 dpi cotyledons; lane 3, 24 dpi roots; lane 4, 15 dpi nodules; lane 5, 20 dpi nodules; lane 6, 24 dpi nodules; lane 7, 40 dpi nodules; lane 8, 24 dpi stems; lane 9, 24 dpi younger leaves, lane 10, 24 dpi older leaves; lane 11, 65 dpi senescent leaves; lane 12, 36 dpi flowers; lane 13, 65 dpi pods; lane 14, 1 pg cDNA clone 133c; lane 15, 10 pg cDNA clone 133c; lane 16, 100 pg gDNA clone 133g; lane 17, negative control.

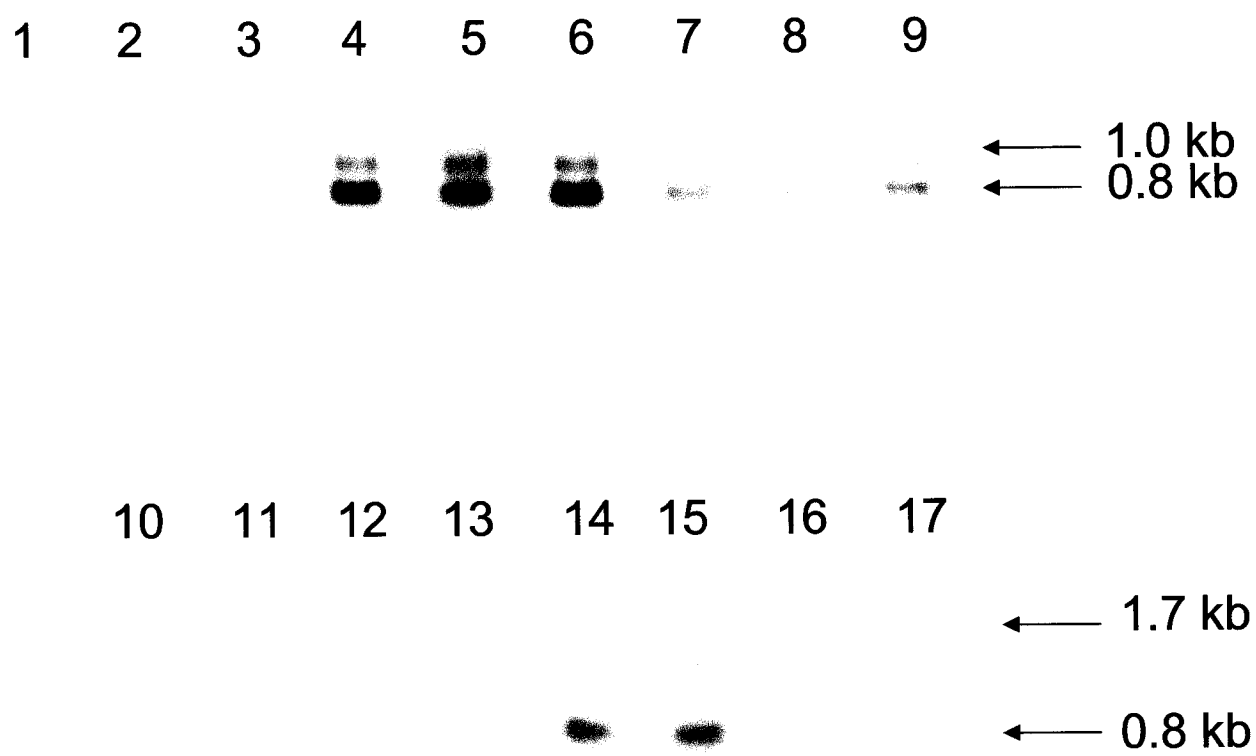

1 2 3 4 5 6 7 8 9
1.0 kb
0.8 kb
10 11 12 13 14 15 16 17
1.7 kb
0.8 kb

Figure 2: The mRNA expression of 7/2 in various tissues (Set B1).

