WHOLE-GENOME COMPARATIVE PROMOTER SEQUENCE ANALYSIS IN PLANTS

Nadia Indriana Chaidir

Department of Plant Science
McGill University
Montréal, Québec, Canada

March 2014

A thesis submitted to McGill University in partial fulfillment of the requirements for the degree of Master of Science.

© Nadia Chaidir (2014)

Table of Contents

Table of Contents ii

List of Figures v

List of Tables vii

List of Appendix viii

Abstract ix

Résumé x

Acknowledgements xii

List of Abbreviations xiii

Chapter 1: Introduction 1

Chapter 2: Hypotheses and Objectives 2

Hypothesis 1 2

Hypothesis 2 2

Objective 1 2

Objective 2 3

Objective 3 3

Chapter 3: Literature Review 4

3.1. Increasing availability of plant genome sequence 4

3.2. Essential regulatory elements of gene expression 5

3.2.1. Transcription factors and their roles 6

3.2.2. Types of promoters used to regulate gene expression 7

3.2.3. Research interests in promoter motif analysis 9

3.3. Approaches in plant promoter investigation 10

3.3.1. Molecular analysis of plant promoter 10

3.3.2. *In silico* prediction of *cis*-regulatory motifs 12

3.3.2.1. Bioinformatics tools for *cis*-regulatory motif prediction 13

3.3.2.2. Public databases for *cis*-regulatory elements 22

Chapter 4: Materials and methods 23

4.1. Motif prediction in promoters of orthologous genes 23

4.1.1. Obtain genome, peptide and coding DNA sequence of 18 plants 23

4.1.2. Generate Clusters of Orthologous Genes across 18 plant species 25

4.1.2.1. All-against-all Smith-Waterman peptide seq alignment 25

4.1.2.2. Generate cluster of orthologous genes between two species 25

4.1.2.3. Cluster orthologous genes across 18 plant species 26

4.1.2.4. Profiling *in silico* built clusters of orthologous genes with Gene Ontology 26

4.1.3. Retrieve promoter sequences corresponding to orthologous genes 26

4.1.4. *De novo* motif prediction in promoters of orthologous genes 27

4.1.5. Match *in silico* predicted motifs with known motifs 27

4.2. Motif prediction in promoters of co-regulated plant genes 28

4.2.1. Obtain list of co-regulated genes from soybean RNA-Seq data and retrieve their promoter sequences 28

4.2.2. *De novo* motif prediction in promoters of co-regulated genes using Seeder 29

4.2.3. *De novo* motif prediction in promoters of co-regulated genes using MEME 29

4.2.4. *De novo* motif prediction in promoters of co-regulated genes using BioProspector 30

4.2.5. Validate *in silico* predicted motifs with known motifs 30

4.3. Promologous: Database of *cis*-regulatory motifs across 18 plants 30

4.3.1. Create MySQL database containing genome-promoter information 30

Chapter 5: Results 32

5.1. Promoters of orthologous genes across 18 plants contain conserved *cis*-regulatory motifs 32

5.2. Promoters of co-regulated genes in shoot epidermis and shoot apical meristem tissues of soybean contain conserved *cis*-regulatory motifs 37

5.2.1. *De novo* motif discovery in promoters of co-regulated genes using Seeder 38

5.2.2. *De novo* motif discovery in promoters of co-regulated genes using MEME 41

5.2.3. *De novo* motif discovery in promoters of co-regulated genes using BioProspector 48

5.2.4. Similar *cis*- regulatory motifs detected in the same tissue-cultivar set by different motif prediction tools 55

5.3. Promologous: Database of *cis*-regulatory motifs across 18 plants 55

Chapter 6: Discussion 62

Chapter 7: Conclusion 68

Chapter 8: Future Research Directions 69

Chapter 9: Contribution to Knowledge 70

References 71

Appendix 92

Supplementary Figures 110

List of Figures

Figure 3.1 **14**

Illustration of all-against-all peptide sequence alignment between two genomes using Smith-Waterman algorithm

Figure 3.2 **14**

Illustration showing how InParanoid generate clusters of orthologous genes

Figure 3.3 **15**

Illustration showing how InParanoid assign confidence value (S) to each ortholog pair

Figure 3.4 **17**

Illustration showing how QuickParanoid merge pairwise COGs from InParanoid output

Figure 3.5 **21**

Comparison of Seeder v.0.01, Weeder v.1.3.1, BioProspector v.1, MEME v.3.5.4, Gibbs Motif Sampler v.3.03.003, and Motif Sampler v.3.2

Figure 5.3.1 **57**

MySQL Relational Database schema

Figure 5.3.2 **58**

Promologous web interface

Figure 5.3.3 **59**

Promologous query page

Figure 5.3.4 **60**

Overview of Promologous database structure

Figure 5.3.5 **61**

Output example in Promologous database

Supplementary Figure 6.2.2.1 **110**

Sequence logos for motifs found by MEME in promoters of co-regulated genes shoot epidermis tissue of cv. Clark

Supplementary Figure 6.2.2.2 **111**

Sequence logos for motifs found by MEME in promoters of co-regulated genes shoot epidermis tissue of 5-Leaflet mutant

Supplementary Figure 6.2.2.3 **112**

Sequence logos for motifs found by MEME in promoters of co-regulated genes shoot epidermis tissue of Glabrous mutant

Supplementary Figure 6.2.2.4 **113**

Sequence logos for motifs found by MEME in promoters of co-regulated genes shoot apical meristem tissue of cv. Clark

Supplementary Figure 6.2.2.5 **114**

Sequence logos for motifs found by MEME in promoters of co-regulated genes shoot apical meristem tissue of 5-Leaflet mutant

Supplementary Figure 6.2.2.6 **116**

Sequence logos for motifs found by MEME in promoters of co-regulated genes shoot apical meristem tissue of Glabrous mutant

Supplementary Figure 6.2.3.1 **119**

Sequence logos for motifs found by BioProspector in promoters of co-regulated genes shoot epidermis tissue of cv. Clark

Supplementary Figure 6.2.3.2 **120**

Sequence logos for motifs found by BioProspector in promoters of co-regulated genes shoot epidermis tissue of 5-Leaflet mutant

Supplementary Figure 6.2.3.3 **122**

Sequence logos for motifs found by BioProspector in promoters of co-regulated genes shoot epidermis tissue of Glabrous mutant

Supplementary Figure 6.2.3.4 **123**

Sequence logos for motifs found by BioProspector in promoters of co-regulated genes shoot apical meristem tissue of cv. Clark

Supplementary Figure 6.2.3.5 **125**

Sequence logos for motifs found by BioProspector in promoters of co-regulated genes shoot apical meristem tissue of 5-Leaflet mutant

Supplementary Figure 6.2.3.6 **127**

Sequence logos for motifs found by BioProspector in promoters of co-regulated genes shoot apical meristem tissue of Glabrous mutant

List of Tables

Table 4.1.1.1 **24**

List of 18 plant genomes used in this study along with their common names, version and genome paper references

Table 5.1.1 **33**

List of 18 plant species used in this study along with their genome paper reference, number of genes, peptide sequences and promoter sequences

Table 5.1.2 **36**

Motifs found in the promoters of orthologous genes

Table 5.2.1.1 **40**

DNA motifs discovered by Seeder in the promoters of co-regulated genes in different tissues of three soybean cultivars

Table 5.2.2.1 **45**

DNA motifs discovered by MEME in the promoters of co-regulated genes in different tissues of three soybean cultivars

Table 5.2.3.1 **52**

DNA motifs discovered by BioProspector in the promoters of co-regulated genes in different tissues of three soybean cultivars

List of Appendix

Appendix 4.1.2.1 **92**

All-against-all Smith-Waterman peptide sequence alignment

Appendix 4.1.2.2 **95**

Remove all orthologous gene members that have confidence value lower than 1.0 in InParanoid results

Appendix 4.1.2.3 **96**

Run QuickParanoid

Appendix 4.1.3 **97**

Retrieve promoter sequence

Appendix 4.1.4 **98**

De novo motif discovery using Seeder

Appendix 4.2.1.1 **102**

Analyze differential expression RNA-Seq data using DESeq

Appendix 4.2.1.2 **106**

Retrieve promoter sequence of each co-regulated genes

Appendix 4.2.2 **107**

De novo motif discovery in promoters of co-regulated genes using Seeder

Appendix 4.2.3 **108**

De novo motif discovery in promoters of co-regulated genes using MEME

Appendix 4.3.1 **109**

Build MySQL relational database for genome-promoter exploration in 18 plant species

Abstract

Large-scale genome-wide comparative analyses are now made possible by the increasing number of publicly available high-quality genome sequence data for numerous plant species. To understand the mechanisms of transcriptional regulation, computational analysis tools were used to find overrepresented and conserved DNA sequences, i.e. *cis*-regulatory elements. Datasets used as positive input for computational identification of regulatory regions commonly include promoters of co-regulated genes or promoters of orthologous genes (Wang and Stormo, 2003).

We discovered *de novo* motif using two approaches, seperately; 1) discovery based on orthology relationship of the genes in 18 plant species and 2) discovery based on co-regulated genes in specific tissues from soybean gene expression RNA-Seq data. In the first approach, a combination of several bioinformatics tools were used to predict motifs in promoter region based on clusters of orthologous genes in whole-genome datasets of *Arabidopsis lyrata, Arabidopsis thaliana, Brachypodium distachyon, Carica papaya, Chlamydomonas reinhardtii, Glycine max, Linus usitatissimum, Malus domestica, Manihot esculenta, Medicago truncatula, Oryza sativa, Physcomitrella patens, Populus trichocarpa, Selaginella moellendorfii, Sorghum bicolor, Vitis vinivera, Volvox carteri and Zea mays*. The results have shown that many promoters of orthologous plant genes contain similar *cis*-regulatory motifs. In addition, inclusion of more evolutionary distant organism led to detection of very conserved motifs, i.e. motifs that have similar function in wider variety of organisms. In the second approach, bioinformatics tools were used to find motifs in promoter region of co-regulated genes in shoot apical meristem and shoot epidermis of three soybean cultivars. The results have shows that promoters of co-regulated genes in specific tissues contain similar *cis*-regulatory motifs.

Since generating genome-scale datasets requires extensive computational resources that are not always readily available, we created a relational database that houses pre-computed and post-processed whole-genome comparative analysis of promoter regions. The database contains motif sequences, annotations, clusters of orthologous genes and other useful information associated with them, for 18 plant genomes.

Résumé

L’étude d’association pangénomique est maintenant rendue possible par le nombre de séquences génétiques de hautes qualités qui sont disponibles pour plusieurs espèces végétales. Pour comprendre les mécanismes de régulation de la transcription, un nombre d'outils d'analyses informatiques ont été développé pour identifier les éléments *cis*-régulatoires. Les bases de données utilisées comme saisie positive pour l’identification informatique des régions de régulation incluent communément les promoteurs des gènes co-régulés ainsi que des gènes orthologues (Wang et Stormo, 2003).

Pour découvrir les motifs *de novo,* nous avons utilisé deux techniques 1) une découverte basée sur la relation orthologue des gènes de 18 espèces végétales et 2) une découverte basée sur les gènes co-régulés dans certains tissus végétales spécifiques provenant de données de séquençage d’ARN de soja. Dans la première approche nous avons utilisé une combinaison de plusieurs outils bioinformatiques pour prédire les motifs des promoteurs basés sur des groupes de gènes orthologues trouvés dans les bases de données des génomes entiers d’*Arabidopsis lyrata, Arabidopsis thaliana, Brachypodium distachyon, Carica papaya, Chlamydomonas reinhardtii, Glycine max, Linus usitissimum, Malus domestica, Manihot esculenta, Medicago truncutula, Oryza sativa, Physcomitrella patens, Populus trichocarpa, Selaginella moellendorfii, Sorghum bicolor, Vitis vinivera, Volvox carteri* et *Zea mays*. Les résultats ont démontré que, dans les plantes, plusieurs promoteurs de gènes orthologues contiennent des motifs *cis*-régulatoires similaires. En plus, en incluant des espèces évolutivement éloignées dans les analyses, nous avons été capable de démontrer que ces motifs sont conservés. Dans la deuxième partie, nous avons fait une analyse comparant les séquences des promoteurs co-régulés dans les méristèmes apicaux ainsi que dans l’épiderme de trois cultivars de soja; Clark sauvage, mutant a 5-feuilles et mutant glabre. Les résultats ont démontré que les promoteurs des gènes co-régulés en différents tissus contiennent des motifs *cis*-régulatoires similaires.

Générer des données à l’échelle génomique demande une puissance informatique énorme qui n’est pas toujours disponible. En conséquence, nous avons créé une base de données pour 18 génomes de plantes composée de séquences de promoteurs, de motifs,

d’annotations et des groupes de gènes orthologues ainsi que d’autres informations associées avec ceux-ci.

Acknowledgements

This thesis would not have been possible without the help of many people in so many ways. There are a number of people without whom this thesis might not have been written, and to whom I am greatly indebted.

I would like to express the deepest appreciation of my supervisor, Professor Martina Strömvik who has been a very good advisor. Without her guidance and persistent help this thesis would not have been possible.

I would like to thank my committee chairs, Professor Jean-Benoit Charron and Professor Reza Salavati, whose advices have been very valuable and providing constructive criticism. I also would like to thank our lab alumni, François Fauteux for the useful advice, the administration staff of the Department of Plant Science, Carolyn Bowes, and other helpful faculty members at McGill University.

I would like to thank my lab mates who are now my friends; Yevgeniy Zolotarov who has been giving me so many useful suggestions and ideas regarding bioinformatics-related work, Haritika Majithia who has been so kind, supportive and generous to share her data with me, Muhammad Chragh, Andrew Blakney, and our lab neighbours, François Gagne-Bourque and Tanya Copley. Had I not made the acquaintance with all of them, this journey would not have been as enjoyable and fun.

I am very grateful to my loving parents, Alex J. Chaidir and Vera I. Judarta, my dear siblings, Ariane M. Lestari and Fahmi Indrawan, who have been an enormous source of love, encouragement and inspiration to me throughout my life and also for the myriad of ways in which have been actively supported me in my determination to find and release my true potential. To my dear best friend, Ryan Gibson whom I cannot thank enough for the encouragement, trust, practical and emotional support he has been giving me in the past few years.

I would like to acknowledge the McGill Recruitment Excellence Fellowship for providing funding through a graduate fellowship.

List of Abbreviations

BAC	Bacterial Artificial Chromosome
cDNA	Complementary DNA
CDS	Coding DNA Sequence
COG(s)	Cluster(s) of Orthologous Genes
CRE(s)	Cis-Regulatory Element(s)
cv	cultivar
DNA	Deoxyribonucleic Acid
DPE(s)	Downstream core Promoter Element(s)
EM	Expectation-Maximization
GFP	Green Fluorescent Protein
GO	Gene Ontology
HD	Hamming Distance
HMM	Hidden Markov Model
IDs	Identification(s)
INR	Initiator Element
IUPAC	International Union of Pure and Applied Chemistry
JGI	Joint Genome Institute
MEME	Multiple EM for Motif Elicitation
MLP	Major Latex Protein
mRNA	Messenger Ribonucleic Acid
NCBI	National Center for Biotechnology Information
NTP	Nucleoside Triphosphate
PLACE	Plant Cis-Acting Regulatory DNA Elements
PO	Plant Ontology
PWM	Position Weight Matrix
PSWM	Position Specific Weight Matrix
RFP	Red Fluorescent Protein
RNA	Ribonucleic Acid
SMD	Substring Minimal Distance

SQL	Structured Query Language
TAC	Transformation-competent Artificial Chromosome
TAIR	The Arabidopsis Information Resource
TBP(s)	TATA-box Binding Protein(s)
TF(s)	Transcription Factor(s)
TFBS(s)	Transcription Factor Binding Site(s)
TSS	Transcription Start Site
TSV	Tab-Separated Values
UAS	Upstream Activating Sequence
UTR	Untranslated Region
QTL	Quantitative Trait Loci

Chapter 1: Introduction

Whole-genome sequencing has contributed to the demise of a concept that a single gene or at most few of them encode each trait or characteristic of an organism. Gene regulation is one of the important factors of diversity in the phenotypes seen in nature, and one of the goals of studying gene regulation is to understand how a plant regulates transcription of 14,000-66,000 genes in the proper patterns. To decipher the mechanisms, numerous processes influencing transcription need to be well understood, and one of them is how genes are turned on and off at different locations and times. With the increasing number of high quality plant genomes sequenced, a comparative DNA sequence analysis in multiple species is becoming possible and there is a tremendous amount of information that we can learn from this. Comparison of sequences between multiple species has been a very useful method to identify functional regions in the genome (Loots *et al*., 2000), more powerful than pairwise DNA sequence comparisons (Dubchak and Frazer, 2003).

In this study, we were interested in investigating if promoters of orthologous plant genes have similar *cis*-regulatory motifs, which are important factors to the regulation of those genes. Bioinformatics tools were used to predict promoter motifs in clusters of orthologous genes (COGs) from 18 plant species; *Arabidopsis lyrata, Arabidopsis thaliana, Brachypodium distachyon, Carica papaya, Chlamydomonas reinhardtii, Glycine max, Linus usitatissimum, Malus domestica, Manihot esculenta, Medicago truncatula, Oryza sativa, Physcomitrella patens, Populus trichocarpa, Selaginella moellendorfii, Sorghum bicolor, Vitis vinivera, Volvox carteri,* and *Zea mays*. In addition, by utilizing RNA sequencing data from soybean, we were able to see if promoters of genes co-regulated in specific tissues have similar *cis*-regulatory motifs, which could give us more insight on the gene expression pattern.

In silico prediction results from this study can serve as a guide to establish the function expression and provide a good starting point for further experimental analysis (Fauteux and Strömvik, 2009).

Chapter 2: Hypotheses and Objectives

Hypothesis 1:

Promoters of orthologous plant genes have similar *cis*-regulatory motifs.

Hypothesis 2:

Promoters of genes co-regulated in specific tissues have similar *cis*-regulatory motifs.

Objective 1: To predict motifs in orthologous promoters in 18 plant species

Objective 1.1.

Build MySQL in-house database as bioinformatics environment to explore information related to promoters in 18 plant genomes and to predict *cis*-regulatory motif by downloading genome, peptide and coding DNA sequences of 18 plant species from Phytozome v8.0 (Goodstein *et al.,* 2012)

Objective 1.2.

Build Clusters of Orthologous Genes (COGs) of 18 plant species using InParanoid (Remm *et al*., 2001) and QuickParanoid (http://pl.postech.ac.kr/QuickParanoid/)

Objective 1.3.

Profile bioinformatically built COGs using Gene Ontology (GO) annotations (The Gene Ontology Consortium, 2000) to investigate the functional relationship between the gene members in each cluster

1.3.1. Create GO table in relational database table consisting of gene name, GO ID, GO terms and GO description

1.3.2. Using gene ID as the correlating key, query the database to see what similarities do each COG share in terms of the function

Objective 1.4.

Run *de novo* motif discovery using Seeder (Fauteux et al., 2008)

1.4.1. Generate index and background sets of all promoters in 18 species.

1.4.2. Set up script to find motif in each COG then filter significant motifs

Objective 1.5.

Match found significant motifs with experimentally characterized motifs in the

PLACE database (Higo *et al.,* 1999)

Objective 2: To predict motifs in promoters of co-regulated genes in specific tissues (epidermis *vs.* shoot apical meristem) in different soybean cultivars

Objective 2.1.

Obtain a list of upregulated genes from soybean RNA-Seq gene expression data from each tissue in each cultivar

Objective 2.2.

Retrieve promoter sequences of each co-regulated genes using BioPython (Cock *et al*., 2009) from in-house MySQL database

Objective 2.3.

Run *de novo* motif discovery using BioProspector (Liu et al., 2001), MEME (Bailey and Elkan, 1994) and Seeder (Fauteux *et al*. 2007)

2.3.1. Obtain soybean promoter sequence from in-house database for background sequence file

2.3.2. Find *cis*-regulatory motifs in promoters of upregulated genes and filter significant motifs

Objective 2.4.

Match found significant motifs with experimentally characterized motifs in PLACE database (Higo *et al.,* 1999) using STAMP (Mahony and Benos, 2007)

Objective 3: To build a public relational database for exploration of the plant genome-promoter data

Objective 3.1.

Build MySQL relational database that contains pre-computed whole-genome comparative data of promoter sequences across 18 plant species, as well as other associated genome information

Chapter 3: Literature Review

3.1. Increasing availability of plant genome sequence

A high quality reference genome has become an important resource to illuminate the function of genes in development, drive genomics-based approaches to systems biology and identify genomic variations of an organism. Since the first higher eukaryotic genome sequenced was from a plant, *Arabidopsis thaliana* (The Arabidopsis Genome Initiative, 2000), the cost of sequencing has been decreasing by 10,000 times during the past ten years (Collins, 2010). Genome sequencing continues to become more affordable and as a result, more genome sequences are being published.

Discoveries in plant genome sequences have been enabling researchers to conduct studies in functional and comparative genomics. The comparison of whole genome sequences provides a highly detailed view of how organisms are related to each other at the genetic level, distinguishes different life forms from each other. Comparative genomes also offers a powerful tool for studying evolutionary changes among organisms, helping to identify genes that are conserved among species, as well as genes that give each organism its unique characteristics. Information obtained from these studies can lead to practical applications in crop improvement such as increasing quality and crop yield, increasing tolerance of environmental pressures like salinity, drought, extreme temperature, resistance to viruses, fungi and bacteria, increasing tolerance to insect pests and herbicides.

The genome of *Arabidopsis thaliana* was successfully sequenced in the late 2000. The genome sequencing covered 115.4 megabases of the 125-megabase of the first complete plant genome that has many advantages for genome analysis including short generation of time, small size, and large number of offspring (The Arabidopsis Genome Initiative, 2000). During the sequencing of this first plant genome, researchers utilized the large-insert bacterial artificial chromosome (BAC), phage (P1), and transformation-competent artificial chromosome (TAC). The genome sequence of *Arabidopsis thaliana* has greatly contributed to the progress in genomics-based Arabidopsis research as well as the exploitation of annotated genes to explore orthologous genes in other plants (Feuillet *et al*., 2010). In January 2010, the genome sequence of the palaeopolyploid soybean, *Glycine max* L. (Merr.), was publicly available and this was the largest whole-genome shotgun sequenced plant genome so far (Schmutz *et al.*, 2010). With the increasing availability of sequenced plant genomes and progress of genetic association studies, the identification and characterization of quantitative trait loci (QTL) becomes easier and this has become a crucial factor to improve crop production to fulfill the food and energy security demands of the world in the future (Rounsley *et al.*, 2009).

3.2. Essential regulatory elements of gene expression

The theory of a gene is an essential part of all the fields in biology that includes genetics, molecular biology, and evolutionary biology. The way classical genetics defined a gene was fairly abstract: a unit of inheritance that passes along the trait from parents to the progeny. By looking from the biochemistry perspective, those trait and characteristics were linked with enzymes or proteins, and with the rise of molecular biology, those genes became real, physical things — sequences of DNA with information that is transcribed into strands of messenger RNA, which are used as the recipe for which building blocks are needed for protein assembly, piece by piece (Pearson, 2006). Although biologists have put tremendous efforts into unraveling the gene concept over the years, still a great deal remains to be discovered.

Gene expression coordination in eukaryotes is similar to prokaryotes but in a much more convoluted way (Griffiths *et al.*, 2000). It involves every path from initiation of mRNA synthesis to the last part of protein products. The regulation of transcription is important because it controls when and where a gene should be expressed or not (Wray *et al.*, 2003). In a typical gene, a proximal non-coding region in genomic DNA that facilitates RNA polymerase to initiate the transcription and regulation of a particular gene is called promoter. It is composed of specific short conserved DNA sequences called *cis*-regulatory elements or motifs, which are recognized and bound by specific transcription factors that can regulate the gene transcription level and pattern (De Boer *et al.*, 1999).

Promoters can have varying sizes, numbers of motifs, and locations (Potenza *et al.*, 2004). The minimal region of DNA in eukaryotes that engages the basal transcription machinery to an accurate and efficient transcription initiation is called core promoter (Yang *et al.*, 2007). It is compact, composed of 60 bp straddling the transcription start site and located immediately adjacent to and upstream of the gene. A typical core promoter encompasses DNA sequences between approximately -40 and +50 relative to Transcription Start Site (TSS) (Smale, 1994). The core promoter serves as the binding region for RNA Polymerase II (Pol II) and its accessory factors, and directs them to begin transcribing at the proper start site. It contains TATA box, initiator element (INR), downstream core promoter element (DPE), and TFIIB recognition element (Smale and Kadonaga, 2003). Roughly half of core promoters contain a TATA-box, which serves as a binding site for TATA-box binding protein (TBP) (Struhl *et al.*, 1998). The upstream activating sequences (UAS) that consist of two or three closely connected binding sites for one or two distinct sequence-specific transcription factors regulate the TATA-box Binding Protein (De Bruin *et al.*, 2001). Although the core promoter plays the important role in this case, gene expression would not be significant without any additional control motifs or transcription factor binding sites (Potenza *et al.*, 2004; Wray *et al.*, 2003).

Another essential regulatory element called enhancer is a short region of DNA where proteins bind to elevate the expression of an adjacent coding region. Though enhancers are not well defined, an idea of what an enhancer could entail is as follows; 500 bp in length, and has ten binding sites for at least three different sequence-specific transcription factors, typically one repressor and two different activators (Davidson, 2001).

Other events such as DNA methylation and histone acetylation can also alter the gene expression pattern in cells by suppressing or decreasing the level of expression. In addition, chromatin structure can also perturb in the surrounding of expressed genes – most obvious in promoter and enhancer regions (Felsenfeld *et al.*, 1996). Chromatin is the association of DNA and histone proteins, which are tightly bonded by the attraction between negatively charged DNA and the positively charged histones (Phillips, 2008). This assembly contributes to the varying levels of complexity in gene regulation and allows simultaneous regulation of functionally or structurally related genes that tend to be present in widely spaced clusters or domains on eukaryotic DNA (Sproul *et al.*, 2005).

3.2.1. Transcription factors (TFs) and their roles

Transcription factors (TFs) are proteins, typically 5-20 aa long, which bind to short DNA sequences. These factors can act alone or as a part of protein complex to repress or activate the recruitment of RNA Polymerase II to promoter regions. Promoter regions commonly contain 10-50 Transcription Factor Binding Sites (TFBS) for 5-15 different TFs (Arnone and Davidson, 1997).

3.2.2 Types of promoters used to regulate gene expression

A highly structured organization of gene regulation allows comprehensive control of gene expression. The interest of promoter investigation derives from the infinite opportunities for controlling gene expression since it has opened up the chances of modulating gene expression in homologous and heterologous organisms where foreign promoters are inserted together with the genes of interest. The promoter has been one of the key determinants used in plant genetic engineering applications to design a transformation-cassette that would enable an accurate control of transgene activity, whether it is spatial and/or temporal expression (Venter, 2007).

Depending on the type of control desired to drive certain gene expression, promoters can be categorized into four; constitutive promoter, inducible promoter, tissue-specific or development-stage-specific promoter, and synthetic promoter.

A constitutive promoter drives the expression in the entire tissues and irrespective of developmental, biotic and abiotic factors. This type of promoter is normally active across species and even kingdoms. Some examples of this type of promoter are *CaMV 35S* promoter and Opine promoter.

An inducible promoter allows a gene to be induced in the presence of biotic or abiotic factors. This promoter may be physically or chemically induced. Chemically-induced promoters typically modulated by chemical compounds that either turn on or off gene expression, such as alcohol-regulated promoters, e.g. alcohol dehydrogenase I *alcA* gene promoter (Roslan *et al.,* 2001, Roberts *et al.,* 2005); tetracycline-regulated promoters, e.g. *Tn10*-encoded *tet* repressor (Gatz and Quail, 1988); steroid-regulated promoters (Schena *et al.,* 1991) and metal-regulated promoters, e.g. *PvSR2* promoters (Qi *et al.,* 2007). In 2007, Qi *et al* reported on the characterization of a novel plant promoter *PvSR2* gene specifically stimulated by heavy metal and identification of promoter regions conferring heavy metal responsiveness in bean *(Phaseolus vulgaris)*. Results of serial promoter deletions and GUS assay showed that sequence 74-bp fragment sequence between -222 and -147 relative to TSS was needed for heavy metal-specific induction of *PvSR2* promoter. In addition, *PvSR2* promoter region of transformed tobacco *(Nicotina tabacum)* was also analyzed using GUS assay and showed that heavy metal-specific responsive activity relies on the type and concentration of the heavy metal and the type of organ. This study sheds some light in the *PvSR2* gene regulation and provides a new heavy-metal-inducible promoter system in transgenic plants (Qi *et al*., 2007).

On the other hand, physically induced promoters are influenced by environmental factors such as water or salt stress, temperature, oxygen level and light. In 1989, Czarnecka *et al* identified the regulatory region of Gmhsp17.5-E heat shock promoter in soybean by insertion-deletion mutagenesis with transgenic expression monitored in *Agrobacterium tumefaciens*-incited tumors of sunflower. There are four regions contributing to promoter activity and they all map within 244 bp from the transcription start site (Czarnecka *et al.,* 1989).

Tissue-specific or development-stage-specific promoters are explained by the name itself. These types of promoter operate in certain tissues, such as fruits (Guillet *et al.,* 2012; Hiwasa-Tanase *et al.,* 2012), roots (Jeong *et al.,* 2010; Li *et al.,* 2013), seeds (Zavallo *et al.,* 2010), flowers (Macknight *et al.,* 2002; Yoo *et al.,* 2011); or only at particular developmental stages, such as vegetative stage (Weinhold *et al.,* 2013; Dutt *et al.,* 2012; Zhang *et al*. 2013) and reproductive stage (Ren *et al.,* 2005; Huda *et al.,* 2013). An example of a study on root-specific promoter by Souza *et al*., 2009, reported on the isolation of promoter sequence of *Mec1* gene in cassava that codes for a glutamic-acid rich protein that is expressed differently in storage root, *Pt2L4* by using Inverse PCR. *In silico* analysis showed that putative *cis*-regulatory elements in *Mec1* gene modulates its gene expression in vascular cambium and starch-rich parenchyma cells that participates in some mechanisms related to second growth. A transient expression experiment was performed in cassava and results showed that ATATT-motif identified by *in silico* analysis are necessary for vascular expression in roots. In addition, DNA sequences identified in this study is a new promoter that can be a potential candidate for genetic engineering of cassava roots. Computational analysis results can be used as a guide to establish the function expression in order to identify regions that exhibits tissue-specific

Mec1 promoter activity (Souza *et al*., 2009).

An example of a study on promoter that regulates expression in vegetative tissue by Saeed *et al*., 2008, reported on the promoters of soybean seed lectin homologues *Le2* and *Le3* in *Arabidopsis*. Lectin promoters were isolated, cloned, sequenced, fused with *gusA* reporter gene, and floral dip-transformed using *Agrobacterium*. GUS expression analysis indicates that *Le3* promoter is found predominantly in vegetative tissues including root, where as *Le2* promoter is very low expressed in all tissues except roots. The expression profile of *Le1* and *Le3* shows some correlation that when *Le1* promoter effect decreases, the effect of *Le3* promoter increases. The same method was being used to *Le1, Le2,* and *Le3* promoters that includes the signal peptides from three respective genes. Results showed the same patterns as the ones from constructs without signal peptides and they are consistent with computational predictions (Saeed *et al*., 2008).

Lastly, a synthetic promoter contains the consensus of DNA sequences of common natural promoter elements from diverse origins (Venter, 2007). In 2011, Liu *et al*, engineered tobacco transgenic plants for the purpose of plant pathogen infection early detection; this was accomplished by using a synthetic pathogen inducible promoters fused to reporter genes in tobacco leaves. The synthetic promoter constructs were capable to confer the inducibility of the RFP reporter in response to infection by pathogens (Liu *et al.,* 2011).

3.2.3. Research interests in promoter motif analysis

Due to the high degree of variation observed in the gene- and species-specific architectures of the regulatory sequences, the precise identification and characterization or promoters and transcription start site localization remains a major challenge in bioinformatics (Azad *et al*., 2011).

Pavesi *et al*., 2004, defined motif as “a group of the same size or with the same degree of similarity among the oligonucleotides forming it”. Short promoter sequences or motifs that serve as transcription factor binding sites largely determine the gene expression pattern. Therefore in order to elucidate the gene expression regulatory networks, it is very important to identify and characterize the regulatory motifs (Wyrick, 2002).

For *de novo* motif discovery, many approaches have been used. The first one is focused approach: assemble a small set of sequences and search for overrepresented patterns in the sequences relative to background model. There are many examples of algorithm that use this approach, for instance MEME (Bailey and Elkan, 1994) and Gibbs Sampler (Thompson *et al.,* 2003). The second approach is focused discriminative approach: assemble two sets of sequences and look for patterns relatively over-represented in one of the input sets (Sinha, 2003; Workman and Stormo, 2000). The third approach uses phylogenetic information: use sequence conservation information about the sequences in a single input set. Some examples are PhyME (Sinha *et al.,* 2004) and PhyloGibbs (Siddharthan *et al.,* 2001). The fourth approach is whole-genome method: look for over-represented, conserved patterns in multiple alignments of the genomes of one or more species (Xie *et al.,* 2005; Kellis *et al.*, *2004*).

3.3. Approaches in plant promoter investigation

Elucidating transcriptional regulation of plant genes is essential in gaining knowledge on how to orchestrate and engineer the gene expression in order to improve qualities of a crop such as yield and resistance to external factors (e.g. diseases and extreme temperatures). There are two different ways on investigating promoters. Due to an increasing number of publicly available whole genome sequences, computational analyses on promoters have been made possible. Bioinformaticians have been using different algorithm to create computer programs that could find specific motifs or patterns that lies within promoter sequences. This information could tell us the location of TFBS and what type of TF binds to it. These computational results can serve as a guideline as a starting point for molecular biologists to do molecular analysis on promoters.

3.3.1. Molecular analysis of plant promoter

In order to understand the mechanisms that control gene expression, many experimental studies that identify and characterize *cis*-acting regulatory elements (CRE) had been done. Promoter deletion analysis, GUS reporter assay, GFP reporter assay, nuclear run-off assay are commonly used to investigate promoter effect on gene expression.

Promoter deletion analysis is used to experimentally identify which region of the promoter plays role in any specific expression in the plants. Promoter 5' and 3' deletion constructs are being inserted into a specific vector for further amplification, cloning and digestion series. As an example, promoter deletion analysis elucidated the role of *cis* elements and 5' UTR intron in spatiotemporal regulation of high-affinity phosphate transporter *AtPht1;4* expression in Arabidopsis (Karthikeyan *et al.,* 2009).

In GUS reporter assay, the bacterial *uidA* gene from *Escherichia coli* that codes for ß-glucuronidase (GUS) is used to investigate promoter activity in terms of gene expression visually and quantitatively. During GUS histochemical staining, the ß-glucuronidase enzyme has the ability to cleave the colorless 5-bromo-4-chloro-3-indolyl glucuronide (X-Gluc) substrate into a blue-colored product and other colorless product, which is detected at the site of enzyme activity. The *uidA* gene can also be used in quantitative assays of a promoter, when used with the fluorescent 4-methylumbelliferyl-glucuronide (MUG) substrate. Stromvik *et al* (1999), reported on the isolation of *Msg* gene in soybean that possesses a high homology to Major Latex Protein (MLP) and very well expressed in developing soybean pods. This promoter was analyzed with 14 distinct promoter fragments that range from 0.65 kb to 2.26 kb, fused with *uidA* (GUS) gene. High transient expression experiment and transformation using *Agrobacterium* shows that *Msg* promoter is completely active only on the 2.26 kb-fragment promoter transformants and GUS was expressed in most part of the plants, but not in mature leaves. The proximal 650bp TATA-containing region in Arabidopsis and soybean is dependent and can be deleted without any changes in the expression pattern. Unlike most of the tissue-specific elements that was located closer to the TATA box, the *Msg* promoter tissue-specific region was located in the distal 5' region upstream of the dispensable TATA box (Stromvik *et al*., 1999).

Another reporter gene, the green fluorescent protein (GFP), was used as a reporter to investigate an *in vivo* expression level in transgenic plants. The fluorescence level of GFP can be quantified to study underlying differences in gene expression. It was shown by the increment of both GFP fluorescent intensity and GFP mRNA level in proportion to the GFP gene copy number (Soboleski *et al*., 2005). It is also reliable in measuring the activity of weak promoters (Ducrest *et al*., 2002). Yingzhen *et al*., 2008, reported on the identification of *pGCI* promoter in Arabidopsis and tobacco plants. Results of YC-3.60, which is a GFP-based calcium FRET reporter, showed that the expression was adequately strong to illustrate intracellular Ca^{2+} dynamics in guard cells of intact plants and resolved spontaneous calcium transients in guard cells. The region of promoter that drives a strong reporter expression in guard cells of wildtype Arabidopsis, the *Too Many Mouths* (TMM) mutant and tobacco, was narrowed down by serial promoter deletions. In addition, the *GCI* promoter was proved to be a effective tool to alter plant performance under stress environment and manipulating specific gene expression in guard cells by using anti-sense approach (Yingzhen *et al*., 2008).

Another method called the nuclear run-off assay allows scientist to see which genes in a population of cells are expressed at a certain time. The RNA transcripts of interest are radiolabeled with Phosporous-32 NTPs as they are elongated in isolated nuclei where new transcripts are not initiated. The radiolabeled run-off transcripts are used as hybridization probes with DNA from recombinant plasmids that contain the gene of interest to detect if a specific gene is expressed.

3.3.2. *In silico* prediction of *cis*-regulatory motifs

Computational analysis is necessary for finding genes and partitioning exons in genes; however, genome annotations has been more focused on identifying protein coding regions and function of the genes rather than predicting *cis*-regulatory elements in non-coding regions. In contrast to coding sequences that have a direct association with their immediate phenotype, *cis*-regulatory sequences have an indirect, non-linear relationship with their phenotype, which is a particular profile of transcription (Wray *et al*., 2003).

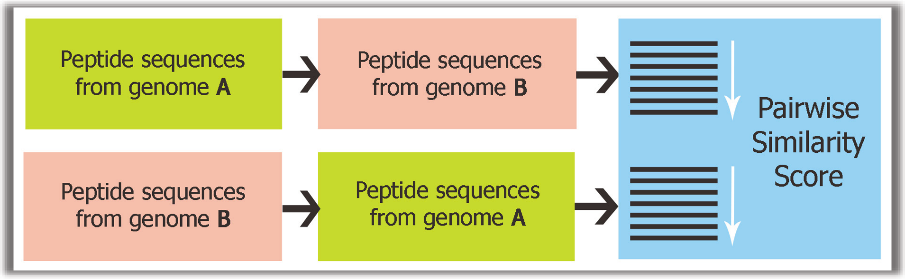
In silico studies can help scientists to elucidate intrinsic sequence properties of plant promoters. Many methods and algorithms have been developed to predict common patterns or motifs, which is the general way to predict *cis*-regulatory elements.

Regulatory elements in DNA are among the most important biological features that are represented by sequence motifs. Biological sequence motifs are nucleotide or amino acid sequence patterns that are short and usually fixed-length. Motifs can represent transcription factor binding sites (TFBSs), splice junctions, and binding domains in DNA, RNA, and protein molecule, respectively. The discovery of sequence motifs can shed a light in understanding transcriptional regulation, mRNA splicing, and formation of protein complexes (Keith, 2008).

3.3.2.1. Bioinformatics tools for *cis*-regulatory motif prediction

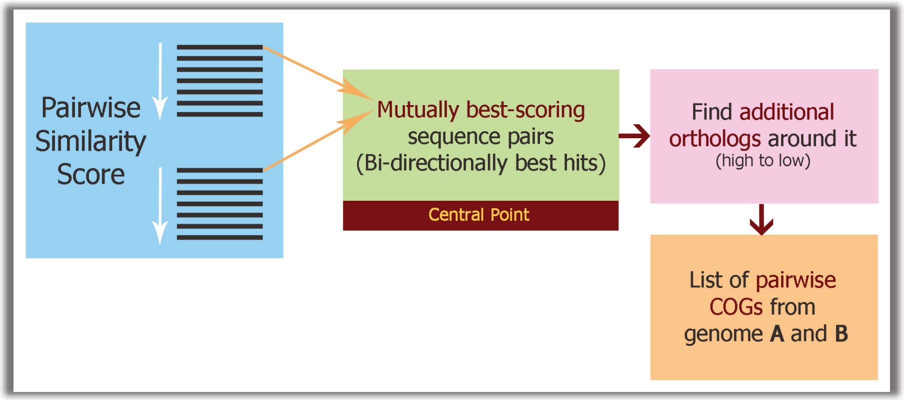
A common approach to find *cis*-regulatory motif is by using promoters of orthologous gene groups as the input sets. Prior to finding motifs, peptide sequence of two or more genomes can be aligned using the all-against-all Smith-Waterman algorithm. The algorithm aligns query locus from one genome (e.g. genome A) against target locus in the other genome (e.g. genome B), then assign Pairwise Similarity Score for each locus pair alignment (Figure 3.1). This was done repetitively until all loci in all genomes were processed. Then, tools that cluster orthologous genes can be used, for instance InParanoid (Remm *et al.,* 2001), QuickParanoid (http://pl.postech.ac.kr/QuickParanoid/), MultiParanoid (Alexeyenko *et al.,* 2006) and OrthoMCL (Li *et al.,* 2003). From the pairwise similarity score in each locus pair alignment assigned by the Smith-Waterman algorithm, program such as InParanoid takes the mutually best pairwise similarity score from the species pair, then finds additional orthologs around that score from high to low and lists pairwise cluster of orthologous genes from genome species A and B (Figure 3.2 and Figure 3.3).

Figure 3.1. Illustration of all-against-all peptide sequence alignments between two genomes using Smith-Waterman algorithm (adapted from Remm *et al*., 2001)


Peptide sequences from genome **A**
Peptide sequences from genome **B**
Peptide sequences from genome **B**
Peptide sequences from genome **A**
Pairwise Similarity Score

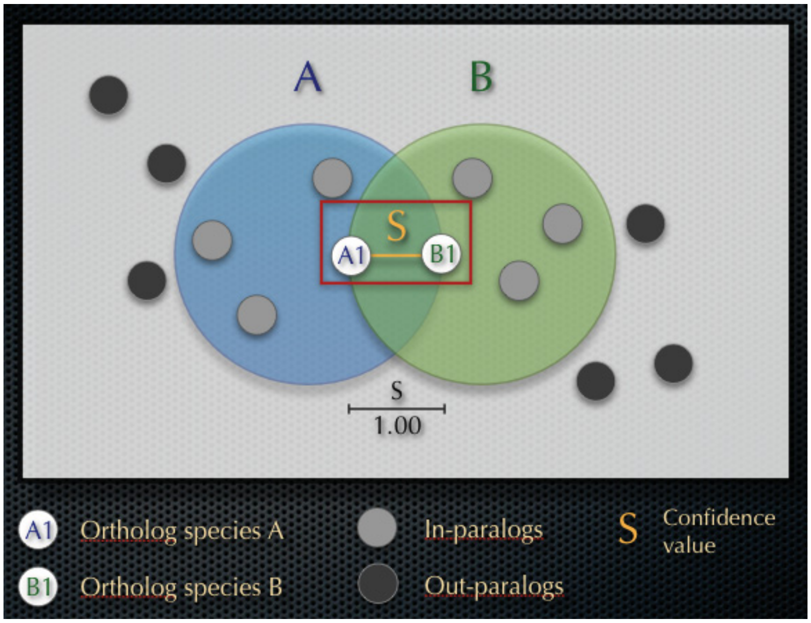
Smith-Waterman algorithm aligns every query locus in genome A against target locus in genome B. It then outputs Pairwise Similarity Score and repeat the alignments for other locuses until the whole input genome are processed.

Figure 3.2. Illustration showing how InParanoid generate pairwise clusters of orthologous genes (COGs) from pairwise similarity score generated by all-against all Smith Waterman peptide alignment (adapted from Remm *et al*., 2001)


Pairwise Similarity Score
Mutually best-scoring sequence pairs (Bi-directionally best hits)
Central Point
Find additional orthologs around it
(high to low)
List of pairwise COGs from genome **A** and **B**

InParanoid utilizes pairwise similarity score given my Smith-Waterman algorithm to find mutually best-scoring sequence pairs that are bi-directionally best hits and then find additional orthologs around the observed score. InParanoid then output list of pairwise clusters of orthologous genes from genome A and B.

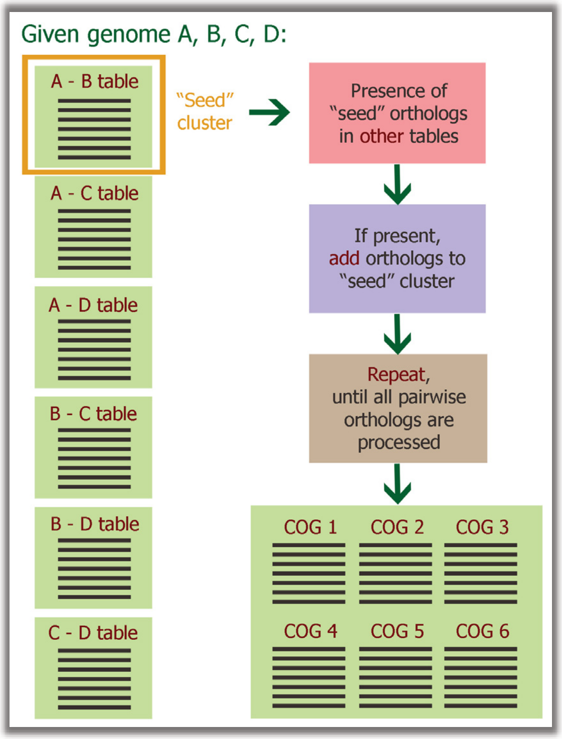
Figure 3.3. Illustration showing how InParanoid assign confidence value (S) to each ortholog pair (adapted from Remm *et al*., 2001)



InParanoid assigns confidence values (S) to each ortholog pairs to differentiate true orthologs from in-paralogs and out-paralogs.

QuickParanoid (http://pl.postech.ac.kr/QuickParanoid/) takes pairwise COGs results from InParanoid and merges the clusters with similar orthologs together into multiple species COGs. The program chooses one pairwise COG SQL table from one pair of species and uses this table as the “Seed Cluster” (Figure 3.4). It then finds presence of seed orthologs in other SQL tables. If present, the QuickParanoid adds the orthologs to Seed Cluster. This process will repeat until all pairwise orthologs in InParanoid results are processed. The result of ortholog clustering is displayed in the MultiParanoid program (Alexeyenko *et al*, 2006) output format and should be redirected to a text format, which later on was parsed to display only COG ID, species and gene name.

Figure 3.4. Illustration showing how QuickParanoid merge pairwise COGs from InParanoid output into multiple species COGs


Given genome A, B, C, D:
A - B table
"Seed" cluster
Presence of "seed" orthologs in other tables
A - C table
If present, add orthologs to "seed" cluster
A - D table
Repeat, until all pairwise orthologs are processed
B - C table
B - D table
C - D table
COG 1
COG 2
COG 3
COG 4
COG 5
COG 6

QuickParanoid merges clusters of orthologous genes from two species into multiple species cluster of orthologous genes. It chooses a pair of orthologous genes from two species SQL table, e.g. in genome A and B, then uses this table as "seed cluster". Then, it finds presence of seed orthologs in SQL tables of other pairwise species and adds them to seed cluster.

Commonly, pattern matching and pattern discovery algorithms are mostly used for regulatory motif prediction. In pattern matching, a common pattern is obtained from a collection of TFBSs. These binding sites are aligned and a consensus sequence or Position Weight Matrix (PWM), also called Position Specific Weight Matrix (PSWM) is generated and used as a representation of the motif or common pattern. The nucleotide base arrangement of each column of the alignment, which is represented by International Union of Pure and Applied Chemistry or IUPAC (http://www.iupac.org/) letter codes, is assigned for each location of the consensus sequence. However, consensus sequences do not capture the quantitative variability of TFBS; hence, the specificity and sensitivity of this approach is not optimal. Most consensus sequences of plant TFBS models are available in the public databases, such as TRANSFAC (Wingender *et al*., 2000), JASPAR (Vlieghe *et al*., 2006), PLACE (Higo *et al.,* 1998) and PlantCARE (Lescot *et al*., 2002) which are databases of transcription factors binding sites and DNA-binding profiles in eukaryotes.