RT-PCR samples were separated on a 1.5 % agarose and process for Southern hybridization as described in the Materials and Methods. The probe was full length cDNA. Lane 1, seedlings; lane 2, 7 dpi cotyledons; lane 3, 24 dpi roots; lane 4, 15 dpi nodules; lane 5, 20 dpi nodules; lane 6, 24 dpi nodules; lane 7, 40 dpi nodules; lane 8, 24 dpi stems; lane 9, 24 dpi younger leaves, lane 10, 24 dpi older leaves; lane 11, 65 dpi senescent leaves; lane 12, 36 dpi flowers; lane 13, 65 dpi pods; lane 14, 1 pg cDNA clone 133c; lane 15, 10 pg cDNA clone 133c; lane 16, 100 pg gDNA clone 133g; lane 17, negative control.

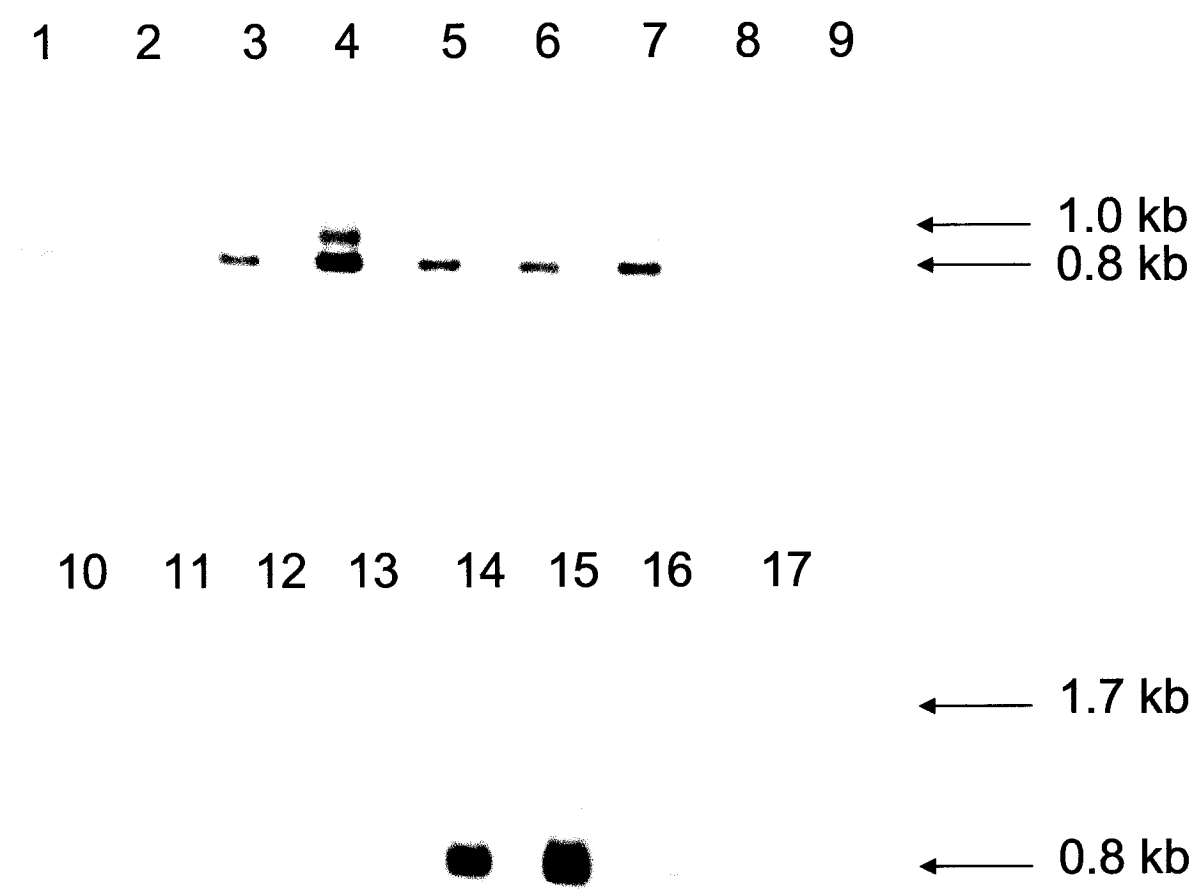

1 2 3 4 5 6 7 8 9
1.0 kb
0.8 kb
10 11 12 13 14 15 16 17
1.7 kb
0.8 kb