Another approach is called pattern discovery and with this method motifs are detected from a collection of unaligned sequences. Although the objective functions for both pattern matching and pattern discovery are quite alike, the method of searching the space of potential alignments is very distinct. The MEME (Bailey and Elkan, 1994) and Gibbs Sampler (Lawrence *et al*., 1993) software are some examples of the tools used for motif discovery. MEME is an Expectation-Maximization (EM) method that takes into account all sites of the training data at the same time and converges to a local maximum (Bailey and Elkan, 1994). It allows scientists to discover signals or motifs in DNA or protein sequences by inputting FASTA formatted sequences of interest that are suspected to have some motif similarity, such as a set of promoters from co-expressed and orthologous genes (Bailey *et al*., 2006; Lyons *et al*., 2000). MEME seeks up to three DNA motifs and chooses the width and number of occurrences of each motif so that the 'E-value' of the motif is minimized (Bailey *et al*., 2006).

The EM method has also been implemented in Gibbs Sampler. Gibbs sampler starts with random alignments, and then uses random sampling for one sequence at a time to gradually improve the alignments (Lawrence *et al*., 1993). The difference between Gibbs Sampler and MEME is that EM chooses the single instance that maximizes the expected value in MEME, whereas in Gibbs every instance has a certain probability to become selected. Although MEME and Gibbs Sampler are routinely fast, these methods are not guaranteed to yield the best solution or global optimum (Stormo, 2000).

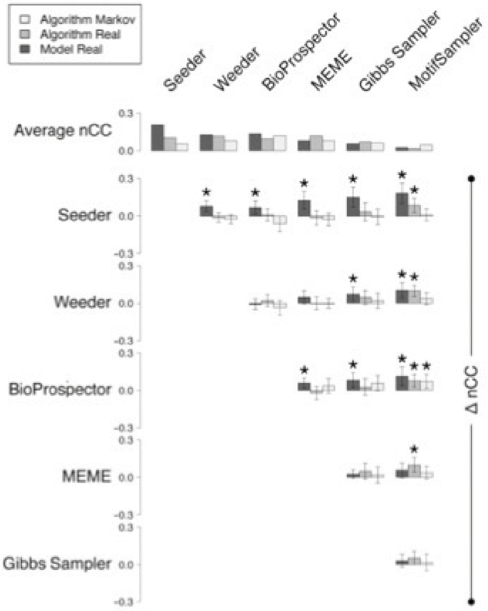
Cis-regulatory elements can be predicted by comparing and matching motif models in one or more DNA sequences. However, to predict new or unknown *cis*-regulatory element in eukaryotic promoters and identify patterns that are enriched, *de novo* motif finding algorithms are needed. Two examples of discriminative seeding *de novo* DNA motif discovery tools are Seeder (Fauteux *et al*., 2008) and Weeder (Pavesi *et al*., 2004). The Seeder algorithm uses sensitive and reliable statistics for the selection of motif seeds, a background model that is based on empirical distribution of SMD (Substring Minimal Distance), a good data structure that makes the computational approach for motif and background models relatively fast at moderate seed lengths (Fauteux *et al*., 2008). Seeder automatically avoids repetitive patterns such as low-complexity sequences, making it a good tool to find global optimum (Fauteux *et al*., 2008). It determines any conserved regions overrepresented in a group of promoters compared with a background set of promoters, e.g. all promoters in a genome. Seeder also processes a group of genes that have something in common, such as being co-regulated or members of the same pathway or paralogues. Seeder output some candidates for important DNA motifs, which later on can be run against other databases that have collections of the corresponding experimentally validated data.

Weeder was developed to allow scientists to characterize a conserved TFBS by setting parameter and statistical evaluation for DNA motif discovery (Pavesi *et al*., 2004). Since Weeder does not require any prior TFBS information and only require a set of at least two promoter regions and untranslated regions (UTRs), it can be used as a tool for *de novo* DNA motif discovery (Pavesi *et al*., 2004). The algorithm of this program utilizes a consensus that is generated by using the most repeated nucleotide in each location of the sites, but allows matches with a defined number of substitutions. The overlapping sequences among over-represented sequences are jointed to create longer motifs (Pavesi *et al*., 2004).

BioProspector (Liu *et al*., 2001) is a C-program that uses Gibbs sampling approach, and which uses Markov background to model base dependencies of non-motif bases. This approach significantly improves specificity of the reported motifs. It predicts regulatory sequence motifs from the upstream region of genes in the same gene expression pattern group. The parameters of the Markov background model are either estimated from user-specified sequences or pre-computed from the whole genome sequences (Liu *et al*., 2001)

Figure 3.5. Comparison of Seeder v.0.01 (Fauteux *et al*., 2008), Weeder v.1.3.1 (Pavesi *et al*., 2004), BioProspector v.1 (Liu *et al*., 2001), MEME v.3.5.4 (Bailey and Elkan, 1994), Gibbs Motif Sampler v.3.03.003 (Lawrence *et al*., 1993), and Motif Sampler v.3.2 (Thijs *et al*., 2001)

Average benchmarking scores and pairwise differences between motif discovery tools.



Fauteux F et al. Bioinformatics 2008;24:2303-2307

© 2008 The Author(s)

Figure above shows average benchmarking scores and pairwise differences between motif discovery tools. Average nucleotide-level Pearson correlation coefficient (nCC) and pairwise differences (Δ nCC) for six motif discovery tools tested on three benchmark suites. Error bars correspond to 95% confidence intervals. Stars indicate significant differences (α=0.05) between scores (Fauteux *et al*., 2008).

In order to discover DNA motifs, it is important to consider the length of the motif. Known motifs are commonly 4-20 bp long. The longer a motif is, and the more similarities the motif instances have, the easier it is to detect it in a background. Short motifs are more difficult to identify, as it is more likely to look like background or rather that it will occur randomly in the sample sequences (Bailey *et al*., 2006).

Although detecting motifs looks pretty straightforward, to differentiate true sites from the background is not easy (Fauteux *et al*., 2008). Matching multiple motifs could give a more accurate prediction to identify potential *cis*-regulatory elements (Guhathakurta, 2006).

3.3.2.2. Public databases for *cis*-regulatory elements

There are numerous online public databases that provide information on *cis*-regulatory elements. Some of them are as follows. PlantCARE database can provide locations for known *cis*-regulatory elements, enhancers and repressors and also provide tools for *in silico* analysis or promoter sequences (Lescot *et al*., 2002, Rombauts *et al*., 1999). PlnTFDB is an integrative plant transcription factor database that harbors large sets of transcription factors of several plant species (Riano-Pachon *et al*., 2007). PLACE is a database that collects and provides various *cis*- and *trans*- acting regulatory DNA elements explained in earlier studies. Plant promoter sequences can be downloaded from public database such as PlantProm DB (Shahmuradov *et al*., 2002). Various plant whole-genome sequences that are publically available can be downloaded on Phytozome v8.0 (http://www.phytozome.net/) (Goodstein *et al.,* 2012).

Since there are many public databases available, it becomes easier to collect different information given by different databases that can enhance our knowledge regarding plant promoters and motifs. In this study, a comparative promoter sequence analysis across 18 plant species was performed. Motif discovery was done based on orthologous relationship in between genes across 18 plant species and co-expression of genes in two specific tissues of soybean. By utilizing computational tools and existing information provided by public databases, the goal of this study was to understand better the transcription machinery that can regulate gene expression.

Chapter 4: Materials and Methods

4.1: Motif prediction in promoters of orthologous genes

4.1.1. Obtain genome, peptide and coding DNA sequence of 18 plant species

Publicly available genomic sequences of 18 plant species listed in Table 4.1.1.1 were obtained from Phytozome v8.0 (http://www.phytozome.net/) (Goodstein *et al.,* 2012) and used in accordance to their data usage policy. Analysis that include the identification of complete or whole genome sets of genomic features such as genes, gene families, regulatory elements, repeat structures, GC content, and whole-genome comparisons of regions of evolutionary conservation were permitted for genome sequences of 18 plant species included in this study. As genome assembly and annotation are continuously being improved, it is important to know which version of the assembly and annotation of each of the genome sequence used for the analysis.

Coding DNA Sequence (CDS), peptide sequence and promoter sequence 1000bp upstream of Transcription Start Site (TSS) were downloaded in FASTA format using the following steps: Tools > Biomart > Choose Dataset > Phytozome 8.0 Genomes > Filters > *Select Species* > Attributes > *Select Attributes* > Export Results to TSV > Go. All sequences were stored in the in-house MySQL database.

Table 4.1.1.1. List of 18 plant genomes used in this study along with their common names, version and genome paper references.

Species	Common name	Version	References
Arabidopsis lyrata	Lyre-leaved rock cress	JGI v1.0	Hu *et al.,* 2011
Arabidopsis thaliana	Thale cress	TAIR v10	AGI 2000; Swarbeck *et al.,* 2008
Brachypodium distachyon	Purple false brome	JGI /MIPS v1.0	International Brachypodium Initiative, 2010
Carica papaya	Papaya	ASGPB 2007	Ming *et al.,* 2008
Chlamydomonas reinhardtii	Green alga	JGI v5.0 assembly, annot v5.3.1 (Augustus u11.6)	Merchant *et al.,* 2007
Glycine max	Soybean	JGI Glyma1 assembly and Glyma 1.0 annotation	Schmutz *et al.,* 2010
Linus usitatissimum	Flax	BGI v1.0 on assembly v1.0	Wang *et al., 2012*
Malus domestica	Apple	GDR prediction v1.0 on Malusxdomestica assembly v1.0	Velasco *et al.,* 2010
Manihot esculenta	Cassava	Assembly version 4, JGI annotation v4.1	Prochnik *et al.,* 2012
Medicago truncatula	Barrel medic	Mt3.5v4 on assembly MedtrA17_3.5 from Medicago Genome Sequence Consortium	Young *et al.,* 2011
Oryza sativa	Rice	MSU Release 7.0 of the Rice Genome Annotation	Ouyang *et al., 2007*
Physcomitrella patens	Moss	JGI v1.1 and COSMOSS annotation v1.6	Rensing *et al.,* 2008
Populus trichocarpa	Poplar	JGI v3.0	Tuskan *et al.,* 2006
Selaginella moellendorfii	Spikemoss	JGI v1.0	Banks *et al.,* 2011
Sorghum bicolor	Sweet sorghum	MIPS/PASA v1.0	Paterson *et al.,* 2009
Vitis vinifera	Grapevine	March 2010 12x assembly and annotation from Genoscope	French-Italian Public Consortium for Grapevine Genome Characterization, 2007
Volvox carteri	Volvox	JGI v2.0	Prochnik *et al.,* 2010
Zea mays	Maize	Maize Golden Path B73 v2	Schnable *et al.*, 2009

4.1.2. Generate Clusters of Orthologous Genes (COGs)

4.1.2.1. All-against-all Smith-Waterman peptide sequence alignment

Peptide sequences of 18 plant species were submitted to an all-against-all sequence alignment using the Smith-Waterman algorithm with an E-value threshold of 1.0e-5, performed with a hardware-accelerated TimeLogic® Decipher® system (Active Motif Inc., Carlsbad, CA). The purpose was to determine the relationship of each peptide sequence of each genome versus each of the other genomes. Smith-Waterman algorithm aligned query locus against target locus and gave Pairwise Similarity Score for each locus pair alignment. Refer to Appendix 4.1.2.1 for the programming script. Parameters used were as follows; score = 50, max_scores = 50 and max_alignments = 500.

4.1.2.2. Generate cluster of orthologous genes between two species

InParanoid (Remm *et al.*, 2001) was used to generate clusters of orthologous genes between two species. From the pairwise similarity score in each locus pair alignment assigned by the Smith-Waterman algorithm, InParanoid took the mutually best pairwise similarity score from the species pair, then found additional orthologs around that score from high to low and lists pairwise COGs from genome species A and B.

Since peptide sequence was aligned using Smith-Waterman algorithm, BLAST step was skipped in InParanoid program. Parameters used were as follows; run_blast = 0, run_inparanoid = 1, use_bootstrap = 0, use_outgroup = 0, matrix = BLOSUM62, mysqltable = 1, score_cutoff = 40, confidence_cutoff = 40, group_overlap_cutoff = 0.5, grey_zone = 0 and segment_coverage_cutoff = 0.25. InParanoid produced output in SQL table format that contained cluster ID number, InParanoid score, species name, seed score, and sequence name was used to generate input files for QuickParanoid.

For each mutually best pairwise similarity score, InParanoid assigned a confidence value of 1.0 to assume that these pair is the true ortholog. Any genes with InParanoid confidence values less than 1.0 were eliminated from the cluster list, assuming that these are considered as paralogs and not orthologs. Refer to appendix 4.1.2.2 for programming script.

4.1.2.3. Generate clusters of orthologous genes across 18 plant species

To obtain COGs of 18 plant species, the QuickParanoid program (http://pl.postech.ac.kr/QuickParanoid/) was used to merge pairwise COGs from the InParanoid results into multiple species COGs by taking SQL tables produced by InParanoid. The QuickParanoid output listed out all COG IDs with each orthologous gene member in it. Refer to Appendix 4.1.2.3 for programming script.

4.1.2.4. Profiling *in silico* built clusters of orthologous genes with Gene Ontology

Gene Ontology annotations (The Gene Ontology Consortium, 2000) were used to investigate the functional relationship between the gene members in each cluster. Ontology files were downloaded from Gene Ontology website (http://www.geneontology.org/GO.downloads.ontology.shtml). Gene ID, GO ID, GO type and GO description were stored as MySQL table in the in-house database. GO IDs for each of the gene ID in every COG was retrieved to see if the gene members in the cluster have a similarity in function.

4.1.3. Retrieval of promoter sequences corresponding to orthologous genes in the cluster

Promoter sequences corresponding to the orthologous genes in the COG of 18 species were then retrieved from the 1000bp promoter sequences downloaded from Phytozome v8.0 (Goodstein *et al*., 2012) using BioPython Bio.SeqIO (Cock *et al*., 2009). Refer to Appendix 4.1.3 for programming script. The output file contained COG ID, Species, Gene ID and their corresponding promoter sequence. In addition, this output file contained the number of genes found in the COG file minus the genes in the *errors_18sp.txt*. The error file contains Gene IDs that were not found in the promoter files. Most of these promoters are less than 1000 bp, therefore it was eliminated when we obtained the promoter sequences from Phytozome v8.0. Some of these genes also have incomplete coding sequences that do not start with an ATG and some limitation on automated sequence annotation such as errors related to text processing.

Due to the large size of this file (i.e. promoter sequence of orthologous genes in 18 species), sequences were stored in the in-house MySQL database for easier query when predicting promoter motifs on the later step.

4.1.4. *De novo* motif prediction in promoters of orthologous genes

The three types of required files for running *de novo* motif finding with Seeder (Fauteux *et al*., 2008) were used: 1) an index file of words was generated to improve the performance of Hamming Distance (HD) calculation; 2) a background distribution file that contains the whole promoter sequence set of the 18 species; and 3) the promoter sequence file for each COG, containing all promoters from the member genes. These promoter sequence files for each COG constitutes the "positive set" in which there are assumed conserved motifs. The whole 18 species set of promoter sequences were used for the computation of background model only. Using a Python script (Refer to appendix 5.1.4), every promoter sequence of each gene members of each COGs were queried through the in-house MySQL database, where a temporary FASTA file was written and the percentage of numbers of *N* (i.e. ambiguous nucleotide base) in the sequence was calculated. Any promoter sequence that contains more than 50% of *N* was excluded. The Seeder motif finder script (*finder.pl*) was also included in this Python script (refer to Appendix 4.1.4 for programming script), therefore Seeder performed motif discovery in promoters of each COG automatically. Parameters used were as follows; seed_width = 6, strand = revcom, motif_width = 12 and n_motif = 10.

Motifs found by Seeder with *Q*-value < 1.0e-02 were considered significant and were used in the remainder of the analysis.

4.1.5. Match *in silico* predicted motifs with known motifs

Predicted significant motifs from Seeder were matched with experimentally validated motifs in PLACE: Plant *Cis*-Acting Regulatory DNA Elements Database (Higo *et al*., 1999). A table that contains information on motif abbreviation, motif sequence and PLACE motif ID was downloaded from PLACE database (http://ftp.dna.affrc.go.jp/pub/dna_place/place.dat). BioPython script using the dictionary module was then used to match the hexamer motif sequence found by Seeder to the motif sequence in the PLACE table.

4.2: Motif prediction in promoters of co-regulated plant genes

4.2.1. Obtain a list of co-regulated genes from the soybean RNA-Seq data and retrieving their promoter sequences

The RNA samples extracted from epidermis tissue and shoot apical meristem tissue of soybean cv. Clark, the 5-Leaflet mutant and the Glabrous mutant was sequenced using Illumina sequencing platform. Raw counts data that contain information on gene models, CDS hits, cDNA length and number of hits was obtained and used as one of the input file for RNA-Seq data analysis. This data was obtained and used in Haritika Majithia's M.Sc. thesis at McGill University; Determining Cell-Specific Gene Expression in two Soybean Mutants using Laser Capture Microdissection and Real-time PCR (Majithia, 2013).

RNA-Seq raw counts data was analyzed using DESeq (Anders and Huber., 2010). The DESeq package was downloaded from Bioconductor open-source website (http://www.bioconductor.org/packages/2.12/bioc/html/DESeq.html). DESeq estimates variance-mean dependence in count data from high-throughput sequencing assays and tests for differential expression based on a model using the negative binomial distribution. This program required two input files. The first one was a design file that describes which samples for which condition. The second file was the raw counts file. Raw counts file should not be normalized values. Technical replicate counts were summed up to get a single column that corresponds to a unique biological replicate. Since the raw counts file contained counts for more than one analysis (i.e. all three cultivars and two different tissues in one file), each individual file for each individual analysis was created. For example, one raw counts file that contains counts for soybean cv. Clark in epidermis tissue and shoot apical meristem tissue; another raw counts file that contains counts for the soybean 5-Leaflet mutant in epidermis tissue and shoot apical meristem; and another raw counts file that contains counts for the soybean Glabrous mutant in epidermis tissue and shoot apical meristem tissue. Duplicates of reads were removed using Unix command-line. The format of the input file for the analysis included gene model, raw counts from hits in epidermis tissue and raw counts from hits in shoot apical meristem tissue. Then, the following R script template was executed to run the program:

```
Rscript deseq.R -d design_filename -c rawcount_filename -o output/
```

Complete R script for DESeq can be found in Appendix 4.2.1.1.

Genes that were upregulated in shoot epidermis and shoot apical meristem tissues in soybean from cv. Clark, 5-Leaflet mutant and Glabrous mutant were obtained from RNA-Seq data analysis. Promoter sequences of each of the genes was retrieved from in-house MySQL database using BioPython (Cock *et al*., 2009). Refer to Appendix 4.2.1.2.

4.2.2. D*e novo* motif prediction in the promoters of co-regulated genes using Seeder

The three different files that were required for running *de novo* motif finding with Seeder (Fauteux *et al*., 2008) were generated: 1) an index file of words was generated to improve the performance of Hamming Distance (HD) calculation; 2) a background distribution file that contains the whole promoter sequence set of the Soybean; and 3) the promoter sequence file for each type of analysis, i.e. promoters sequence of co-regulated genes in epidermis tissue of soybean cv. Clark, promoters sequence file of co-regulated genes in shoot apical meristem tissue of soybean cv. Clark; and the same tissue respectively for the 5-Leaflet mutant and Glabrous mutant. These promoter sequence files for each set constitutes the "positive set" in which there were assumed conserved motifs. Using a Python script, every promoter sequence of each co-regulated genes in each set were queried through the in-house MySQL database and output as plain text FASTA format. A Perl motif finder script was used (Refer to Appendix 4.2.2). Motifs found by Seeder with Q-value < 1.0e-02 were considered significant and were used in the remainder of the analysis.

4.2.3. D*e novo* motif prediction in the promoters of co-regulated genes using MEME

MEME suite (Bailey *et al.,* 2009) was downloaded from http://meme.nbcr.net/meme/. MEME required one input file, which was the set of promoter sequences in FASTA format; and one background sequence file. In this case, soybean whole promoter sequences were used as the background set. Parameters used were as follows; motif_width = 6, number_motif = 10, and strand = revcom. To run MEME, a Unix script was used (Refer to appendix 4.2.3).

4.2.4. D*e novo* motif prediction in promoters of co-regulated genes using BioProspector

BioProspector (Liu *et al*., 2001) was downloaded from http://ai.stanford.edu/~xsliu/BioProspector. BioProspector required one input file, which was the set of promoter sequences in FASTA format; and one background sequence file. In this case, soybean whole promoter sequences were used as the background set. Parameters used were as follows; motif_width = 6, n = 40 and r = 5. BioProspector reinitialized the program to avoid local maxima and only report the top five among 40 tries. The following command template was executed:

```
./BioProspector -i inputfile –W 6 –b backgroundseqfile –n 40 –r 5 –o outputfile/
```

4.2.5. Validate *in silico* predicted motifs with known motifs

Predicted motifs using Seeder, MEME and BioProspector were matched with experimentally validated motifs in PLACE: Plant *Cis*-Acting Regulatory DNA Elements Database (Higo *et al*., 1999). Each Position Weight Matrix (PWM) generated by Seeder, MEME and BioProspector was pasted into STAMP web-interface (Mahony and Benos, 2007) one by one. The program was set to find the five best matches to each of the motifs in PLACE and motif logos were generated.

4.3: Public relational database of *cis*-regulatory motifs across 18 plants

4.3.1. Create MySQL relational database containing genome-promoter information

Schema was first created based on what needed to be included in the database. Database contents should include useful information that would compose a good bioinformatics environment to explore further any genome-promoter related information on 18 plant species used in this study.

Four tables were included in the database. The *cog_info* table contained information on COG ID, gene ID, species abbreviation and version of the COGs. The *go_info* table contained information on gene ID, GO ID and GO description. The *seeder_motif* table contained information on COG ID, motif sequence, q-value, gene reference of the motif, motif ID and version. *place_info* table contained information on motif ID, PLACE motif ID, PLACE motif sequence and gene reference of the motif.

Chapter 5: Results

5.1. Promoters of orthologous genes in 18 plants contain similar *cis*-regulatory motifs

Identifying motif that serves as the binding sites for transcription factors typically begins with creating sets of orthologous genes, as they tend to share functional similarity (Altenhoff *et al.,* 2012). In the first part of this study, motif discovery was performed on the promoters of orthologous genes across 18 plant species.

Promoter sequences consisting of 1000 bp upstream of the transcriptional start site (TSS), peptide sequences and coding DNA sequences (CDS) were obtained from the genome sequences of *Arabidopsis lyrata* (Hu *et al.,* 2011)*, Arabidopsis thaliana* (Swarbeck *et al.,* 2008)*, Brachypodium distachyon* (International Brachypodium Initiative, 2010)*, Carica papaya* (Ming *et al.,* 2008)*, Chlamydomonas reinhardtii* (Merchant *et al.,* 2007)*, Glycine max* (Schmutz *et al.,* 2010)*, Linus usitatissimum* (Wang *et al.,* 2012)*, Malus domestica* (Velasco *et al.,* 2010)*, Manihot esculenta* (Prochnik *et al.,* 2012)*, Medicago truncatula* (Young *et al.,* 2011)*, Oryza sativa* (Ouyang *et al.,* 2007)*, Physcomitrella patens* (Rensing *et al.,* 2008)*, Populus trichocarpa* (Tuskan *et al.,* 2006)*, Selaginella moellendorfii* (Banks *et al.,* 2011)*, Sorghum bicolor* (Paterson *et al.,* 2009)*, Vitis vinivera* (French-Italian Public Consortium for Grapevine Genome Characterization, 2007)*, Volvox carteri* (Prochnik *et al.,* 2010)*,* and *Zea mays* (Schnable *et al.,* 2009); retrieved from Phytozome version 8.0 (http://www.phytozome.net) (Goodstein *et al.,* 2012). All information was stored in the in-house database prior to analysis. A total number of 609,006 promoter sequences and a total number of 625,504 peptide sequences were obtained from the 18 plant genomes (Table 5.1.1). The length of promoter sequences used in this study was 1000bp. Any promoter sequences less than 1000 bp were eliminated from the analysis.

Table 5.1.1. List of 18 plant species used in this study along with their genome paper reference, number of genes, peptide sequences and promoter sequences.

Species	References	Genes	Peptide Seq	Promoter Seq (all)	Promoter Seq (1000bp cutoff)
Arabidopsis lyrata	Hu *et al*., 2011	41,063	32,670	41,063	32,456
Arabidopsis thaliana	Swarbeck *et al*., 2008	35,386	27,383	35,386	27,414
Brachypodium distachyon	International Brachypodium Initiative, 2010	31,029	25,967	31,029	26,548
Carica papaya	Ming *et al*., 2008	27,796	27,793	27,796	26,018
Chlamydomonas reinhardtii	Merchant *et al*., 2007	17,114	17,114	17,114	17,068
Glycine max	Schmutz *et al*., 2010	66,153	46,389	66,153	46,339
Linus usitatissimum	Wang *et al*., *2012*	43,471	43,471	43,471	43,101
Malus domestica	Velasco *et al*., 2010	63,541	63,517	63,541	50,305
Manihot esculenta	Prochnik *et al*., 2012	30,666	30,666	30,666	30,117
Medicago truncatula	Young *et al*., 2011	53,423	50,962	53,423	50,902
Oryza sativa	Ouyang *et al*., *2007*	66,338	55,766	66,338	55,986
Physcomitrella patens	Rensing *et al*., 2008	38,354	32271	38,354	31,779
Populus trichocarpa	Tuskan *et al*., 2006	45,033	40,688	45,033	40,340
Selaginella moellendorfii	Banks *et al*., 2011	22,285	22,285	22,285	22,174
Sorghum bicolor	Paterson *et al*., 2009	36,338	27,608	36,338	27,580
Vitis vinifera	French-Italian Public Consortium for Grapevine Genome Characterization, 2007	26,346	26,346	26,346	26,345
Volvox carteri	Prochnik *et al*., 2010	14,971	14,971	14,971	14,878
Zea mays	Schnable *et al*., 2009	39,656	39,637	39,656	39,656
Total		**698,963**	**625,504**	**698,963**	**609,006**

Local sequence alignments of 625,504 peptide sequence were performed using all-against-all Smith-Waterman algorithm and pairwise similarity scores were obtained. These were used as input for InParanoid (Remm et al., 2001) to generate Clusters of Orthologous Genes (COGs) of two species. A confidence value of 1.000 was assigned to the best match in order to differentiate true orthologs from paralogs. Output SQLtables generated by InParanoid were used by QuickParanoid to merge pairwise COGs to create composite COGs from 18 plant genome sequences. One dataset consisted of pairwise InParanoid output files between all species for analysis. To analyze *N* species, a dataset of *N (N-1)/2* InParanoid output files was needed. In the case of 18 species, 153 SQL table files were used in the analysis.

In total, 32,158 COGs were obtained. Most gene members in each COGs shared similarities in function, corroborated by Gene Ontology (GO) annotations. Out of 691,307 genes, 609,006 genes were included in those 32,158 COGs. There were 82,301 genes that were excluded in the remainder of the clustering orthologous genes step. To be included in the analysis, the promoter of those genes should have 1000 bp length, therefore any genes whose promoter was less than 1000 bp length were eliminated when we obtained the promoter sequences from Phytozome v8.0. Most of these genes have incomplete coding sequences that do not start with an ATG and there were some limitations on automated annotations such as errors obtained from text processing. The range of number of members in COGs is in between 1 to 19,461. There were 4,915 clusters that contained only 1 member and only 1 cluster contained 19,461 members, which is COG 57.

The corresponding promoter sequences of the orthologous genes included in the COGs were retrieved from in-house database using BioPython Bio.SeqIO module (Cock, et al., 2009). Due to short sequence length (less than 1000bp), 5,649 promoter sequences were not included in the data retrieved from Phytozome v8.0 and therefore gene members associated to them were eliminated from the COGs and further analysis.

Discriminative seeding *de novo* DNA motif discovery using Seeder (Fauteux et al., 2008) was performed separately for the set of promoter sequences of each of the COGs using a background model based on the complete set of promoters of 18 plants. Among all motifs predicted by Seeder, only motifs that have *Q*-value lower than 0.01 (*Q*-

value < 1.0e-02) were considered to be significant, which means that these motifs are present in a higher frequency in the positive set than in the background set (Fauteux et al., 2008). Significant motifs were identified in 805 COGs out of 32,158 COGs of 18 plants, i.e. only 2.4% COGs considered to have significant motifs and 97.60% COGs contain no significant motifs.

A table that contains PLACE motif abbreviations and ID with each corresponding motif sequences were downloaded from PLACE and stored into in-house database. Discovered motifs were then matched to consensus sequences of experimentally characterized plant *cis*-regulatory elements from the PLACE database (Higo et al., 1998). Seeder has detected 2,823 significant motifs, in which 2,191 of them matched to the experimentally validated motifs in PLACE, while 632 other motifs had no match, i.e. 22.4% of found motifs had no match and 77.6% matched with known motifs. They are thus novel putative *cis*-regulatory motifs. Table 5.1.2. below shows some examples of common motif that was found in the promoters of orthologous genes within each cluster. Each cluster shares the same functional annotation represented by GO ID and GO term.

Table 5.1.2. Motifs found in promoters of orthologous genes.

COG ID	Species	Gene ID	GO ID	GO Term	Motif Seq	Motif Annot.
6405	*A. lyrata*	939346	6457	Protein Folding	ACACGT	ABRELATERD1
	A. thaliana	AT3G62600				
	B. distachyon	Bradi2g34950				
	C. papaya	evm.TU.supercontig_18.17				
	G. max	Glyma02g01730				
	L. usitatissimum	Lus10022297.g				
	M. domestica	MDP0000330032				
	M. esculenta	cassava4.1_010980m.g				
	O. sativa	LOC_Os05g06440				
	P. patens	Pp1s298_70V6				
	S. moellendorfii	109399				
	S. bicolor	Sb09g004380				
	V. vinifera	GSVIVG01028094001				
	Z. mays	GRMZM2G129218				
8411	*A. lyrata*	483266	3723	RNA Binding	CGGCCC	SORLIP2AT
	A. thaliana	AT1G58380				
	B. distachyon	Bradi1g04660				
	C. papaya	evm.TU.supercontig_1.357				
	G. max	Glyma02g09370				
	L. usitatissimum	Lus10014909.g				
	M. domestica	MDP0000245640				
	M. esculenta	cassava4.1_013889m.g				
	O. sativa	LOC_Os03g59310				
	P. patens	Pp1s10_190V6				
	S. moellendorfii	67484				
	S. bicolor	Sb01g004250				
	V. carteri	Vocar20008601m.g				
	Z. mays	GRMZM2G168149				
2078	*A. lyrata*	327156	5975	Carbohydrate Metabolism	CACATG	MYCATERD1
	A. thaliana	AT1G02800				
	B. distachyon	Bradi1g35220				
	C. papaya	evm.TU.contig_27379.2				
	G. max	Glyma02g01990				
	L. usitatissimum	Lus10001666.g				
	M. domestica	MDP0000147635				
	M. esculenta	cassava4.1_003698m.g				
	O. sativa	LOC_Os01g12070				
	P. patens	Pp1s100_137V6				
	S. moellendorfii	117027				
	S. bicolor	Sb02g024050				
	V. vinifera	GSVIVG01009881001				
	V. carteri	Vocar20000570m.g				

5.2 Promoters of co-regulated genes in shoot epidermis and shoot apical meristem of soybean contain conserved *cis*-regulatory motifs

Another approach for motif discovery typically begins with a group of putatively co-regulated genes. Promoters of co-regulated genes are likely to be responsive to the same pathway and therefore to share common regulatory motifs (Hudson *et al.,* 2003). In the second part of this study, motif discovery was performed in the promoters of co-regulated genes in three different Soybean cv. Clark, 5-Leaflet mutant and Glabrous mutant.

The RNA samples extracted from shoot epidermis tissue and shoot apical meristem tissue of soybean cv. Clark, the 5-Leaflet mutant and the Glabrous mutant were sequenced using the Illumina HiSeq sequencing platform. The experimental part was done by Haritika Majithia for her M.Sc. thesis at McGill University. Raw counts data that contain information on gene models, CDS hits, cDNA length and number of hits were obtained and used as one of the input files for RNA-Seq data analysis. Differential expression test based on a model using the negative binomial distribution using DESeq (Anders and Huber, 2010) was performed in three sets of data. Differential expression comparison was performed between one tissue to the other tissue within each cultivar, i.e. shoot epidermis *vs*. shoot apical meristem; for each different cultivar (cv. Clark, 5-Leaflet mutant and Glabrous mutant), independently. The main biological question was to see which genes are upregulated in each tissue (i.e. comparing shoot epidermis with shoot apical meristem) in each cultivar.

To identify genes that were significantly upregulated, a cutoff q-value of 0.05 was used. Results of differential expression analysis using DESeq (Anders and Huber, 2010) indicated that there were; 153 upregulated genes in shoot apical meristem tissue of cv. Clark as compared to shoot epidermis tissue of the same cultivar; 72 upregulated genes in shoot apical meristem tissue of 5-Leaflet mutants as compared to shoot epidermis tissue of the same cultivar; 134 upregulated genes in shoot apical meristem tissue of Glabrous mutant as compared to shoot epidermis tissue of the same cultivar; 214 upregulated genes in shoot epidermis tissue of cv. Clark as compared to shoot apical meristem tissue of the same cultivar; 218 upregulated genes in shoot epidermis tissue of 5-Leaflet mutant as compared to shoot apical meristem tissue of the same cultivar; and 110 upregulated genes of shoot epidermis tissue of Glabrous mutant as compared to shoot apical meristem tissue of the same cultivar.

Promoters of the upregulated genes were then retrieved from in-house database using BioPython Bio.SeqIO module (Cock, et al., 2009). Since we are interested in the 1000 bp region upstream of transcription start site, some of the promoter sequences were eliminated due to their short sequence length. Corresponding to the upregulated genes in shoot epidermis tissue compared with shoot apical meristem, the following numbers of promoter sequences were retrieved: 202 for cv. Clark, 199 for the 5-Leaflet mutant and 101 for the Glabrous mutant. In shoot apical meristem, 141 promoters sequence were retrieved for the cv. Clark, 65 for the 5-Leaflet mutant and 127 for the Glabrous mutant.

5.2.1. *De novo* motif discovery in promoters of co-regulated genes using Seeder

DNA motif discovery using Seeder (Fauteux *et al.,* 2008) was performed in the promoters of co-regulated genes in each of the tissue-cultivar set separately, i.e. shoot epidermis tissue of cv. Clark, shoot epidermis tissue of the 5-Leaflet mutant, shoot epidermis tissue of Glabrous mutant, shoot apical meristem tissue of cv. Clark, shoot apical meristem tissue of 5-Leaflet mutants, and shoot apical meristem of Glabrous mutant. Total number of motif searches was 6. The background model was based on the complete set of promoters of soybean that is 53,452 sequences. Among all motifs predicted by Seeder, only motifs that have Q-value lower than 0.01 (Q-value < 1.0e-02) were considered to be significant, which means that these motifs are present in a higher frequency in the positive set than in the background set (Fauteux et al., 2008). Statistically significant conserved *cis*-regulatory motifs were identified in gene promoters within tissue-cultivar sets. Discovered motifs were matched to consensus sequences of experimentally characterized plant *cis*-regulatory elements from PLACE database (Higo *et al*., 1998) using the STAMP suite of tools (Mahony and Benos, 2007). In default, STAMP returns the best five hits of motif.

In the promoters corresponding to upregulated genes in shoot epidermis tissue of cv. Clark, no motif was detected (Table 5.2.1.1).

In the promoters corresponding to upregulated genes in shoot epidermis tissue of the 5-Leaflet mutants, two motifs were detected (Table 5.2.1.1); AGATAT/ATATCT and

GGTACA/TGTACC. The first motif, AGATAT/ATATCT, was found in the promoters of 199 upregulated genes and the motif matched to one known motif in PLACE database; GATABOX (Lam *et al.,* 1989). The second motif, GGTACA/TGTACC, was found in the promoters of 199 upregulated genes and the motif matched to one known motif in PLACE database; CURECORECR (Quinn *et al.,* 2000).

In the promoters corresponding to upregulated genes in shoot epidermis tissue of Glabrous mutants, no motif was detected (Table 5.2.1.1).

In the promoters corresponding to upregulated genes in shoot apical meristem tissue of cv. Clark, motif CACGTA/TACGTG was detected (Table 5.2.1.1). This motif was found in the promoters of 141 upregulated genes and the motif matched to one known motif in PLACE database; ABRELATERD1 (Simpson *et al.,* 2003).

In the promoters corresponding to upregulated genes in shoot apical meristem tissue of the 5-Leaflet mutants, motif CACTGC/GCAGTG was detected (Table 5.2.1.1). This motif was found in the promoters of 65 upregulated genes and the motif matched to one known motif in PLACE database; CACTFTPPCA1 (Gowik *et al.,* 2004).

In the promoters corresponding to upregulated genes in shoot apical meristem tissue of Glabrous mutants, no motif was detected (Table 5.2.1.1).

Table 5.2.1.1 DNA motifs discovered by Seeder in the promoters of co-regulated genes in different tissues of three soybean cultivars

Cv.	Tissue	Seeder Motif	Motif Sites	Seeder *Q*-value[1]	PLACE ID[2]	STAMP *E*-value[3]
Clark	SE[4]	-	-	-	-	-
	SAM[5]	CACGTA	141	4.05E-02	ABRELATERD1	9.79E-01
5-Leaflet mutant	SE[4]	AGATAT	199	4.23E-02	GATABOX	9.68E-01
		GGTACA	199	8.47E-02	CURECORECR	9.64E-01
	SAM[5]	CAGTAG	65	2.86E-02	CACTFTPPCA1	9.81E-01
Glabrous mutant	SE[4]	-	-	-	-	-
	SAM[5]	-	-	-	-	-

[1] *Q*-value represents statistical significance of motif found by Seeder. [2] PLACE ID is the identifier of PLACE consensus sequence matching motif. [3] STAMP *E*-value represents expectation value of the STAMP alignment. [4] SE is Shoot Epidermis. [5] SAM is Shoot Apical Meristem.

5.2.2. *De novo* motif discovery in promoters of co-regulated genes using MEME

DNA motif discovery using MEME (Bailey *et al.,* 2009) was performed in the promoters of co-regulated genes in each of the tissue-cultivar set separately, i.e. shoot epidermis tissue of cv. Clark, shoot epidermis tissue of the 5-Leaflet mutant, shoot epidermis tissue of Glabrous mutant, shoot apical meristem tissue of cv. Clark, shoot apical meristem tissue of 5-Leaflet mutants, and shoot apical meristem of Glabrous mutant. Total number of motif searches was 6. The background model was based on the complete set of promoters of soybean that is 53,452 sequences. Among all motifs predicted by MEME, only motifs with *E*-values less than 0.001 were considered to be significant, as they are very unlikely to be random sequence artifacts. Significance of a motif is a function of the length of the pattern, number of times it occurs and degree of similarity among the occurrences (Bailey *et al.,* 2009). Statistically significant conserved *cis*-regulatory motifs were identified in gene promoters within tissue-cultivar sets. Discovered motifs were matched to consensus sequences of experimentally characterized plant *cis*-regulatory elements from PLACE database (Higo *et al*., 1998) using the STAMP suite of tools (Mahony and Benos, 2007). STAMP returns the best five hits of motif.

In the promoters corresponding to upregulated genes in shoot epidermis tissue of cv. Clark, two motifs were detected (Table 5.2.2.1); CCCC[A/C]C or G[T/G]GGGG, and CC[A/T]CCC or GGG[T/A]GG. The first motif, CCCC[A/C]C or G[T/G]GGGG, was found in promoters of 73 upregulated genes and it matched to five known motifs in PLACE database; SE1PVGRP18 (Keller *et al.,* 1994), BOXCPSAS1_3 (Ngai *et al.,* 1997), GCBP2ZMGAPC4 (Geffers *et al.,* 2000), SITEIIBOSPCNA (Kosugi *et al.,* 1995) and ABREAZMRAB28 (Busk *et al.,* 1997). The second motif, CC[A/T]CCC or GGG[T/A]GG, was found in the promoters of 73 upregulated genes and it matched to four known motifs in PLACE database; MYBPZM (Grotewold *et al.,* 1994), SITEIOSPCNA (Kosugi *et al.,* 1995), ABRECE1HVA22 (Shen *et al.,* 1995) and POLLEN2LELAT52 (Bate *et al.,* 1998). Refer to Supplementary Figure 5.2.2.1 for motif logos.

In the promoters corresponding to upregulated genes in shoot epidermis tissue of the 5-Leaflet mutants, motif CCCC[A/T]C or G[T/A]GGGG was detected (Table

5.2.2.1). This motif was found in the promoters of 88 upregulated genes and it matched to five known motifs in PLACE database; SE1PVGRP18 (Keller *et al.,* 1994), BOXCPSAS1_3 (Ngai *et al.,* 1997), GCBP2ZMGAPC4 (Geffers *et al.,* 2000), SITEIIBOSPCNA (Kosugi *et al.,* 1995) and ABREAZMRAB28 (Busk *et al.,* 1997). Refer to Supplementary Figure 5.2.2.2 for motif logo.

In the promoters corresponding to upregulated genes in shoot epidermis tissue of the Glabrous mutant, motif CCCC[A/C]C or G[T/G]GGGG was detected (Table 5.2.2.1). This motif was found in the promoters of 33 upregulated genes and it matched to five known motifs in PLACE database; SE1PVGRP18 (Keller *et al.,* 1994), BOXCPSAS1_3 (Ngai *et al.,* 1997), GCBP2ZMGAPC4 (Geffers *et al.,* 2000), SITEIIBOSPCNA (Kosugi *et al.,* 1995) and ABREAZMRAB28 (Busk *et al.,* 1997). Refer to Supplementary Figure 5.2.2.3 for motif logo.

In the promoters corresponding to upregulated genes in shoot apical meristem tissue of cv. Clark , two motifs were detected (Table 5.2.2.1); CCC[T/A]CC or GG[A/T]GGG, and [C/G]CCCCA or TGGGG[G/C]. The first motif, CCC[T/A]CC or GG[A/T]GGG, was found in the promoters of 94 upregulated genes and it matched to five known motifs in PLACE database; ACIPVPAL2 (Hatton *et al.,* 1995), BOXCPSAS1_3 (Ngai *et al.,* 1997), AMMORESIIUDCRNIA1 (Loppes and Radoux, 2001), PALBOXAPC (Logemann *et al.,* 1995) and SBOXATRBCS (Acevedo-Hernandez *et al.,* 2005). The second motif, [C/G]CCCCA or TGGGG[G/C], was found in the promoters of 52 upregulated genes and it matched to four known motifs in PLACE database; ARELIKEGHPGDFR2 (Eloma *et al.,* 2003), SE1PVGRP18 (Keller *et al.,* 1994), AMMORESVDCRNIA1 (Loppes and Radoux, 2001) and SEF3MOTIFGM (Allen *et al.,* 1989). Refer to Supplementary Figure 5.2.2.4 for motif logo.

In the promoters corresponding to upregulated genes in shoot apical meristem tissue of the 5-Leaflet mutants, three motifs were detected (Table 5.2.2.1); [A/G]GAG[A/G]G or C[T/C]CTC[T/C], CC[A/C]C[A/C]C or G[T/G]G[T/G]GG, and CCACCA or TGGTGG. The first motif, [A/G]GAG[A/G]G or C[T/C]CTC[T/C], was found in the promoters of 160 upregulated genes and it matched to five known motifs in PLACE database; CTRMCAMV35S (Pauli *et al.,* 2004), GAGA8HVBKN3 (Santi *et al.,* 2003*)*, GAGAGMGSA1 (Sangwan *et al.,* 2002), SORLIP5AT (Hudson and Quail, 2003) and PE3ASPHYA3 (Bruce and Quail., 1990). The second motif, CC[A/C]C[A/C]C or G[T/G]G[T/G]GG, was found in the promoters of 88 upregulated genes and it matched to five known motifs in PLACE database; RBCSBOX2PS (Fluhr *et al.,* 1986), LREBOX2PSRBCS3 (Green *et al.,* 1987), SE1PVGRP18 (Keller *et al.,* 1994), BOXCPSAS1_3 (Ngai *et al.,* 1997) and ACGTSEED2 (Bustos and Thomas., 1993). The third motif, CCACCA or TGGTGG, was found in the promoters of 31 upregulated genes and it matched to five known motifs in PLACE database; POLLEN2LELAT52 (Bate and Twell., 1998), ACIIPVPAL2 (Hatton *et al.,* 1995), SITEIOSPCNA (Kosugi *et al.,* 1995), ABRECE1HVA22 (Shen and Ho., 1995) and ACIPVPAL2 (Hatton *et al.,* 1995). Refer to Supplementary Figure 5.2.2.5 for motif logo.

In the promoters corresponding to upregulated genes in shoot apical meristem tissue of the Glabrous mutant, five motifs were detected (Table 5.2.2.1); GAGAGA or TCTCTC, CCCC[A/C]C or G[T/G]GGGG, CCCTCC or GGAGGG, ACACAC or GTGTGT, and CCA[A/C]CC or GG[T/G]TGG. The first motif, GAGAGA or TCTCTC, was found in the promoter of 212 upregulated genes and it matched to five known motifs in PLACE database; CTRMCAMV35S (Pauli *et al.,* 2004), GAGA8HVBKN3 (Santi *et al.,* 2003*)*, GAGAGMGSA1 (Sangwan *et al.,* 2002), NDEGMSAUR (Li *et al.,* 1994) and ARFAT (Ulmasov *et al.,* 1999). The second motif, CCCC[A/C]C or G[T/G]GGGG, was found in the promoters of 70 upregulated genes and it matched to five known motifs in PLACE database; SE1PVGRP18 (Keller *et al.,* 1994), BOXCPSAS1_3 (Ngai *et al.,* 1997), GCBP2ZMGAPC4 (Geffers *et al.,* 2000), SITEIIBOSPCNA (Kosugi *et al.,* 1995) and ABREAZMRAB28 (Busk *et al.,* 1997). The third motif, CCCTCC or GGAGGG, was found in the promoters of 38 upregulated genes and it matched to five known motifs in PLACE database; PALBOXAPC (Logemann *et al.,* 1995), SBOXATRBCS (Acevedo-Hernandez *et al.,* 2005), CMSRE1IBSPOA (Morikami *et al.,* 2005), IDRSZMFER1 (Petit *et al.,* 2001) and BOXICHS (Block *et al.,* 1990). The fourth motif, ACACAC or GTGTGT, was found in the promoters regions of 122 upregulated genes and it matched to five known motifs in PLACE database; ACGTSEED2 (Bustos and Thomas., 1993), GLUTEBP1OS (Croissant-Sych and Okita, 1996) BOXCPSAS1_2 (Ngai *et al.,* 1997), NAPINMOTIFBN (Ericson *et al.,* 1991) and SP8BFIBSP8AIB (Ishiguro and Nakamura, 1992). The fifth motif, CCA[A/C]CC or GG[T/G]TGG, was found in the promoters of 74 upregulated genes and it matched to four known motifs in PLACE database; MYBPZM (Grotewold *et al.,* 1994), ACIIPVPAL2 (Hatton *et al.,* 1995), L4DCPAL1 (Takeda *et al.,* 2002), PALBOXLPC (Logemann *et al.,* 1995) and MYBPLANT (Sablowski *et al.,* 1994). Refer to Supplementary Figure 5.2.2.6 for motif logo.