Figure 3: The mRNA expression of 7/2 in various tissues (Set B2).

RT-PCR samples were separated on a 1.5 % agarose and process for Southern hybridization as described in the Materials and Methods. The probe was full length cDNA. . Lane 1, seedlings; lane 2, 7 dpi cotyledons; lane 3, 24 dpi roots; lane 4, 15 dpi nodules; lane 5, 20 dpi nodules; lane 6, 24 dpi nodules; lane 7, 40 dpi nodules; lane 8, 24 dpi stems; lane 9, 24 dpi younger leaves, lane 10, 24 dpi older leaves; lane 11, 65 dpi senescent leaves; lane 12, 36 dpi flowers; lane 13, 65 dpi pods; lane 14, 1 pg cDNA clone 133c; lane 15, 10 pg cDNA clone 133c; lane 16, 100 pg gDNA clone 133g; lane 17, negative control.

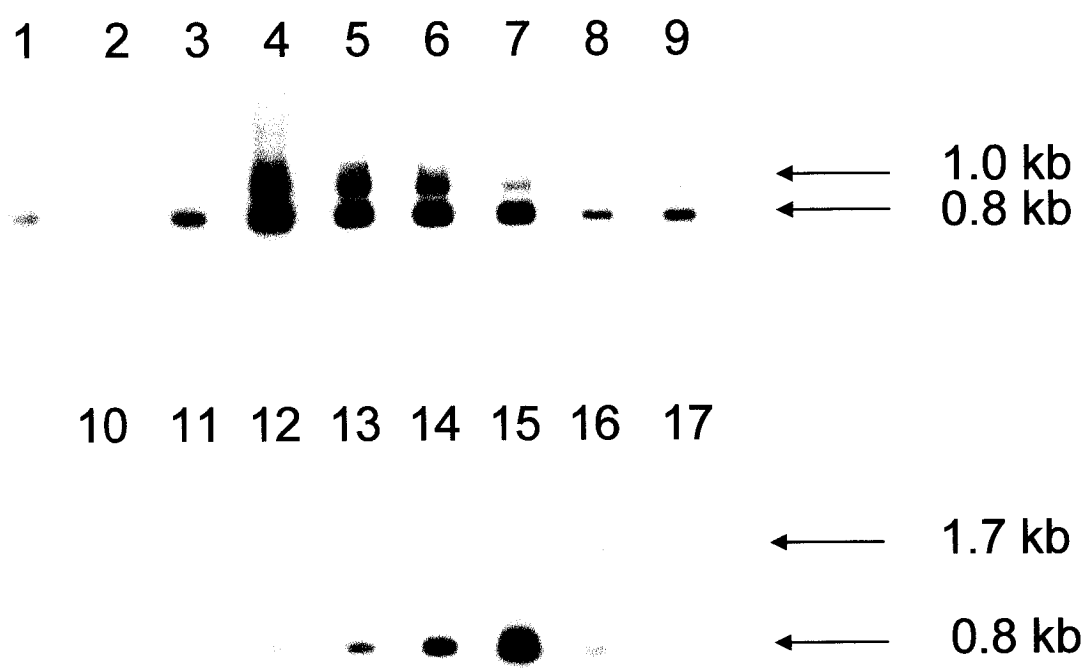
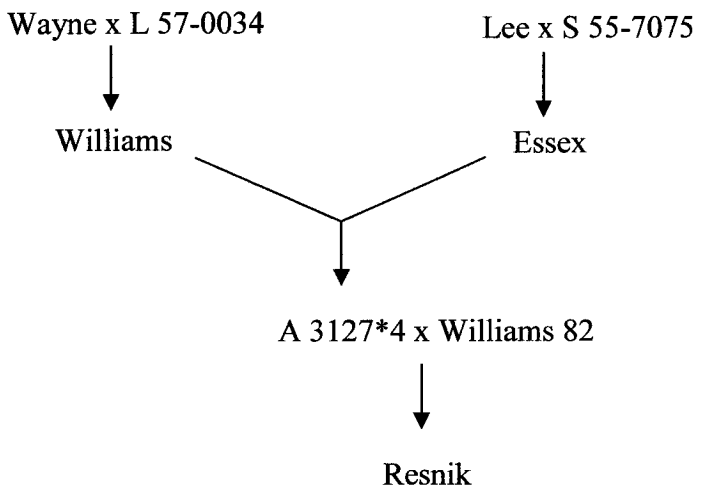

1 2 3 4 5 6 7 8 9
1.0 kb
0.8 kb
10 11 12 13 14 15 16 17
1.7 kb
0.8 kb

Figure 4: Pedigree relation ships giving ancestors of Maple Arrow and Resnik cultivars.


Wayne x L 57-0034
Lee x S 55-7075
Williams
Essex
A 3127*4 x Williams 82
Resnik

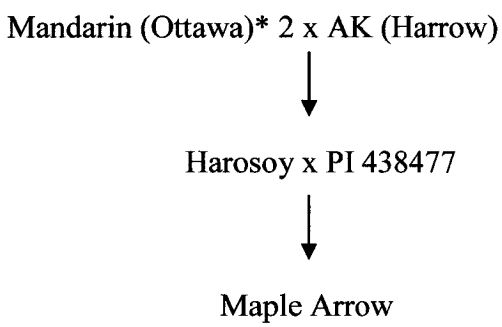

Mandarin (Ottawa)* 2 x AK (Harrow)
Harosoy x PI 438477
Maple Arrow

Table 1: Summary of mismatched positions on the gDNA and cDNA sequences.

F, forward sequence; R, reverse sequence; E, edited sequence. F and R sequences are single pass sequences while edited sequences represent the sequence obtained from the alignment of forward and inverse of reverse sequences. Most of the mismatches are in the introns and two are in coding region.

gDNA (Resnik) 4454bp			gDNA (Maple Arrow) 2726bp			gDNA (Maple Arrow) 1773bp			cDNA (Maple Arrow) 1383bp		
nt position	**nt ID**	**# of calls**	**nt position**	**nt ID**	**# of calls**	**nt position**	**Nt ID**	**# of calls**	**nt position**	**Nt ID**	**# of calls**
2602	A	2F/3R/ 2E	1125	-------	3F/2R	162	A	3E	----------	---------	---------
2608	T	2F/3R/2E	1131	C	3R	168	T	3E	----------	---------	---------
2826	T	3F/4R/2E	1347	-------	3F/3R	386	T	1E	----------	---------	---------
2827	T	3F/4R/2E	1348	-------	3F/3R	387	T	1E	----------	---------	---------
3059	A	2F/2R/3E	1580	G	3F/3R	619	A	3E	----------	---------	---------
3140	T	3F/3R/1E	1661	C	2F/3R	700	T	2E	----------	---------	---------
3638	A	3F/2R/1E	2158	G	1F/3R	1197	A	2E	----------	----------	---------
4041	A	4F/4R/1E	2561	G	4F/4R	1600	A	3E	1183	A	9F/9R
4051	A	4F/4R/1E	2571	G	4F/4R	1610	A	3E	1193	A	8F/8R