Table 5.2.2.1. DNA motifs discovered by MEME in the promoters of co-regulated genes in different tissues of three soybean cultivars

Cv.	Tissue	Motif	Motif Sites	MEME *E*-value[1]	PLACE ID[2]	STAMP *E*-value[3]
Clark	SE[4]	For: CCCC[A/C]C Rev: G[T/G]GGGG	73	2.60E-06	SE1PVGRP18	1.66E-06
					BOXCPSAS1_3	1.09E-05
					GCBP2ZMGAPC4	6.39E-05
					SITEIIBOSPCNA	6.39E-05
					ABREAZMRAB28	1.09E-04
		For: CC[A/T]CCC Rev: GGG[T/A]GG	72	1.70E+05	MYBPZM	1.37E-05
					SITEIOSPCNA	1.63E-04
					ABRECE1HVA22	2.30E-04
					POLLEN2LELAT52	3.24E-04
	SAM[5]	For: CCC[T/A]CC Rev:GG[A/T]GGG	94	4.30E-16	ACIPVPAL2	1.40E-06
					BOXCPSAS1_3	1.65E-04
					AMMORESIIUDCRNIA1	2.28E-04
					PALBOXAPC	2.73E-04
					SBOXATRBCS	3.10E-04
		For: [C/G]CCCCA Rev: TGGGG[G/C]	52	5.70E+04	ARELIKEGHPGDFR2	5.49E-06
					SE1PVGRP18	5.00E-05
					AMMORESVDCRNIA1	9.31E-04
					SEF3MOTIFGM	1.18E-03

Table 5.2.2.1 continued

5-Leaflet Mutant	SE[4]	For: CCCC[A/T]C Rev:G[T/A]GGGG	88	1.60E+04	SE1PVGRP18	3.92E-06
					BOXCPSAS1_3	2.33E-05
					GCBP2ZMGAPC4	1.26E-04
					SITEIIBOSPCNA	1.26E-04
					ABREAZMRAB28	2.11E-04
	SAM[5]	For: [A/G]GAG[A/G]G Rev: C[T/C]CTC[T/C]	160	1.60E-20	CTRMCAMV35S	1.18E-07
					GAGA8HVBKN3	1.93E-06
					GAGAGMGSA1	2.81E-06
					SORLIP5AT	3.44E-04
					PE3ASPHYA3	3.88E-04
		For: CC[A/C]C[A/C]C Rev: G[T/G]G[T/G]GG	88	1.70E-06	RBCSBOX2PS	3.65E-06
					LREBOX2PSRBCS3	5.20E-06
					SE1PVGRP18	1.97E-05
					BOXCPSAS1_3	6.54E-04
					ACGTSEED2	1.09E-03
		For: CCACCA Rev: TGGTGG	31	7.80E+05	POLLEN2LELAT52	1.18E-08
					ACIIPVPAL2	8.44E-08
					SITEIOSPCNA	1.75E-05
					ABRECE1HVA22	3.89E-05
					ACIPVPAL2	6.73E-05

Table 5.2.2.1 continued

Glabrous Mutant	SE[4]	For: CCCC[A/C]C Rev: G[T/G]GGGG	33	1.3+003	SE1PVGRP18	5.13E-06
					BOXCPSAS1_3	2.96E-05
					GCBP2ZMGAPC4	1.56E-04
					SITEIIBOSPCNA	1.56E-04
					ABREAZMRAB28	2.59E-04
	SAM[5]	For:GAGAGA Rev: TCTCTC	212	7.60E-59	CTRMCAMV35S	1.18E-08
					GAGA8HVBKN3	2.31E-07
					GAGAGMGSA1	3.38E-07
					NDEGMSAUR	8.90E-07
					ARFAT	1.30E-05
		For: CCCC[A/C]C Rev: G[T/G]GGGG	70	1.20E-17	SE1PVGRP18	6.73E-06
					BOXCPSAS1_3	3.75E-05
					GCBP2ZMGAPC4	1.94E-04
					SITEIIBOSPCNA	1.94E-04
					ABREAZMRAB28	3.18E-04
		For: CCCTCC Rev: GGAGGG	38	5.10E-01	PALBOXAPC	1.30E-05
					SBOXATRBCS	1.75E-05
					CMSRE1IBSPOA	6.79E-05
					IDRSZMFER1	9.73E-05
					BOXICHS	1.40E-04
		For: ACACAC Rev: GTGTGT	122	2.60E+01	ACGTSEED2	4.90E-08
					GLUTEBP10S	1.48E-07
					BOXCPSAS1_2	3.38E-07
					NAPINMOTIFBN	6.24E-06
					SP8BFIBSP8AIB	1.75E-05
		For: CCA[A/C]CC Rev: GG[T/G]TGG	74	4.70E+06	MYBPZM	2.07E-08
					ACIIPVPAL2	3.62E-07
					L4DCPAL1	3.62E-07
					PALBOXLPC	6.39E-06
					MYBPLANT	1.94E-05

[1] MEME *E*-value represents statistical significance of motif found by MEME. [2] PLACE ID is the identifier of PLACE consensus sequence matching motif. [3] STAMP *E*-value represents expectation value of the STAMP alignment. [4] SE is Shoot Epidermis. [5] SAM is Shoot Apical Meristem.

Motif SITEIIBOSPCNA (Kosugi *et al.,* 1995) was found in the promoters of 72, 88 and 33 upregulated genes corresponding to upregulated genes in shoot epidermis of cv. Clark, 5-Leaflet mutants and Glabrous mutants, respectively; and in the promoters of 70 upregulated genes in shoot apical meristem tissue of Glabrous mutants. Motif GAGAGMGSA1 (Sangwan *et al.,* 2002) and GAGA8HVBKN3 (Santi *et al.,* 2003*)* was found in the promoters of 160 and 212 upregulated genes in shoot apical meristem tissue of the 5-Leaflet and Glabrous mutant, respectively. Motif ACIPVPAL2 (Hatton *et al.,* 1995) and ACIIPVPAL2 (Hatton *et al.,* 1995) were found in the promoters corresponding to upregulated genes in shoot apical meristem of cv. Clark, 5-Leaflet mutants and Glabrous mutants. Motif NDEGMSAUR (Li *et al.,* 1994) and ARFAT (Ulmasov *et al.,* 1999) were found in the shoot apical meristem tissue of Glabrous mutant. Motif AMMORESVDCRNIA1 and AMMORESIIUDCRNIA1 (Loppes and Radoux, 2001) were found in the shoot apical meristem of cv. Clark.

5.2.3. *De novo* motif discovery in promoters of co-regulated genes using BioProspector

DNA motif discovery using BioProspector (Liu et al., 2001) was performed in the promoters of co-regulated genes in each of the tissue-cultivar set separately, i.e. shoot epidermis tissue of cv. Clark, shoot epidermis tissue of the 5-Leaflet mutant, shoot epidermis tissue of Glabrous mutant, shoot apical meristem tissue of cv. Clark, shoot apical meristem tissue of 5-Leaflet mutants, and shoot apical meristem of Glabrous mutant. Total number of motif searches was 6. The background model was based on the complete set of promoters of soybean that is 53,452 sequences. BioProspector output *E*-values that indicates the significance of each motif based on a motif score distribution calculated by a Monte Carlo method (Liu *et al.,* 2001). Statistically significant conserved *cis*-regulatory motifs were identified in gene promoters within tissue-cultivar sets. Discovered motifs were matched to consensus sequences of experimentally characterized plant *cis*-regulatory elements from PLACE database (Higo *et al*., 1998) using the STAMP suite of tools (Mahony and Benos, 2007). STAMP returns the best five hits of motifs.

In the promoters corresponding to upregulated genes in shoot epidermis tissue of cv. Clark, motif CATTCA/TGAATG was detected (Table 5.2.3.1). This motif was found in the promoters of 127 upregulated genes and it matched to five known motifs in PLACE database; ARELIKEGHPGDFR2 (Eloma *et al.,* 2003), RYREPEATBNNAPA (Ezcurra *et al.,* 1999), WUSATAg (Kamiya *et al.,* 2003), RYREPEATGMGY2 (Lalievre *et al.,* 1992) and RYREPEATLEGUMINBOX (Fujiwara *et al.,* 1994). Refer to Supplementary Figure 5.2.3.1 for motif logo.

In the promoters corresponding to upregulated genes in shoot epidermis tissue of the 5-Leaflet mutants, three motifs were detected (Table 5.2.3.1); GATTCA/TGAATC, CATACA/TGTATG, and AGTTAG/CTAACT. The first motif, GATTCA/TGAATC, was found in the promoters of 125 upregulated genes and it matched to five known motifs in PLACE database; EIN3ATERF1 (Solano *et al.,* 1998), GCM40SGLUUB1, L1DCPAL1 (Takeda *et al.,* 2002), AGMOTIFNTMYB2 (Sugimoto *et al.,* 2003), and ARR1AT (Sakai *et al.,* 2000). The second motif, CATACA/TGTATG, was found in the promoters of 122 upregulated genes and it matched to five known motifs in PLACE database; S2FSORPL21 (Zhou *et al.,* 1992), SORLIP4AT (Hudson *et al.,* 2003), RYREPEATBNNAPA (Ezcurra *et al.,* 1999, JASE2ATOPR1 (He *et al.,* 2001), and SORLREP3AT (Hudson *et al.,* 2003). The third motif, AGTTAG/CTAACT, was found in the promoters of 87 upregulated genes and it matched to five known motifs in PLACE database; SITE3SORPS1 (Villain *et al.,* 1994), MYBATRD22 (Abe *et al.,* 1997), MYB2AT (Urao *et al.,* 1993), MYB1LEPR (Chakravarty *et al.,* 2003), and LREBOXIPCCHS1 (Schulze-Lefert *et al*, 1989). Refer to Supplementary Figure 5.2.3.2 for motif logo.

In the promoters corresponding to upregulated genes in shoot epidermis tissue of the Glabrous mutants, two motifs were detected (Table 5.2.3.1); TCATAC/GTATGA and CTAACT/AGTTAG. The first motif, TCATAC/GTATGA, was found in the promoters of 67 upregulated genes and it matched to five known motifs in PLACE database; SORLIP4AT (Hudson *et al.,* 2003), S2FSORPL21 (Lagrange *et al.,* 1997), LS7ATPR1 (Despres *et al.,* 2000), JASE2ATOPR1 (He *et al.,* 2001), and RSEPVGRP1 (Elmayan *et al.,* 1995). The second motif, CTAACT/AGTTAG, was found in the promoters of 73 upregulated genes and it matched to five known motifs in PLACE database; SITE3SORPS1 (Villain *et al.,* 1994), MYB2AT (Urao *et al.,* 1993), MYB1LEPR

(Chakravarty *et al.,* 2003), MYBATRD22 (Abe *et al.,* 1997) and MYB26PS (Uimari *et al.,* 1997). Refer to Supplementary Figure 5.2.3.3 for motif logo.

In the promoters corresponding to upregulated genes in shoot apical meristem tissue of cv. Clark, four motifs were detected (Table 5.2.3.1); GCGTGC/GCACGC, GCAAGC/GCTTGC, GATCGA/TCGATC and CTAACT/AGTTAG. The first motif, GCGTGC/GCACGC, was found in the promoters of 73 upregulated genes and it matched to five known motifs in PLACE database; RHERPATEXPA7 (Kim *et al.,* 2006), ABRE2HVA22 (Shen *et al.,* 1995), GBOX10NT (Ishige *et al.,* 1999), ABASEED1 (Chung and Thomas, 1993) and -284MOTIFZMSBE1 (Kim *et al.,* 1999). The second motif, GATCGA/TCGATC, was found in the promoters of 87 upregulated genes and it matched to four known motifs in PLACE database; ALF1NTPARC (Sakai *et al.,* 1998), ASF1NTPARA (Sakai *et al.,* 1998), PASNTPARA (Kusaba *et al.,* 1996) and CE3OSOSEM (Hobo *et al.,* 1999). The third motif, CTAACT/AGTTAG, was found in the promoters of 82 upregulated genes and it matched to five known motifs in PLACE database; NONAMERATH4 (Chaubet *et al.,* 1996), RNFG1OS (Yin *et al.,* 1995), ABAREG2 (Nunberg *et al*., 1993*)*, AMMORESIVDCRNIA (Loppes and Radoux*,* 2001) and SV40COREENHAN (Weiher *et al.,* 1983). The fourth motif, CTAACT/AGTTAG, was found in the promoters of 99 upregulated genes and it matched to five known motifs in PLACE database; SITE3SORPS1 (Villain *et al.,* 1994), MYB2AT (Urao *et al.,* 1993), MYB1LEPR (Chakravarty *et al.,* 2003), MYBATRD22 (Abe *et al.,* 1997) and MYB26PS (Uimari *et al.,* 1997). Refer to Supplementary Figure 5.2.3.4 for motif logo.

In the promoters corresponding to upregulated genes in shoot apical meristem tissue of the 5-Leaflet mutants, three motifs were detected (Table 5.2.3.1); ACTCGG/CCGAGT, GGCGCG/CGCGCC and AGATAG/CTATCT. The first motif, ACTCGG/CCGAGT, was found in the promoters of 40 upregulated genes and it matched to seven known motifs in PLACE database; DRE1COREZMRAB17 (Busk *et al.,* 1997), LTRECOREATCOR15 (Baker *et al.,* 1994), C1MOTIFZMBZ2 (Bodeau *et al.,* 1996), RSRBNEXTA (Elliott *et al.,* 1998), ABREMOTIFIIIOSRAB16B (Ono *et al.,* 1996), PREATPRODH (Satoh *et al.,* 2002) and GCN4OSGLUB1 (Washida *et al.,* 1999). The second motif, GATCGA/TCGATC, was found in the promoters of 30 upregulated genes and it matched to five known motifs in PLACE database; RE1ASPHYA3 (Bruce *et al.,* 1991), ABRECE3ZMRAB28 (Busk *et al.,* 1997), PE3ASPHYA3 (Bruce *et al.,* 1990), PE2FNTRNR1A (Lincker *et al.,* 2004) and OCTAMOTIF2 (Chaubet *et al.,* 1986). The third motif, AGATAG/CTATCT, was found in the promoters of 48 upregulated genes and it matched to six known motifs in PLACE database; EVENINGAT (Harmer *et al.,* 2000), RGATAOS (Yin *et al.,* 1997), IBOXLSCMCUCUMISIN (Yamagata *et al.,* 2002), SE2PVGRP1 (Elmayan, et al., 1995), CAATBOX2, and CTRMCAMV35S (Pauli *et al.,* 2004). Refer to Supplementary Figure 6.2.3.5 for motif logo.

In the promoters corresponding to upregulated genes in shoot apical meristem tissue of the Glabrous mutants, motif CCAAAC/GTTTGG was detected (Table 5.2.3.1). This motif was found in the promoters of 68 upregulated genes and it matched to five known motif in PLACE database; 23PUASNSCYCB1, MYBPLANT (Sablowski *et al.,* 1994), 2SSEEDPROTBANAPA (Stalberg *et al.,* 1996), AACACOREOSGLUB1 (Wu *et al.,* 2000) and PROXBBNNAPA (Ezcurra *et al.,* 1999). Refer to Supplementary Figure 5.2.3.6 for motif logo.

Table 5.2.3.1. DNA motifs discovered by BioProspector in the promoters of co-regulated genes in different tissues of three soybean cultivars

Cv.	Tissue	Motif	Motif Sites	Motif Score	PLACE ID[1]	STAMP *E*-value[2]
Clark	SE[3]	For: TGAATG Rev:CATTCA	128	6.54	ARELIKEGHPGDFR2	3.53E-07
					RYREPEATBNNAPA	1.30E-05
					WUSATAg	6.75E-05
					RYREPEATGMGY2	6.79E-05
					RYREPEATLEGUMINBOX	6.79E-05
	SAM[4]	For: GCGTGC Rev:GCACGC	73	6.343	RHERPATEXPA7	5.05E-06
					ABRE2HVA22	5.81E-06
					GBOX10NT	5.81E-06
					ABASEED1	9.47E-06
					-284MOTIFZMSBE1	8.36E-05
		For: GCAAGC Rev: GCTTGC	87	6.208	ALF1NTPARC	2.94E-05
					ASF1NTPARA	2.98E-05
					PASNTPARA	4.86E-05
					CE3OSOSEM	2.12E-04
		For: GATCGA Rev: TCGATC	82	6.241	NONAMERATH4	2.44E-06
					RNFG1OS	7.90E-06
					ABAREG2	2.20E-04
					AMMORESIVDCRNIA	5.88E-04
					SV40COREENHAN	1.44E-03
		For: CTAACT Rev: AGTTAG	99	6.217	SITE3SORPS1	7.01E-07
					MYB2AT	3.95E-06
					MYB1LEPR	2.38E-05
					MYBATRD22	2.39E-05
					MYB26PS	6.17E-05

Table 5.2.3.1 continued

5-Leaflet Mutant	SE^3	For: GATTCA Rev:TGAATC	125	6.616	EIN3ATERF1	1.24E-06
					GCN4OSGLUB1	8.78E-05
					L1DCPAL1	2.03E-04
					AGMOTIFNTMYB2	2.11E-04
					ARR1AT	3.54E-04
		For: CATACA Rev: TGTATG	122	6.549	S2FSORPL21	2.19E-07
					SORLIP4AT	3.89E-05
					RYREPEATBNNAPA	1.34E-04
					JASE2ATOPR1	2.04E-04
					SORLREP3AT	3.15E-04
		For: AGTTAG Rev: CTAACT	87	6.541	SITE3SORPS1	5.51E-07
					MYBATRD22	1.12E-06
					MYB2AT	3.12E-06
					MYB1LEPR	6.24E-06
					LREBOXIPCCHS1	1.50E-05
	SAM^4	For: ACTCGG Rev: CCGAGT	41	5.303	DRE1COREZMRAB17	2.82E-05
					LTRECOREATCOR15	1.11E-03
					C1MOTIFZMBZ2	1.25E-03
					RSRBNEXTA	1.29E-03
					ABREMOTIFIIIOSRAB16B	1.65E-03
		For: CCGAGT Rev: ACTCGG	41	5.303	PREATPRODH	8.84E-05
					GCN4OSGLUB1	3.67E-04
		For: GGCGCG Rev: CGCGCC	30	5.305	RE1ASPHYA3	6.88E-08
					ABRECE3ZMRAB28	1.17E-07
					PE3ASPHYA3	1.23E-06
					PE2FNTRNR1A	2.29E-05
					OCTAMOTIF2	5.28E-05
		For: AGATAG Rev: CTATCT	48	5.278	EVENINGAT	7.97E-05
					RGATAOS	7.97E-05
					IBOXLSCMCUCUMISIN	3.92E-04
					SE2PVGRP1	5.29E-04
					CAATBOX2	6.14E-04
					CTRMCAMV35S	3.29E-04

Table 5.2.3.1 continued

Glabrous Mutant	SE[3]	For: TCATAC Rev: GTATGA	67	5.694	SORLIP4AT	2.81E-08
					S2FSORPL21	1.98E-05
					LS7ATPR1	1.24E-04
					JASE2ATOPR1	2.04E-04
					RSEPVGRP1	2.86E-04
		For: CTAACT Rev: AGTTAG	72	6.699	SITE3SORPS1	2.57E-07
					MYB2AT	1.49E-06
					MYB1LEPR	6.24E-06
					MYBATRD22	6.24E-06
					MYB26PS	1.75E-05
	SAM[4]	For: CCAAAC Rev: GTTTGG	68	6.081	23BPUASNSCYCB1	8.43E-07
					MYBPLANT	5.59E-06
					2SSEEDPROTBANAPA	6.24E-06
					AACACOREOSGLUB1	6.24E-06
					PROXBBNNAPA	1.75E-05

Motif Score represents statistical significance of motif distribution estimated by a Monte Carlo method. [1]PLACE ID is the identifier of PLACE consensus sequence matching motif. [2]STAMP E-value represents expectation value of the STAMP alignment. [3]SE is Shoot Epidermis and [4]SAM is Shoot Apical Meristem.

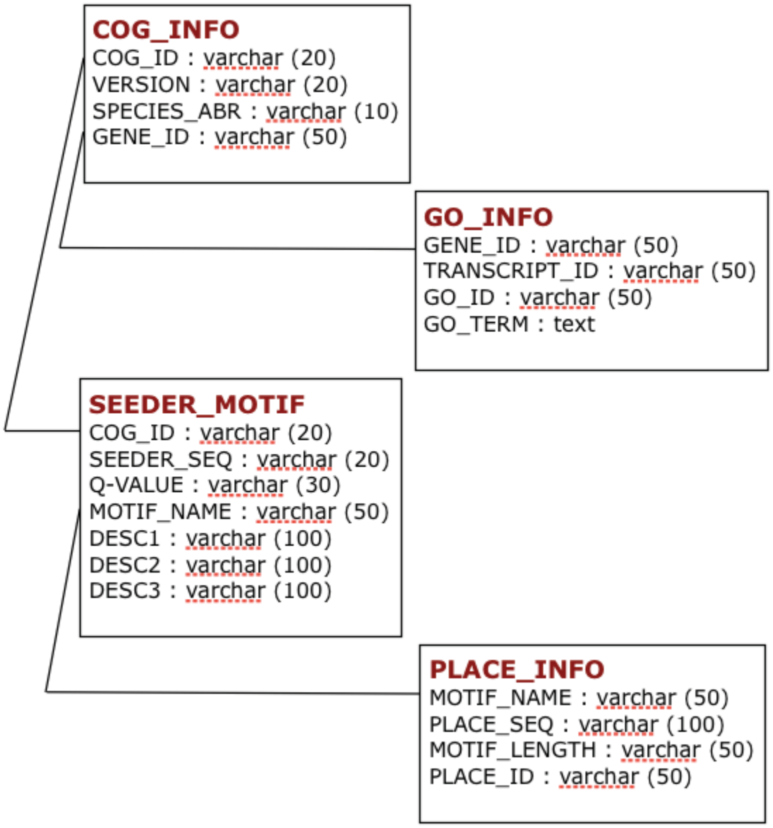
5.2.4. Similar *cis*-regulatory motifs were detected in the same tissue-cultivar set by different motif prediction tools

Motif AMMORESVCDRNIA1, detected by MEME and BioProspector; and motif AMMORESIIUDCRNIA1 (Loppes *et al.,* 2001), detected by MEME, were found in the promoters corresponding to upregulated genes in shoot apical meristem tissue of cv. Clark. Motif SE1PVGRP18 (Keller *et al.,* 1994), detected by MEME; and motif RSEPVGRP1 (Keller *et al.,* 1994), detected by BioProspector; were found in the promoters corresponding to upregulated genes in shoot apical meristem tissue of the 5-Leaflet mutants and shoot epidermis tissue of Glabrous mutants. Motif MYBPZM (Grotewold *et al.,* 1994), detected by MEME; and motif MYBPLANT (Sablowski *et al.,* 1994), detected by BioProspector, were found in the promoters corresponding to upregulated genes in shoot apical meristem tissue of Glabrous mutants.

5.3. Promologous: Database of *cis*-regulatory motifs across 18 plants

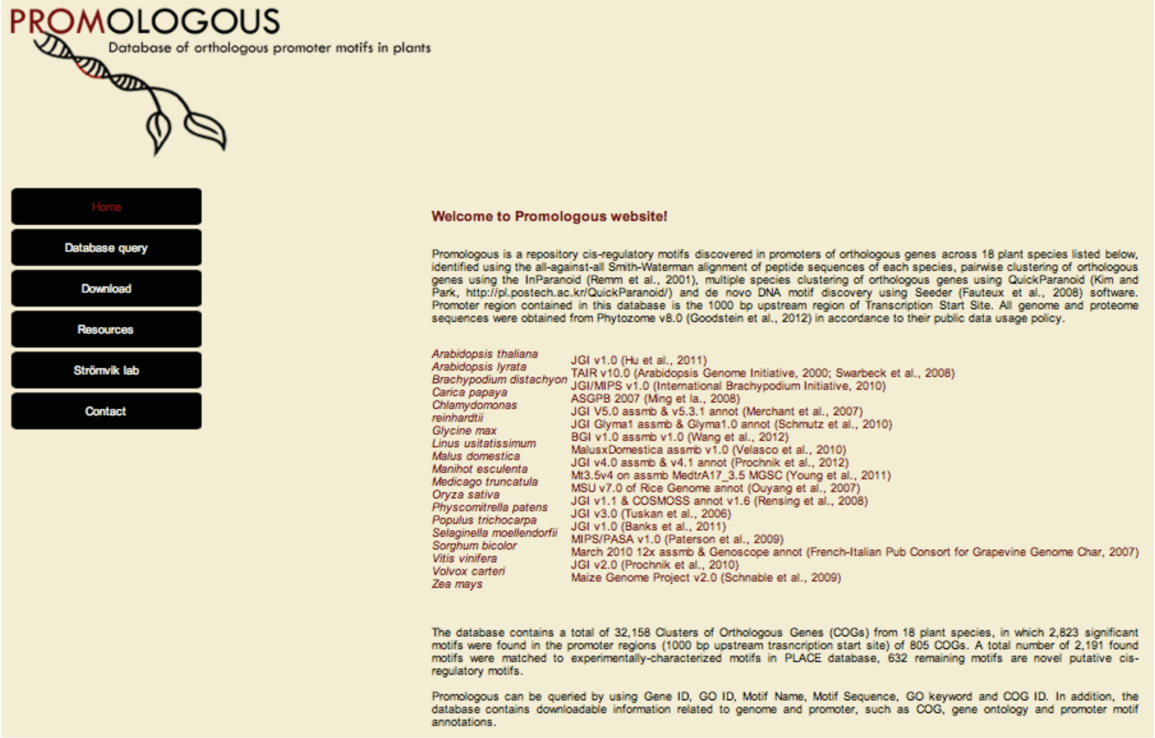
Any data related to genome-promoter that were obtained in this study were made publically available through an online relational database called Promologous (http://stromviklab.agrenv.mcgill.ca/promologous/index.html). Promologous houses a total of 32,158 COGs from 18 plant species, in which 2,823 significant motifs were found in 805 COGs. A total number of 2,191 found motifs were matched to experimentally characterized motifs in PLACE database, whereas 632 remaining motifs are putative *cis*-regulatory motifs. In this database, there are four MySQL tables that are connected to each other by keys assigned to each of them, making genome-promoter exploration more accessible and easier. The *cog_info* table contains COG ID (key), gene ID, (key) species abbreviation and version of the COGs. The *go_info* table contains gene ID (key), GO ID and GO description. The *seeder_motif* table contains COG ID (key), motif sequence, q-value, gene reference of the motif, motif ID (key) and version number. The *place_info* table contains motif ID (key), PLACE motif ID, PLACE motif sequence and gene reference of the motif. Promologous database can be queried by typing gene ID, COG ID, GO ID, GO keyword motif name and motif sequence. In addition, downloadable files include information related to genes that contains gene ID (key), transcript ID (key), description of gene, chromosome name, gene start, gene end and strand; information related to promoter that contains gene ID (key), promoter sequence, strand and version; and information related to transcripts that contains transcript ID (key), pac transcript ID, peptide name, transcript start and transcript end.

Figure 5.3.1. MySQL Relational Database schema



The relational database contains *cog_info*, *go_info*, *seeder_motif* and *place_info* tables. Each table has a component (key) that connects the table to at least one or more tables.

Figure 5.3.2. Promologous web interface



PROMOLOGOUS

Database of orthologous promoter motifs in plants

Welcome to Promologous website!

Promologous is a repository cis-regulatory motifs discovered in promoters of orthologous genes across 18 plant species listed below, identified using the all-against-all Smith-Waterman alignment of peptide sequences of each species, pairwise clustering of orthologous genes using the InParanoid (Remm et al., 2001), multiple species clustering of orthologous genes using QuickParanoid (Kim and Park, http://pl.postech.ac.kr/QuickParanoid/) and de novo DNA motif discovery using Seeder (Fauteux et al., 2008) software. Promoter region contained in this database is the 1000 bp upstream region of Transcription Start Site. All genome and proteome sequences were obtained from Phytozome v8.0 (Goodstein et al., 2012) in accordance to their public data usage policy.

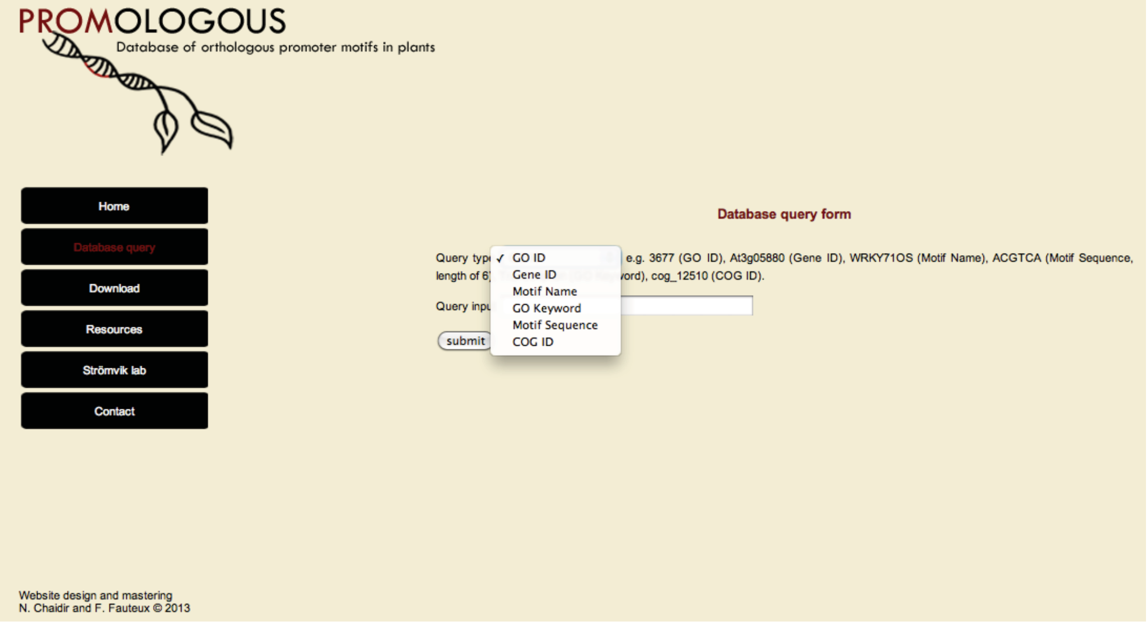
Arabidopsis thaliana	JGI v1.0 (Hu et al., 2011)
Arabidopsis lyrata	TAIR v10.0 (Arabidopsis Genome Initiative, 2000; Swarbeck et al., 2008)
Brachypodium distachyon	JGI/MIPS v1.0 (International Brachypodium Initiative, 2010)
Carica papaya	ASGPB 2007 (Ming et la., 2008)
Chlamydomonas	JGI V5.0 assmb & v5.3.1 annot (Merchant et al., 2007)
reinhardtii	JGI Glyma1 assmb & Glyma1.0 annot (Schmutz et al., 2010)
Glycine max	BGI v1.0 assmb v1.0 (Wang et al., 2012)
Linus usitatissimum	MalusxDomestica assmb v1.0 (Velasco et al., 2010)
Malus domestica	JGI v4.0 assmb & v4.1 annot (Prochnik et al., 2012)
Manihot esculenta	Mt3.5v4 on assmb MedtrA17_3.5 MGSC (Young et al., 2011)
Medicago truncatula	MSU v7.0 of Rice Genome annot (Ouyang et al., 2007)
Oryza sativa	JGI v1.1 & COSMOSS annot v1.6 (Rensing et al., 2008)
Physcomitrella patens	JGI v3.0 (Tuskan et al., 2006)
Populus trichocarpa	JGI v1.0 (Banks et al., 2011)
Selaginella moellendorfii	MIPS/PASA v1.0 (Paterson et al., 2009)
Sorghum bicolor	March 2010 12x assmb & Genoscope annot (French-Italian Pub Consort for Grapevine Genome Char, 2007)
Vitis vinifera	JGI v2.0 (Prochnik et al., 2010)
Volvox carteri	Maize Genome Project v2.0 (Schnable et al., 2009)
Zea mays	

The database contains a total of 32,158 Clusters of Orthologous Genes (COGs) from 18 plant species, in which 2,823 significant motifs were found in the promoter regions (1000 bp upstream trasncription start site) of 805 COGs. A total number of 2,191 found motifs were matched to experimentally-characterized motifs in PLACE database, 632 remaining motifs are novel putative cis-regulatory motifs.

Promologous can be queried by using Gene ID, GO ID, Motif Name, Motif Sequence, GO keyword and COG ID. In addition, the database contains downloadable information related to genome and promoter, such as COG, gene ontology and promoter motif annotations.

Promologous is an online database of *cis*-regulatory motifs discovered in orthologous promoters of 18 plant species.

Figure 5.3.3. Promologous query page



Database query form

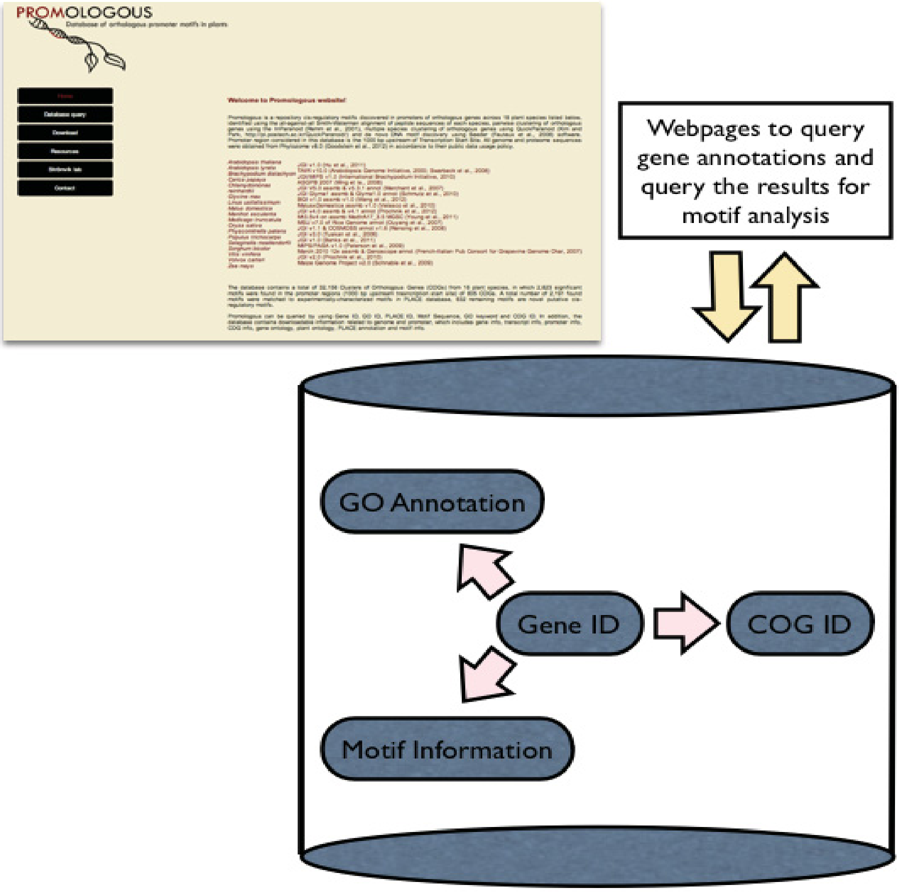
Query typ[...] e.g. 3677 (GO ID), At3g05880 (Gene ID), WRKY71OS (Motif Name), ACGTCA (Motif Sequence, length of 6) [...]ord), cog_12510 (COG ID).

Query inpu

Website design and mastering
N. Chaidir and F. Fauteux © 2013

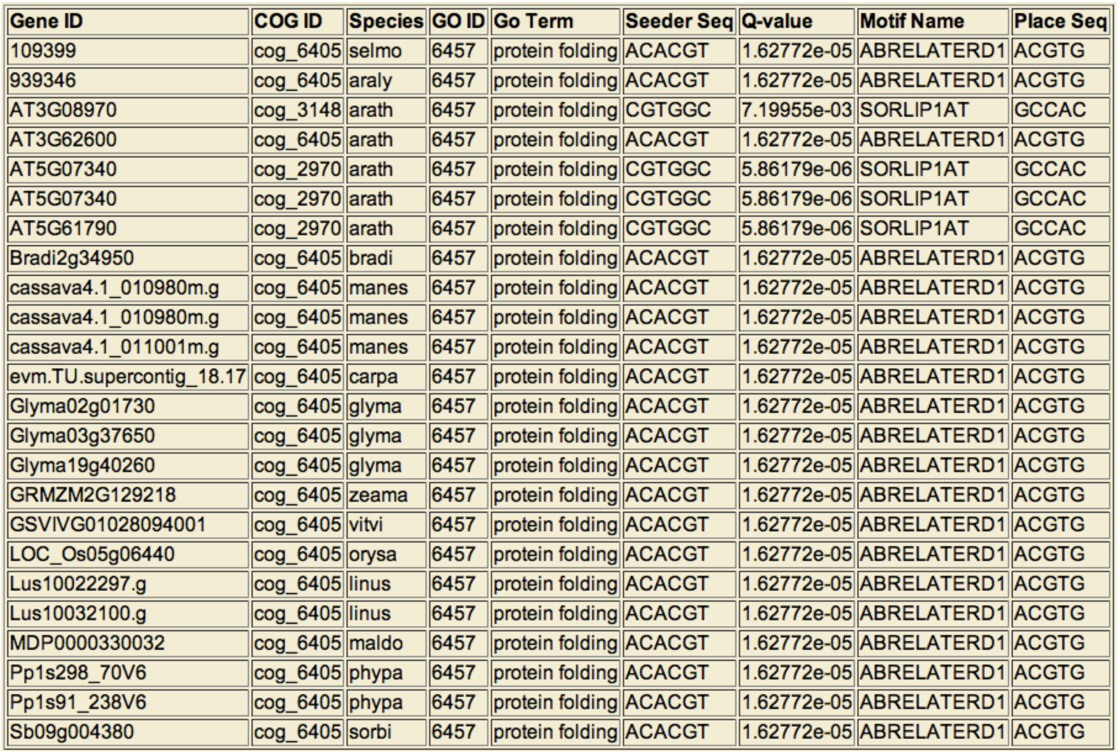
Promologous database can be queried by using GO ID, gene ID, motif name, GO keyword, motif sequence or COG ID.

Figure 5.3.4. Overview of Promologous database structure



Gene ID links COG information with motif information, as well as functional annotation represented by Gene Ontology.

Figure 5.3.5. Example of output in Promologous database



Gene ID	COG ID	Species	GO ID	Go Term	Seeder Seq	Q-value	Motif Name	Place Seq
109399	cog_6405	selmo	6457	protein folding	ACACGT	1.62772e-05	ABRELATERD1	ACGTG
939346	cog_6405	araly	6457	protein folding	ACACGT	1.62772e-05	ABRELATERD1	ACGTG
AT3G08970	cog_3148	arath	6457	protein folding	CGTGGC	7.19955e-03	SORLIP1AT	GCCAC
AT3G62600	cog_6405	arath	6457	protein folding	ACACGT	1.62772e-05	ABRELATERD1	ACGTG
AT5G07340	cog_2970	arath	6457	protein folding	CGTGGC	5.86179e-06	SORLIP1AT	GCCAC
AT5G07340	cog_2970	arath	6457	protein folding	CGTGGC	5.86179e-06	SORLIP1AT	GCCAC
AT5G61790	cog_2970	arath	6457	protein folding	CGTGGC	5.86179e-06	SORLIP1AT	GCCAC
Bradi2g34950	cog_6405	bradi	6457	protein folding	ACACGT	1.62772e-05	ABRELATERD1	ACGTG
cassava4.1_010980m.g	cog_6405	manes	6457	protein folding	ACACGT	1.62772e-05	ABRELATERD1	ACGTG
cassava4.1_010980m.g	cog_6405	manes	6457	protein folding	ACACGT	1.62772e-05	ABRELATERD1	ACGTG
cassava4.1_011001m.g	cog_6405	manes	6457	protein folding	ACACGT	1.62772e-05	ABRELATERD1	ACGTG
evm.TU.supercontig_18.17	cog_6405	carpa	6457	protein folding	ACACGT	1.62772e-05	ABRELATERD1	ACGTG
Glyma02g01730	cog_6405	glyma	6457	protein folding	ACACGT	1.62772e-05	ABRELATERD1	ACGTG
Glyma03g37650	cog_6405	glyma	6457	protein folding	ACACGT	1.62772e-05	ABRELATERD1	ACGTG
Glyma19g40260	cog_6405	glyma	6457	protein folding	ACACGT	1.62772e-05	ABRELATERD1	ACGTG
GRMZM2G129218	cog_6405	zeama	6457	protein folding	ACACGT	1.62772e-05	ABRELATERD1	ACGTG
GSVIVG01028094001	cog_6405	vitvi	6457	protein folding	ACACGT	1.62772e-05	ABRELATERD1	ACGTG
LOC_Os05g06440	cog_6405	orysa	6457	protein folding	ACACGT	1.62772e-05	ABRELATERD1	ACGTG
Lus10022297.g	cog_6405	linus	6457	protein folding	ACACGT	1.62772e-05	ABRELATERD1	ACGTG
Lus10032100.g	cog_6405	linus	6457	protein folding	ACACGT	1.62772e-05	ABRELATERD1	ACGTG
MDP0000330032	cog_6405	maldo	6457	protein folding	ACACGT	1.62772e-05	ABRELATERD1	ACGTG
Pp1s298_70V6	cog_6405	phypa	6457	protein folding	ACACGT	1.62772e-05	ABRELATERD1	ACGTG
Pp1s91_238V6	cog_6405	phypa	6457	protein folding	ACACGT	1.62772e-05	ABRELATERD1	ACGTG
Sb09g004380	cog_6405	sorbi	6457	protein folding	ACACGT	1.62772e-05	ABRELATERD1	ACGTG

Promologous outputs gene ID, COG ID, species abbreviation, GO ID, GO term, Seeder motif sequence, Q-value of motif, motif annotation in PLACE and its corresponding sequence. Database query was made by using GO ID 6457 as an input.

Chapter 6: Discussion

This thesis had three main objectives; to discover gene regulatory motifs in orthologous genes in several plant whole genome sequences, to discover motifs in groups of promoters from co-regulated genes and finally to develop a relational database that will allow researchers to explore this collection of data.

Transcription factor binding sites, or regulatory motifs, are commonly located within 1000bp upstream of transcription start site, and this is the length of promoter sequences used in our study. In the first objective, 609,006 out of 698,963 promoter sequences across 18 plant genomes were included in the data set. The remaining 89,957 promoter sequences were of short length, i.e. less than 1000 bp, and were thus eliminated. A total number of 359,025 promoter sequences were then retrieved from orthologous genes in 32,158 clusters and analyzed for motif prediction. This would make an overall coverage of 51.4% of the promoters in the whole set of 18 plant species.

Regulatory DNA motifs were discovered in the promoters of 805 COGs across 18 plant species. Among 32,158 bioinformatically profiled COGs in 18 plant species, only. 2.5% of the COGs contained conserved motifs in their promoters. This number seems to be reasonable because in such a wide range of species, only main regulatory genes will be conserved, such as housekeeping genes that are required for the maintenance of basic cellular function and are expressed in most cells of an organism under normal conditions. Although only 805 COGs were profiled, a large number, i.e. 2,823 *cis*-regulatory motifs, were found in the promoters of those orthologous gene groups, and of which 2,191 matched known motifs in the PLACE database. Interestingly, 632 other motifs had no match to known motifs and these are considered putative novel motifs. In the first part of this study, it has became obvious that inclusion of more evolutionary distant organism has led to detection of only very conserved motifs, i.e. motifs that have similar function in wider variety of organisms. The *de novo* predicted motif data and related clusters of orthologous gene annotations obtained in this study will be useful to annotate and characterize novel *cis*-regulatory elements or motifs with no close similarity to known *cis*-regulatory elements.

The other objective of this study was to discover DNA motifs in promoters of co-regulated genes in epidermis and shoot apical meristem tissue of soybean cv. Clark, the 5-Leaflet mutant and the Glabrous mutant. Significant motifs were found in the promoters of the upregulated genes of each tissue. Some tissue-specific motifs, experimentally characterized by other researchers were found in the promoters of co-regulated genes in the analysis.

Using BioProspector software, Motif LIDCPAL1 was detected in the promoters corresponding to epidermis tissue of the 5-Leaflet mutant. This promoter motif was proven to drive strong expression in vascular tissues of roots and leaves, but repressing activities in the meristem tissue (Ohl, et al., 1990). Motif NONAMERATH1, a Nonamer motif of *Arabidopsis thaliana* histone H4 promoter was detected in the shoot apical meristem tissue of cv. Clark. This motif, AGATCGACG, was proved to be essential for meristem-specific expression (Chaubet, et al., 1996). Motif SE2PVGRP1 was detected by using BioProspector software in the shoot apical meristem tissue of the 5-Leaflet mutant. This motif is known as the binding site of SE2 (Stem Element 2) that is proven to drive strong expression in stem (Elmayan, et al., 1995). The motifs MYB2AT were detected by BioProspector software in shoot epidermis tissue of the 5-Leaflet mutant and the Glabrous mutant, as well as in the shoot apical meristem tissue of cv. Clark . MYB2AT motif is known to be the binding site of ATMYB2 that shows considerable homology to plant MYB-related proteins, such as GL1, which is required for initiation of differentiation of hair cells or trichomes (Oppenheimer, et al., 1991).

Using MEME software, motif NDEGMSAUR (Li *et al.,* 1994) and ARFAT (Ulmasov *et al.,* 1999) were detected in the promoters corresponding to upregulated genes in shoot apical meristem tissue of Glabrous mutant. These motifs were related to auxin-responsive gene promoter (Goda *et al.,* 2004). Numerous pharmacological and genetic studies have demonstrated that auxin promoter also plays role in trichome formation (Rahman *et al.,* 2002; Masucci *et al.,* 1994; Vernous *et al.,* 2010). Motif SITEIIBOSPCNA (Kosugi *et al.,* 1995) was detected by MEME software in the promoters corresponding to upregulated genes in shoot epidermis of cv. Clark , 5-Leaflet mutants and Glabrous mutants; and shoot apical meristem tissue of Glabrous mutants. Motif SITEIBOSPCNA (Kosugi *et al.,* 1995) was detected by MEME software in the promoters corresponding to upregulated genes in shoot epidermis of cv. Clark and the 5-

Leaflet mutants. Both of these motifs were proven to be required for a meristem tissue-specific expression as a binding site for two nuclear protein; PCF1 and PCF2 (Kosugi and Ohashi, 1994). Motif GAGAGMGSA1 (Sangwan *et al.,* 2002) and GAGA8HVBKN3 (Santi *et al.,* 2003*)* were detected by MEME software in the promoters corresponding to upregulated genes in shoot apical meristem tissue of the 5-Leaflet and Glabrous mutant. GAGAGMGSA1 was proven to be a GAGA binding protein in heme and chlorophyll gene *Gsa1* in soybean (Sangwan *et al.,* 2002). GAGA8HVBKN2 was proven to be a GA octodinucleotide repeat found in intron IV or barley gene *Bkn3* (Santi *et al.,* 2003). Motif ACIPVPAL2 and ACIIPVPAL2 (Hatton *et al.,* 1995) were detected by MEME software in the promoters corresponding to upregulated genes in shoot apical meristem of cv. Clark, 5-Leaflet mutants and Glabrous mutants. These motifs were shown to be related to ACI or ACII element in PAL2 promoter that is required for vascular gene expression (Hatton *et al*., 1995). Motif AMMORESVDCRNIA1 and AMMORESIIUDCRNIA1 (Loppes and Radoux, 2001) were detected by MEME software in the promoters corresponding to upregulated genes in shoot apical meristem of cv. Clark. These motifs were proven to be found in *Chlamydomonas reinhardtii Nia1* gene promoter, which also encodes for cytosolic minor isoform of nitrate reductase in *Arabidopsis thaliana,* involved in the first step of nitrate assimilation as it contributes about 15% of the nitrate reductase activity in shoots (Desikan *et al.,* 2002). These correlations show possibilities that the *cis*-regulatory elements that were found play a role in the transcriptional regulation of these genes.

Some different motifs were discovered by three different motif discover software; Seeder (Fauteux *et al.,* 2008), MEME (Bailey *et al.,* 2009) and BioProspector (Liu *et al.,* 2001). The differences might derive from different algorithm basis that each of them has. Seeder uses enumerative-guaranteed optimality of seed selection and background model based on empirical distribution of substing minimal distance (SMD). MEME uses multiple Expectation-Maximization algorithm and searches for repeated, ungapped sequence motifs that present in DNA sequence; while BioProspector uses Gibbs sampling method, zero to third-order Markov background models and significance of motifs found is determined by motif score distribution estimated by a Monte Carlo method. It is best for biologists to use a couple of complementary bioinformatics tools and consider the top few predicted motifs of each tool (Tompa *et al*., 2005).

Genes that contained same motifs in each tissue-cultivar set found by each motif prediction tools were crosschecked with their GO annotations to see if they share some functional similarities. Results have shown that most of the gene members within the cluster do have similar functions. As an example, in COG 6405, all of the gene members share one similar function, that is protein folding. Motif ACACGT and CGTGGC were also found in the promoter regions of its gene members.

Some genes also show correlation between a motif that was found in the promoter region and GO annotation. An example is gene AT3G05880 in *Arabidopsis thaliana* that was known to be induced by dehydration, low temperatures, salt stress, and abscisic acid related (Dai *et al.,* 2007; Medina *et al*., 2005; Mitsuya *et al.,* 2005). It produces a highly hydrophobic protein that bears two potential transmembrane domains (Medina *et al*., 2001). This gene was also annotated in Gene Ontology that has integral function to membrane (GO:16021), response to cold (GO:9409), response to abscisic acid stimulus (GO:9737), hyperosmotif salinity response (GO:42538) and response to cold (GO:9409). Motif ACGGTG was detected by Seeder in the promoter region of this gene and this motif was known for early responsive to dehydration in Arabidopsis (Simpson *et al.,* 2003).

Identifying promoter features that define transcriptional behavior of co-regulated genes is a difficult task for a number of reasons. First, it is not clear how to distinguish a six-symbol DNA alphabet (i.e. motif) that truly affects gene expression from a random sequence (Beer and Tavazoie, 2004). There is a probability of 4,096 random six-letter sequence occurs in random DNA. Second, the relative locations of the transcription factor binding sites (motifs) differ from one promoter to another. Third, similar transcription patterns can be derived from a combination of more than one underlying features, therefore making it more challenging to differentiate the causes of analogous regulatory effects (Pritsker et al., 2004).

Wide varieties of transcription factors compete for binding to *cis*-regulatory elements. *Cis*-regulatory elements and putative transcription factors that are likely to bind can be predicted by using *in silico* prediction linked with available experimental evidence. However, due to the limited number of experimental data, analysis and interpretation of putative regulatory motifs in gene promoter are difficult to decipher. In addition, the affinities for a specific binding site and the nuclear concentrations of active transcription factors may define the actual binding outcome.

There are also many other regulatory elements in a genome and the resulting transcription is the output of many regulatory signals. Other regulatory elements include; DNA methylation where addition of a methyl group at position 5 of cytosines could make a major epigenetic modification and altering transcriptional machinery (Foerster and Scheid, 2010); regulation of chromatin that could influence nuclear processes from replication, recombination and repair to transcriptional machinery (Wagner, 2003); distal enhancers that are found to be important key factors of nuclear organization, contributing to a general model for enhancer function that involves enhancer-promoter contact (Bulger and Goudine, 2011); and insulators that prevent incorrect gene regulation by restricting the action of enhancers and silencers (Raab and Kamakaka, 2010). Furthermore, post-transcriptional regulation by small-RNAs adds a new level of complexity to this, as they are riboregulators that have important roles by repressing gene expression of DNA to guide sequence elimination or chromatin remodeling, or on RNA to guide cleavage and translation repression (Vaucheret *et al.,* 2006).

Narrowing down the possibilities of which regulatory elements come into play is an extremely challenging task and was not within the scope of the work presented in this thesis. The presence or not of specific motifs, remain the basic pre-requisite for whether a transcription factor can bind or not, to a given promoter. *Cis*-regulatory elements are one of the key factors responsible for gene expression regulation. *Cis*-regulatory motifs define specific recruitments of the protein complexes (i.e. Transcription Factors) that will bind. To fully understand how these elements operate, it is also necessary to understand the protein complexes that interact with these *cis*-acting elements. These proteins are encoded by other genes and due to their ability to diffuse; these protein complexes can also act in *trans*, which is why they have effects on any copy of a gene that has an appropriate regulatory sequence within it.

To generate large genome-scale datasets across many species, extensive computational resources are required but not always readily available, not to mention the amount of time it consumes. *In silico* motif predictions can serve as a good starting point for establishing the gene function and expression, which ultimately needs to be validated by experimental analysis. The Promologous database contains pre-computed whole-genome comparative data of promoter sequences across 18 plant species, as well as other genome information associated with them. The information in the database can be potentially used to facilitate genome-wide experimentation for researchers that have interests in engineering and manipulating gene expression for plant biotechnology purposes.

Chapter 7: Conclusion

The results of the first approach in this study shows that many promoters of orthologous plant genes contain similar *cis*-regulatory motifs. In addition, inclusion of more evolutionary distant organism leads to detection of very conserved motifs, i.e. motifs that have similar function in wider variety of organisms.

Furthermore, using the second approach, we detected conserved motifs in promoter region of co-regulated genes in epidermis and shoot apical meristem tissue of soybean cv. Clark, 5-Leaflet mutant and Glabrous mutant. Data obtained from experimentally characterized plant *cis*-acting regulatory DNA elements in PLACE database support the biological relevance of our findings.

A relational database of *cis*-regulatory motifs discovered in the promoters of orthologous genes across 18 plant species called Promologous is available online for researchers who are interested in further exploring genome-promoter information.

Chapter 8: Future Research Directions

The following directions could be taken to go further with the research reported in this thesis:

1. The analysis of conserved motifs in orthologous genes of 18 plants could be improved by a more thorough analysis of paralogs prior to clustering step and by going deeper in the analysis of gene function in relation with promoter motifs composition and experimentally verified *cis*-regulatory elements.
2. As more plant species are sequenced, if would be interesting to create different divisions in finding motifs based on orthologous genes, as a more specific stringency could be applied accordingly to the genetic composition of the species members in each division, for instance, motifs in monocotyledonous orthologous promoters, dicotyledonous orthologous promoters, etc.
3. Motifs found in promoters of co-regulated genes in soybean could be tested experimentally to confirm its function in regulating transcription and examine its relative strength in shoot epidermis *vs* shoot apical meristem.

Chapter 9: Contribution to Knowledge

1. A total number of 32,158 groups of orthologous genes were identified in 18 plant species.
2. A total number of 2,823 conserved motifs were identified in promoters of orthologous genes from 18 plant species and 632 of them were novel putative *cis*-regulatory motifs.
3. Conserved motifs were identified in promoter region of co-regulated genes from shoot epidermis and shoot apical meristem tissues of three soybean cultivars; the cv. Clark, 5-Leaflet mutant and Glabrous mutant.
4. A relational database called Promologous was developed. This database houses pre-computed and post-processed whole-genome comparative analysis of promoter regions. Promologous also contains motif sequences, annotations, clusters of orthologous genes and other useful information associated with them, for 18 plant genomes.

References

Abe, H., Yamaguchi-Shinozaki, K., Urao, T., Iwasaki, T., Hosokawa, D., Shinozaki, K. Role of Arabidopsis MYC and MYB homologs in drought- and abscisic acid-regulated gene expression. *Plant Cell*. 9:1859-1868 (1997)

Acevedo-Hernandez, G.J., Leon, P., Herrera-Estrella, L.R. Sugar and ABA responsiveness of a minimal RBCS light-responsive unit is mediated by direct binding of ABI4. *Plant J*. 43:506-519 (2005)

Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M, Polymeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B., Moreno, R.F., et al. Complementary DNA sequencing: Expressed sequence tags and human genome project. *Science* 252: 1651–1656 (1991)

Alexeyenko, A., Tamas, I., Liu, G., Erik, L., Sonnhammer, L. Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics* 22: e9-e15 (2006)

Allen RD, Bernier F, Lessard PA, Beachy RN. Nuclear factors interact with a soybean beta-conglycinin enhancer. *Plant Cell*. 1:623-631 (1989)

Alwine, J.C., Kemp, D.J., & Stark, G.R. Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl- paper and hybridization with DNA probes. *Proceeding National Academy of Sciences: Biochemistry*. USA 74, 5350–5354 (1977)

Anders, S., Huber, W. Differential expression analysis for sequence count data. *Genome Biology*: **11**, R106 (2010)

Arnone, M. I. and Davidson, E. H. The hardwiring of development: organization and function of genomic regulatory systems. *Development,* 124(10):1851–1864 (1997)

Azad, A.K.M., Shahid, S., Noman, N., Lee, H. Prediction of plant promoters based on hexamers and random triplet pair analysis. *Algorithms for Molecular Biology*, 6:19 (2011)

Bailey, T.L. and Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol*, 2, 28-36

(1994)

Bailey, T.L., Williams, N., Misleh, C. Li, W.W. MEME: discovering and analyzing DNA and protein sequence motifs. (2006) *Nucleic Acids Research*, Vol. 34, Web Server issue W369–W373

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

Bao, J.Y., Lee, S., Chen, C., Zhang, X.Q., Zhang, Y., Liu, S.Q., Clark, T., Wang, J., Cao, M.L., Yang, H.M., Wang, S.M., Yu, J. Serial Analysis of Gene Expression Study of a Hybrid Rice Strain (LYP9) and Its Parental Cultivars. *Plant Physiology*, Vol. 138, pp. 1216–1231 (2005)

Bate N, Twell D. Functional architecture of a late pollen promoter: pollen-specific transcription is developmentally regulated by multiple stage-specific and co-dependent activator elements. *Plant Mol Biol*. 37:859-869 (1998)

Baum, L.E., Petrie, T., Soules, G., Weiss, N. "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains", *Ann. Math. Statist.,* vol. 41, no. 1, pp. 164–171, (1970)

Berezikov, E. et al. Diversity of microRNAs in human and chimpanzee brain. *Nature Genetics*, 38, 1375–1377 (2006)

Block A, Dangl JL, Hahlbrock K, Schulze-Lefert P. Functional borders, genetic fine structure, and distance requirements of cis elements mediating light responsiveness of the parsley chalcone synthase promoter. *Proc Natl Acad Sci USA*. 87:5387-5391(1990)

Bodeau JP, Walbot V. Structure and regulation of the maize Bronze2 promoter. *Plant Mol Biol*. 32:599-609 (1996)

Bohnert, R., Behr, J., Rätsch, G. Transcript quantification with RNA-Seq data. *BMC Bioinformatics,* 10(Suppl 13):P5 (2009)

Brown, P.O., Botstein, D. Exploring the new world of the genome with DNA microarrays. *Nature Genetics* 21, 33-37 (1999)

Bruce WB, Deng XW, Quail PH. A negatively acting DNA sequence element mediates phytochrome-directed repression of phyA gene transcription. *EMBO J*. 10:3015-

3024 (1991)

Bruce WB, Quail PH. cis-Acting elements involved in photoregulation of an oat phytochrome promoter in rice. *Plant Cell*. 2:1081-1089 (1990)

Bulger, M., Groudine, M. Functional and mechanistic diversity of distal transcription enhancers. *Cell,* Volume 144, Issue 3, 327-339 (2011)

Busk PK, Jensen AB, Pages M. Regulatory elements in vivo in the promoter of the abscisic acid responsive gene rab17 from maize. *Plant J*. 11: 1285-1295 (1997)

Busk PK, Pages M. Protein binding to the abscisic acid-responsive element is independent of VIVIPAROUS1 in vivo. *Plant Cell*. 9:2261-2270 (1997)

Campbell, T.N., Choy, F.Y. RNA interference:past present and future. *Current Issues Molecular Biology*, 7(1):1-6 (2005)

Chakravarthy S, Tuori RP, DAscenzo MD, Fobert PR, Despres C, Martin GB. The tomato transcription factor Pti4 regulates defence-related gene expression via GCC box and non-GCC box cis elements. *Plant Cell*. 15: 3033-3050 (2003)

Chaubet N, Flenet M, Clement B, Brignon P, Gigot C. Identification of cis-elements regulating the expression of an Arabidopsis histone H4 gene. *Plant J*. 10:425-435 (1996)

Chaubet N, Philipps G, Chaboute M-E, Ehling M, Gigot C. Nucleotide sequences of two corn histone H3 genes. Genomic organization of the corn histone H3 and H4 genes. *Plant Mol Biol*. 6:253-263 (1986)

Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, and de Hoon MJ. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*. Jun 1; 25(11) 1422-3 (2009)

Cocquet, J., Chong, A., Zhang, G.,Veitia, R. A. Reverse transcriptase template switching and false alternative transcripts. *Genomics* 88, 127–131 (2006)

Collins, F. Has the revolution arrived? *Nature* 464, 674–675 (2010)

Croissant-Sych Y, Okita TW. Identification of positive and negative regulatory cis-elements of the rice glutelin Gt3 promoter. *Plant Science*. 116:27-35 (1996)

Czarnecka, E., Key, J. L., Gurley, W. B. Regulatory domains of the Gmhsp17,5-E heat shock promoter of soybean. *Mol Cell Bio*. 9(8): 3457-3463 (1989)

Dai, Xiaoyan, Xu, Yunyuan, Ma, Qibin, Xu, Wenying, Wang, Tai, Xue, Yongbiao, Chong, Kang. Overexpression of a R1R2R3 MYB Gene, OsMYB3R-2, Increases Tolerance to Freezing, Drought, and Salt Stress in Transgenic Arabidopsis. *Plant Physiology;* 143(4):1739-51 (2007)

Davidson, E. H. Genomic Regulatory Systems: Development and Evolution. *Academic*, New York (2001)

De Boer, G.J., Testerink, C., Pielage, G., Nijkamp, H.J., Stuitje, A.R. Sequences surrounding the transcription initiation site of the Arabidopsis enoyl-acyl carrier protein reductase gene control seed expression in transgenic tobacco. *Plant Mol Biol*, 39(6):1197-1207 (1999)

De Bruin, D., Zaman, Z., Liberatore, R.A. & Ptashne, M. Telomere looping permits gene activation by a downstream UAS in yeast. *Nature* 409, 109–113 (2001)

Desikan, R., Griffiths, R., Hancock, J., Neill, S. ABA inhibits stomatal opening in Nia1 Nia2 as in the wildtype, revealing hat not all stomatal responses to ABA are reduced. *PNAS USA*. V99(25):16314-16318 (2002)

Despres C, DeLong C, Glaze S, Liu E, Fobert PR. The Arabidopsis NPR1/NIM1 protein enhances the DNA binding activity of a subgroup of the TGA family of bZIP transcription factors. *Plant Cell*. 12: 279-290 (2000)

Ducrest, A.L., Amacker, M., Lingner, J., Nabholz, M. Detection of promoter activity by flow cytometric analysis of GFP reporter expression. (2002) *Nucleic Acids Research*, Vol. 30 No. 14 e65

Duggan, D.J., Bittner, M., Chen, Y., Meltzer, P., Trent, J.M. Expression profiling using cDNA microarrays. *Nature Genetics* 21, 10-14 (1999)

Dutt, M., Ananthakrishnan, G., Jaromin, M.K., Branlsky, R.H., Grosses, J.W. Evaluation of four phloem-specific promoters in vegetative tissues of transgenic citrus plants. *Tree Physiol*. Jan;32(1):83-93 (2012)

Ekman, D.R., Lorenz, W.W., Przybyla, A.E., Wolfe, N.L., Dean, J.F. SAGE analysis of transcriptome responses in Arabidopsis roots exposed to 2,4,6-trinitrotoluene. *Plant Physiology,* 133: 1397–1406 (2003)

Elliott KA , Shirsat AH. Promoter regions of the extA extensin gene from Brassica napus control activation in response to wounding and tensile stress. *Plant Mol Biol*.

37:675-687 (1998)

Elmayan T, Tepfer M. Evaluation in tobacco of the organ specificity and strength of the rol D promoter, domain A of the 35S promoter and the 35S^2 promoter. *Transgenic Res*. 4:388-396 (1995)

Elomaa P, Uimari A, Mehto M, Albert VA, Latinen RAE, Teeri TH. Activation of anthocyanin biosynthesis in Gerbera hybrida (Asteraceae) suggests conserved protein-protein and protein-promoter interactions between the anciently diverged monocots and eudicots. *Plant Physiol*. 133: 1831-1842 (2003)

Ericson ML, Muren E, Gustavsson H-O, Josefsson L-G, Rask L. Analysis of the promoter region of napin genes from Brassica napus demonstrates binding of nuclear protein in vitro to a conserved sequence motif. *Eur J Biochem*. 197:741-746 (1991)

Ezcurra I, Ellerstrom M, Wycliffe P, Stalberg K, Rask L. Interaction between composite elements in the napA promoter: both the B-box ABA-responsive complex and the RY/G complex are necessary for seed-specific expression. *Plant Mol Biol,* 40:699-709 (1999)

Fauteux, F., Blanchette, M., Stromvik, M.V. Seeder: discriminative seeding DNA motif discovery. *Bioinformatics*. 24 (20): 2303-2307 (2008)

Felsenfeld, G., Boyes, J., Chung, J., Clark, D., Studitsky, V. Chromatin structure and gene expression. *Proceeding of the National Academics of Sciences of the United States of America: V*ol. 93 no. 18 9384-9388 (1996)

Feuillet, C., Leach, J.E., Rogers, J., Schnable, P.S., Eversole, K. Crop genome sequencing: lessons and rationales. *Trends in Plant Science*, Volume 16, Issue 2, 77-88 (2010)

Fizames, C., Munos, S., Cazettes, C., Nacry, P., Boucherez, J., Gaymard, F., Piquemal, D., Delorme, V., Commes, T., Doumas, P., et al The Arabidopsis root transcriptome by serial analysis of gene expression. Gene identification using the genome sequence. *Plant Physiology,* 134: 67–80 (2004)

Fluhr R, Kuhlemeier C, Nagy F, Chua N-H. Organ-specific and light-induced expression of plant genes. *Science*. 232:1106-1112 (1986)

Foerster, A.M., Scheid, O.M. Analysis of DNA methylation in plants by bisulfite sequencing. *Cell,* Volume 144, Issue 3, 327-339 (2010)

Frazer KA, Elnitski L, Church DM, Dubchak I, Hardison RC. Cross-species sequence comparisons: a review of methods and available resources. Genome Res. 2003;13:1–12

Fujiwara T, Beachy RN. Tissue-specific and temporal regulation of a beta-conglycinin gene: roles of the RY repeat and other cis-acting elements. *Plant Mol Biol* 24:261-272 (1994)

Galbraith, D.W. DNA Microarray Analyses in Higher Plants. *OMICS: A Journal of Integrative Biology* 10(4): 455-473 (2006)

Gatz, C., Quail, P.H. Tn10-encoded tet repressor can regulate an operator-containing plant promoter. *Proc Natl Acad Sci USA*. March; 85(5):1394-1397 (1988)

Geffers R, Cerff R, Hehl R. Anaerobiosis-specific interaction of tobacco nuclear factors with cis-regulatory sequences in the maize GapC4 promoter. *Plant Mol Biol*. 43: 11-21 (2000)

Gibbings, J.G., Cook, B.P., Dufault, M.R., Madden, S.L., Khuri, S., Turnbull, C.J., Dunwell, J.M. Global transcript analysis of rice leaf and seed using SAGE technology. *Plant Biotechnology Journal,* 1: 271–285 (2003)

Goodstein, D.M., Shu, S., Howson, R., Neupane, R., Hayes, R.D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., Rokhsar, D.S. Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res*. 2012; 40(d1):D1178-1186

Gowda, M., Jantasuriyarat, C., Dean, R.A., Wang, G.L. Robust-LongSAGE (RL-SAGE): a substantially improved LongSAGE method for gene discovery and transcriptome analysis. *Plant Physiology,* 134: 890–897 (2004)

Gowik U, Burscheidt J, Akyildiz M, Schlue U, Koczor M, Streubel M, Westhoff P. cis-Regulatory elements for mesophyll-specific gene expression in the C4 plant Flaveria trinervia, the promoter of the C4 phosphoenolpyruvate carboxylase gene. *Plant Cell*. 16:1077-1090(2004)

Green PJ, Kay SA, Chua N-H. Sequence-specific interactions of a pea nuclear factor with light-responsive elements upstream of the rbcS-3A gene. *EMBO J*. 6:2543-2549 (1987)

Griffiths, A., Miller, J.H., Suzuki, D.T., et al. *An Introduction to Genetic Analysis*. 7th edition (2000)

Grotewold E, Drummond BJ, Bowen B, Peterson T. The myb-homologous P gene controls phlobaphene pigmentation in maize floral organs by directly activating a flavonoid biosynthetic gene subset. *Cell*. 76:543-553 (1994)

Gubler, U. Second-strand cDNA synthesis: classical method. *Methods Enzymology*. 152, 325–329 (1987).

GuhaThakurta, D. Computational identification of transcriptional regulatory elements in DNA sequence. *Nucleic Acids Res,* 34(12), 3585-3598 (2006)

Gubler, U. Second-strand cDNA synthesis: mRNA fragments as primers. *Methods Enzymology*. 152, 330–335 (1987).

GuhaThakurta, D. Computational identification of transcriptional regulatory elements in DNA sequence. *Nucleic Acids Res,* 34(12), 3585-3598 (2006)

Guillet, C., Aboul-Soud, M.A., Le Menn, A., Viron, N., Pribat, A., Germain, V., Just, D., Baldet, P., Rousselle, P., Lemaire-Chamley, M., Rothan, C. Regulation of the fruit-specific PEP carboxylase SIPPC2 promoter at early stages of tomato fruit development. *PLoS One*. Vol.7(5):e36795 (2012)

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

Hatton D, Sablowski R, Yung MH, Smith C, Schuch W, Bevan M. Tow classes of cis sequences contribute to tissue-specific expression of a PAL2 promoter in transgenic tobacco. *Plant J*. 7:859-876 (1995)

He Y, Gan S. Identical promoter elements are involved in regulation of the OPR1 gene by senescence and jasmonic acid in Arabidopsis. *Plant Mol Biol*, 47: 595-605 (2001)

Heid, C.A., Stevens, J., Livak, K.J., Williams, P.M. Quantitative Real Time PCR. *Genome Res* 6:986-994 (1996)

Hiwasa-Tanase, K., Kuroda, H., Hirai, T., Aoki, K., Takane, K., Ezura, H. Novel promoters that induce specific transgene expression during the green to ripening stages of tomato fruit development. *Plant Cell Rep*. Aug;31(8):1415-1424 (2012)

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

Huda, K.M., Banu, M.S., Pathi, K.M., Tuteja, N. Reproductive organ and vascular specific promoter of the rice plasma membrane Ca2+ATPase mediates environmental stresses responses in plants. *PLoS One*. Vol.8(3):e57803 (2013)

Hudson ME, Quail PH. Identification of promoter motifs involved in the network of phytochrome A-regulated gene expression by combined analysis of genomic sequence and microarray data. *Plant Physiol*. 133: 1605-1616 (2003)

Hughes, T.R., Mao, M., Jones, A.R., Burchard, J., Marton, M.J., Shannon, K.W., Lefkowitz, S.M., Ziman, M., Schelter, J.M., Meyer, M.R., Kobayashi, S., Davis, C., Dai, H., He, Y.D., Stephaniants, S.B., Cavet, G., Walker, W.L., West, A., Coffey, E., Shoemaker, D.D., Stoughton, R., Blanchard, A.P., Friend, S.H., Linsley, P.S. Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *National Biotechnology*, 19:342-7 (2001)

Illumina inc. Digital Gene Expression: Small RNA Discovery and Analysis. *Illumina Data Sheet: Sequencing*. Pub. No. 770-2007-005 (2008)

Ishige F, Takaichi M, Foster R, Chua NH, Oeda K. A G-box motif (GCCACGTGCC) tetramer confers high-level constitutive expression in dicot and monocot plants. *Plant J*. 18:443-448 (1999)

Ishiguro S, Nakamura K. The nuclear factor SP8BF binds to the 5'-upstream regions of three different genes coding for major proteins of sweet potato tuberous roots. *Plant Mol Biol*. 18:97-108 (1992)

Ishii, M., Hashimoto, S., Tsutsumi, S., Wada, Y., Matsushima, K., Kodama, T., Aburatani, H. Direct comparison of GeneChip and SAGE on the quantitative accuracy in transcript profiling analysis. *Genomics*, 68: 136 –143 (2000)

Jeong, J.S., Kim, Y.S., Baek, K.H., Jung, H., Ha, S-H., Choi, Y.D., Kim, M., Reuzeau, C., Kim, J-K. Root specific expression of *OsNAC10* improves drought tolerance and grain yield in rice under field drought conditions. *Plant Physiol*. May;Vol. 153(1):185-197 (2010)

Kamiya N, Nagasaki H, Morikami A, Sato Y, Matsuoka M. Isolation and characterization of a rice WUSCHEL-tyope homoebox gene that is specifically expressed in the central cells of a quiescent center in the root apical meristem. *Plant J*. 35: 429-441 (2003)

Kane, M.D., Jatkoe, T.A., Stumpf, C.R., Lu, J., Thomas, J.D., Madore, S.J. Assessment of the sensitivity and specificity of oligonucleotide (50mer) microarrays. *Nucleic Acids Res*. 28:4552-7 (2000)

Karthikeyan, A.S., Ballachanda, D.N., Raghothama, K.G. Promoter deletion analysis elucidates the role of cis elements and 5' UTR intron in spatiotemporal regulation of AtPht1;4 expression in Arabidopsis. *Physiol Plant*. May;136(1):10-18 (2009)

Keith, J.M. Bioinformatics, Volume I: Data, Sequence Analysis, and Evolution, vol. 452 (2008)

Keller B, Heierli D. Vascular expression of the grp1.8 promoter is controlled by three specific regulatory elements and one unspecific activating sequence. *Plant Mol Biol*. 26: 747-756 (1994)

Kellis, M., Patterson, N., Birren, B., et al. (2004) Methods in comparative genomics: genome correspondence, gene identification and regulatory motif discovery. *J Comput Biol*. 11, 319–355.

Kim DW, Lee SH, Choi SB, Won SK, Heo YK, Cho M, Park YI, Cho HT. Functional Conservation of a Root Hair Cell-Specific cis-Element in Angiosperms with Different Root Hair Distribution Patterns. *Plant Cell*. 18:2958-2970 (2006)

Kim, H.L. Comparison of oligonucleotide-microarray and serial analysis of gene expression (SAGE) in transcript profiling analysis of megakaryocytes derived from CD34+ cells. *Experimental and Molecular Medicine,* Vol. 35, No. 5, 460-466 (2003)

Kim K-N, Guiltinan MJ. Identification of cis-acting elements important for expression of the starch-branching enzyme I gene in Maize endosperm. *Plant Physiol*. 121: 225-236 (1999)

Kosugi S, Suzuka I, Ohashi Y. Two of three promoter elements identified in a rice gene for proliferating cell nuclear antigen are essential for meristematic tissue-specific expression. *Plant J*. 7:877-886 (1995)

Kusaba M, Takahashi Y, Nagata T. A multiple-stimuli-responsive as-1-related element of parA gene confers responsiveness to cadmium but not to copper. *Plant Physiol*. 111: 1161-1167 (1996)

Lagrange T, Gauvin S, Yeo HJ, Mache R. S2F, a leaf-specific trans-acting factor, binds to a novel cis-acting element and differentially activates the RPL21 gene. *Plant Cell*. 9:1469-1479 (1997)

Lam E, Chua NH. ASF-2: A factor that binds to the cauliflower mosaic virus 35S promoter and a conserved GATA motif in cab promoters. *Plant Cell*. 1:1147-1156 (1989)

Lau, N.C., Lim, L.P., Weinstein, E.G., Bartel, D.P. An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science,* 294, 858–862 (2001)

Lawrence, C.E., Reilly, A.A. An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins,* 7 (1), 41-51 (1990)

Lawrence, C.E., Altschul, S.F., Boguski, M.S., Liu, J.S., Neuwald, A.F., Wooton, J.C. "Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignment". *Science*; 262(5131), 208-214 (1993)

Lee, J.Y., Lee, D.H. Use of serial analysis of gene expression technology to reveal changes in gene expression in Arabidopsis pollen undergoing cold stress. *Plant Physiology,* 132: 517–529 (2003)

Lelievre J-M, Oliveira LO, Nielsen NC. 5'-CATGCAT-3' elements modulate the expression of glycinin genes. *Plant Physiol*. 98:387-391 (1992)

Lescot, M., Déhais, P., Thijs, G., Marchal, K., Moreau, Y., Van de Peer, Y., Rouzé, P., and Rombauts, S. PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. *Nucleic Acids Res*. 30, 325–327 (2002)

Li, Y., Liu, S., Yu, Z., Liu, Y., Wu, P. Isolation and characterization of two novel root-specific promoters in rice (*Oryza sativa* L.). *Plant Sci*. June; Vol.207:37-44 (2013)

Li Y, Liu ZB, Shi X, Hagen G, Guilfoyle TJ. An auxin-inducible element in soybean SAUR promoters. *Plant Physiol*. 106:37-43 (1994)

Lincker F, Philipps G, Chaboute ME. UV-C response of the ribonucleotide reductase large subunit involves both E2F-mediated gene transcriptional regulation and protein subcellular relocalization in tobacco cells. *Nucleic Acids Res.* 32: 1430-1438 (2004)

Linsen, S.E.V., Wit, E., Janssens, G., Heater, S., Chapman, L., Parkin, R.K., Fritz, B., Wyman, S.K., Bruijn, E., Voest, E.E., Kuersten, S., Tewari, M., Cuppen, E. Limitations and possibilities of small RNa digital gene expression profiling. *Nature Methods*, Vol.6 no.7, 474-476 (2009)

Liu, W., Mazarei, M., Rudis, M.R., Fether, M.H., Stewart, C.N. Rapid *in vivo* analysis of synthetic promoters of plant pathogen phytosensing. *BMC Biotechnology,* 11:108 (2011)

Liu, X., Brutlag, D.L., Liu, J.S. Bioprospector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac Symp Biocomput,* 127-138 (2001)

Lockhart DJ, Winzeler EA. Genomics, gene expression and DNA arrays. *Nature,* 405, 827-36 (2000)

Lodish, H., Berk, A., Zipursky, S., Matsudaira, P., Baltimore, D., Darnell, J. Molecular Cell Biology, Chapter 10:Regulation of Transcription Initiation. 4. Edition (2000)

Logemann E, Parniske M, Hahlbrock K. Modes of expression and common structural features of the complete phenylalanine ammonia-lyase gene family in parsley. *Proc Natl Acad Sci USA*. 92:5905-5909 (1995)

Loots, G.G. e*t al*. Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science,* 288. 136-140 (2000)

Loppes R, Radoux M. Identification of short promoter regions involved in the transcriptional expression of the nitrate reductase gene in Chlamydomonas reinhardtii. *Plant Mol Biol*. 45: 215-227 (2001)

Lyons,T.J., Gasch,A.P., Alex Gaither,L., Botstein,D., Brown,P.O., Eide,D.J. Genome-wide characterization of the Zap1p zinc-responsive regulon in yeast. (2000). *Proc. Natl Acad. Sci. USA,* 97, 7957–7962

Macknight, R., Duroux, M., Laurie, R., Dijkwel, P., Simpson, G., Dean, C. Functional significance of the alternative transcript processing of the Arabidopsis floral promoter PCA. *Plant Cell*. Apr14(1):877-888 (2002)

Mader, R. M. et al. Reverse transcriptase template switching during reverse transcriptase-polymerase chain reaction: artificial generation of deletions in ribonucleotide reductase mRNA. *J. Lab. Clin. Med.* 137, 422–428 (2001).

Marsit, C.J., Eddy, K., Kelsey, K.T. MicroRNA responses to cellular stress. *Cancer Research,* 66(22):10843-10848 (2006)

Masucci JD, Schiefelbein JW. The rhd6 mutation of Arabidopsis thaliana alters root- hair initiation through an auxin- and ethylene-associated process. *Plant Physiol.* 106:1335– 46 (1994)

Matsumura, H., Nirasawa, S., Terauchi, R. Technical advance: transcript profiling in rice (Oryza sativa L.) seedlings using serial analysis of gene expression (SAGE). *Plant Journal,* 20: 719–726 (1999)

Matsumura, H., Reich, S., Ito, A., Saitoh, H., Kamoun, S., Winter, P., Kahl, G., Reuter, M., Kruger, D.H., Terauchi, R. Gene expression analysis of plant host-pathogen interactions by SuperSAGE. *Proceedings of the National Academy of Sciences USA,* 100: 15718–15723 (2003)

Matsumura, H., Ito, A., Saitoh, H., Winter, P., Kahl, G., Reuter, M., Krüger, D.H., Terauchi, R. SuperSAGE. *Cell Microbiology* 7 (1): 11–8 (2005)

Medina, J., Ballesteros, M.L., Salinas, J. Characterization of a new Arabidopsis gene family encoding highly conserved hydrophobic proteins. *12th International Conference on Arabidopsis Research* (2001)

Medina, J., Rodriguez-Franco, M., Penalosa, A., Carrascoasa , M.J., Neuhaus, G., Salinas, J. Arabidopsis mutants deregulated in RCI2A expression reveal new signaling pathways in abiotic stress responses. *Plant J*; ;42(4):586-97 (2005)

Mitsuya, S., Taniguchi, M., Miyake, H., Takabe, T. Disruption of RCI2A leads to over-accumulation of Na(+) and increased salt sensitivity in Arabidopsis thaliana plants. *PLANTA*: 222(6):1001-9 (2005)

Morikami A, Matsunaga R, Tanaka Y, Suzuki S, Mano S, Nakamura K. Two cis-acting regulatory elements are involved in the sucrose-inducible expression of the sporamin gene promoter from sweet potato in transgenic tobacco. *Mol Genet Genomics*. 272:690-699 (2005)

Moses, A. M., Chiang, D. Y., Eisen, M. B. (2004) Phylogenetic motif detection by expectation-maximization on evolutionary mixtures. *Pac Symp Biocomput*. 324–335

Murphy, D. Gene expression studies using microarrays: principles, problems, and prospects. *Advance in Physiology Education*. 26:256-270 (2002)

Nacht, M., Ferguson, A.T., Zhang, W., Petroziello, J.M., Cook, B.P., Gao, Y.H., Maguire, S., Riley, D., Coppola, G., Landes, G.M., Madden, S.L., Sukumar, S. Combining serial analysis of gene expression and array technologies to identify genes differentially expressed in breast cancer. *Cancer Research*, 59:5464 –5470 (1999)

Nagalakshmi, U., et al. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*;320:1344–1349 (2008)

Ngai N, Tsai FY, Coruzzi G. Light-induced transcriptional repression of the pea AS1 gene: identification of cis-elements and transfactors. *Plant J*. 12:1021-1234 (1997)

Nunberg A, Li Z, Bogue M, Vivekananda J, Reddy A, Thomas TL. unpublished results (cited in a review by Thomas TL, 1993)

Ono A, Izawa T, Chua N-H, Shimamoto K. The rab16B promoter of rice contains two distinct abscisic acid-responsive elements. *Plant Physiol*. 112: 483-491 (1996)

Oono, Y., Chen, Q.G., Overvoorde, P.J., Kohler, C., and Theologis, A. Age Mutants of Arabidopsis exhibit altered auxin-regulated gene expression. *The Plant cell* 10, 1649- 1662 (1998)

Östlund, G., Schmitt, T., Forslund, K., Köstler,T., Messina, D.N., Roopra, S., Frings, O., Sonnhammer, E.L.L. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. *Nucleic Acids Res*: January 38(Database issue): D196–D203 (2010)

Ozsolak, F., Platt, A.R., Jones, D.R., Reifenberger, J.G., Sass, L.E., McInerney, P., Thompson, J.F., Bowers, J., Jarosz, M. Milo, P.M. Direct RNA sequencing. *Nature* Vol 461 (2009)

Ozsolak, F., Milos, P.M. RNA sequencing: advances, challenges and opportunities. Nature Reviews, *Nature* 461, 814-818 (2009)

Pauli S, Rothnie HM, Chen G, He X, Hohn T. The cauliflower mosaic virus 35S promoter extends into the transcribed region. *J Virol*. 78: 12120-12128(2004)

Pavesi, G., Pesole, G. Using Weeder for the discovery of conserved transcription factor binding sites. *Current protocols in bioinformatics editoral board Andreas D Baxevanis et al*, *Chapter 2*, Unit 2.11 (2006)

Pearson, H. Genetics: what is a gene? *Nature,* 441(7092), 398-401 (2006)

Pennacchio L.A., et al. An apolipoprotein influencing triglycerides in humans and mice revealed by comparative sequencing. *Science,* 294:169-173 (2001)

Perocchi, F., Xu, Z., Clauder-Munster, S., Steinmetz, L. M. Antisense artifacts in transcriptome microarray experiments are resolved by actinomycin D. *Nucleic Acids Res*. 35, e128 (2007).

Petit JM, van Wuytswinkel O, Briat JF, Lobreaux S. Characterization of an iron-dependent regulatory sequence involved in the transcriptional control of AtFer1 and ZmFer1 plant ferritin genes by iron. *J Biol Chem*. 276:5584-5590. (2001)

Phillips, T. Regulation of transcription and gene expression in eukaryotes. *Nature Education 1(1)* (2008)

Potenza, C., Aleman, L., Sengupta-Gopalan, C. Invited Review: Targeting transgene expression in research, agricultural and environmental applicationsL Promoters used in plant transformation. *In Vitro Cellular and Developmental Biology – Plant*: 40:1-22 (2004)

Qi, X., Zhang, Y., Chai, T. Characterization of a Novel Plant Promoter Specifically Induced by Heavy Metal and Identification of the Promoter Regions Conferring Heavy Metal Responsiveness. *Plant Physiology*, 143, 50-59 (2007)

Raab, J.R., Kamakaka, R.T. Insulators and promoters: closer than we think. *Nat Rev Genet*. Jun;11(6):439-46 (2010)

Rahman A, Hosokawa S, Oono Y, Amakawa T, Goto N, Tsurumi S. Auxin and ethylene response interactions during Arabidopsis root hair development dissected by auxin influx modulators. *Plant Physiol*. 130:1908–17 (2002)

Ren, M., Chen, Q., Li, L., Zhang, R., Guo, S. Functional analysis of a reproductive organ predominant expressing promoter in cotton plants. *Sci China C Life Sci*. Oct;48(5):452-459 (2005)

Riano-Pachon DM, Ruzicic S, Dreyer I, Mueller-Roeber B: PlnTFDB: an integrative plant transcription factor database. *BMC Bioinformatics* 2007, 8:42

Roberts, G.R., Garoosi, G.A., Koroleva, O., Ito, M., Laufs, P., Leader, D.J., Caddick, M.X., Doonan, J.H., Tomsett, A.B. The alc-GR system: a modified alc gene switch designed for use in plant tissue culture. *Plant Physiol*. Jul;138(3): 1259-1267 (2005)

Rodriguez, A., Vigorito, E., Clare, S., Warren, M.V., Couttet, P., et al. Requirement of bic/microRNA-155 for normal immune function. *Science,* 316(5824): 530 (2007)

Rombauts, S., Dehais, P., Van Montagu, M., Rouze, P. PlantCARE, a plant cis-acting regulatory element database. *Nucleic Acids Res* 27: 295–296 (1999)

Roslan, H.A., Salter, M.G., Wood, C.D., White, M.R., Croft, K.P., Robson, F., Coupland, G., Doonan, J., Laufs, P., Tomsett, A.B., Caddick, M.X. Characterization of the ethanol-inducible alc gene-expression system in Arabidopsis thaliana. *Plant Journal*. Oct;28(2):225-235 (2001)

Roy, S. W., Irimia, M. When good transcripts go bad: artifactual RT-PCR ‘splicing’ and genome analysis. *Bioessays* 30, 601–605 (2008)

Rounsley, S., Marri, P.R., Yu, Y., He, R., Sisneros, N., Goicoechea, J.L., Lee, S.J., Angelova, A., Kudrna, D., Luo, M., Affourtit, J., Desany, B., Knight, J., Niazi, F., Egholm, M., Wing, R.A. De Novo Next Generation Sequencing of Plant Genomes. *Rice,* 2:35–43 (2009)

Rushton, P.J., Reinstadler, A., Lipka, V., Lippok, B., Somssich, I.E. Synthetic plant promoters containing defined regulatory elements provide novel insights into pathogen- and wound-induced signaling. *Plant Cell*. April; Vol. 14(4);749-762 (2002)

Saeed, H.A., Vodkin, L.O., Stromvik, M.V. Promoters of the soybean seed lectin homologues Le2 and Le3 regulate gene expression in vegetative tissues in Arabidopsis. *Plant Science,* 175, 868-876 (2008)

Sablowski RWM, Moyano E, Culianez-Macia FA, Schuch W, Martin C, Bevan M. A flower-specific Myb protein activates transcription of phenylpropanoid biosynthetic genes. *EMBO J*. 13:128-137 (1994)

Sakai H, Aoyama T, Oka A. Arabidopsis ARR1 and ARR2 response regulators operate as transcriptional activators. *Plant J*. 24: 703-711 (2000)

Sakai T, Takahashi Y, Nagata T. The identification of DNA binding factor specific for as-1-like sequences in auxin-responsive regions of parA, parB and parC. *Plant Cell Physiol*. 39:731-739 (1998)

Sangwan I, O'Brian MR. Identification of a Soybean Protein That Interacts with GAGA Element Dinucleotide Repeat DNA. *Plant Physiol*. 129: 1788-1794 (2002)

Santi L, Wang Y, Stile MR, Berendzen K, Wanke D, Roig C, Pozzi C, Muller K, Muller J, Rohde W, Salamini F. The GA octodinucleotide repeat binding factor BBR participates in the transcriptional regulation of the homeobox gene Bkn3. *Plant J*. 34:813-826 (2003)

Satoh R, Nakashima K, Seki M, Shinozaki K, Yamaguchi-Shinozaki K. ACTCAT, a novel cis-acting element for proline- and hypoosmolarity-responsive expression of the ProDH gene encoding proline dehydrogenase in Arabidopsis. *Plant Physiol*. 130:709-719 (2002)

Schena, M., Shalon, D., Davis, R.W., Brown, P.O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270: 467–470 (1995)

Schena, M., Lloyd, A. M., Davis, R. W. A steroid-inducible gene expression system for plant cells. *Proc. Natl. Acad. Sci. USA. Genetics*. Vol. 88, p10421-10425 (1991)

Schmutz, J., Cannon, S.B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D.L., Song, Q., Thelen, J.J., Cheng, J., Xu, D., Hellsten, U., May, G.D., Yu, Y., Sakurai, T., Umezawa, T., Bhattacharyya, M.K., Sandhu, D., Valliyodan, B., Lindquist, E., Peto, M., Grant, D., Shu, S., Goodstein, D., Barry, K., Futrell-Griggs, M., Abernathy, B., Du, J., Tian, Z., Zhu, L., Gill, N., Joshi, T., Libault, M., Sethuraman, A., Zhang, X.C., Shinozaki, K., Nguyen, H.T., Wing, R.A., Cregan, P., Specht, J., Grimwood, J., Rokhsar, D., Stacey, G., Shoemaker, R.C., Jackson, S.A. Genome sequence of the palaeopolyploid soybean. *Nature* 463(7278):178-83 (2010)

Schulze-Lefert P, Dangl JL, Becker-Andre M, Hahlbrock K, Schulz. Inducible in vivo DNA footprints define sequences necessary for UV light activation of the parsley chalcone synthase gene. *EMBO J*. 8: 651-656 (1989)

Shahmuradov, I., Gammerman, A., Hancock, J., Bramley, P., Solovyev, V. PlantProm: a database of plant promoter sequences. *Nucl. Acids Res*. 31 (1): 114-117 (2003)

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

Siddharthan, R., Siggia, E. D., van Nimwe- gen, E. (2005) PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny. *PLoS Comput Biol*. 1, e67.

Simpson SD, Nakashima K, Narusaka Y, Seki M, Shinozaki K, Yamaguchi-Shinozaki K. Two different novel cis-acting elements of erd1, a clpA homologous Arabidopsis gene function in induction by dehydration stress and dark-induced senescence. *Plant J*. 33: 259-270 (2003)

Sinha, S. Discriminative motifs. *J Comput Biol* 10, 599–615 (2003)

Sinha, S., Blanchette, M., Tompa, M. (2004) PhyME: a probabilistic algorithm for finding motifs in sets of orthologous sequences. *BMC Bioinformatics*. 5, 170.

Smale, S. T. & Kadonaga, J. T. The RNA polymerase II core promoter. *Annual Review of Biochemistry,* 72 : 449-79 (2003)

Soboleski, M.R., Oaks, J., Halford, W.P. Green fluorescent protein is a quantitative reporter of gene expression in individual eukaryotic cells. *(2005)* The FASEB Journal; 19:440-442.

Solano R, Stepanova A, Chao Q, Ecker JR. Nuclear events in ethylene signaling: a transcriptional cascade mediated by ETHYLENE-INSENSITIVE3 and ETHYLENE-RESPONSE-FACTOR1. *Genes Dev*. 12: 3703-3714 (1998)

Sproul, D., Gilbert, N., & Bickmore, W. The role of chromatin structure in regulating the expression of clustered genes. Nature Reviews Genetics 6, 775–781 (2005)

Stalberg K, Ellerstom M, Ezcurra I, Ablov S, Rask L. Disruption of an overlapping E-box/ABRE motif abolished high transcription of the napA storage-protein promoter in transgenic Brassica napus seeds. *Planta*. 199:515-519 (1996)

Stormo, G.D. DNA binding sites: representation and discovery. *Bioinformatics* 16(1), 16-23 (2000)

Stromvik, M.V., Sundararaman, V.P., Vodkin, L.O. A novel promoter from soybean that is active in a complex developmental pattern with and without its proximal 650 base pairs. *Plant Mol Biol*, 41(2), 217-231 (1999)

Struhl, K. Fundamentally different logic of gene regulation in eukaryotes and prokaryotes. *Cell* 98, 1–4 (1999)

Struhl, K., Kadosh, D., Keaveney, M., Kuras, L. & Moqtaderi, Z. Activation and repression mechanisms in yeast. *Cold Spring Harb. Symp. Quant Biol* 63, 413–421 (1998)

Sugimoto K, Takeda S, Hirochika H. Transcriptional activation mediated by binding of a plant GATA-type zinc finger proteinAGP1 to the AG-motif (AGATCCAA) of the wound-inducible Myb gene NtMyb2. *Plant J*, 36: 550-564 (2003)

Takeda J, Ito Y, Maeda K, Ozeki Y. Assignment of UVB-responsive cis-element and protoplastization-(dilution-) and elicitor-responsive ones in the promoter region of a carrot phenylalanine ammonia-lyase gene (gDcPAL1). *Photochem Photobiol*. 76:232-238 (2002)

The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* 408, 796-815 (2000)

Thijs, G., Lescot, M., Marchal, K., Rombauts, S., De Moor, B., Rouze, P., et al. A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics*, 17(12), 1113-1122 (2001)

Thomas TL. Gene expression during plant embryogenesis and germination: An overview. *Plant Cell*. 5:1401-1410 (1993)

Thompson, W., Rouchka, E. C., Lawrence, C. E. Gibbs Recursive Sampler: finding transcription factor binding sites. *Nucleic Acids Res*. 31, 3580–3585 (2003)

Timothy L. Bailey, Mikael Bodén, Fabian A. Buske, Martin Frith, Charles E. Grant, Luca Clementi, Jingyuan Ren, Wilfred W. Li, William S. Noble, "MEME SUITE: tools for motif discovery and searching", *Nucleic Acids Research*, 37:W202-W208 (2009)

Uimari A, Strommer J. Myb26: a MYB-like protein of pea flowers with affinity for promoters of phenylpropanoid genes. *Plant J*. 12:1273-1284 (1997)

Ulmasov T, Hagen G, Guilfoyle TJ. Dimerization and DNA binding of auxin response factors. *Plant J*. 19:309-319 (1999)

Urao T, Yamaguchi-Shinozaki K, Urao S, Shinozaki K. An Arabidopsis myb homolog is induced by dehydration stress and its gene product binds to the conserved MYB recognition sequence. *Plant Cell*. 5:1529-1539 (1993)

Velculescu, V.E., Zhang, L., Vogelstein, B., Kinzler, K.W. Serial analysis of gene expression. *Science* 270: 484–487 (1995)

Vaucheret, H. Post-transcriptional small RNA pathways in plants: mechanisms and regulations. Genes & Dev. 20: 759-771 (2006)

Velculescu, V.E. Tantalizing Transcriptomes--SAGE and Its Use in Global Gene Expression Analysis. *Science*, Vol. 286 no. 5444 pp. 1491-1492 (1999)

Vera, J.C., et al. Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Mol. Ecol*;17:1636–1647 (2008)

Vernoux, T., Besnard, F., Traas, J. Auxin at the shoot apical meristem. *CSH Perspect Biol*. Apr;2(4):a001487 (2010)

Villain P, Clabault G, Mache R, Zhou DX. S1F binding site is related to but different from the light-responsive GT-1 binding site and differentially represses the spinach rps1 promoter in transgenic tobacco. *J Biol Chem,* 269:16626-16630 (1994)

Vlieghe, D., Sandelin, A., Bleser, P. J. D., Vleminckx, K., Wasserman, W. W., van Roy, F., and Lenhard, B. (2006). A new generation of jaspar, the open-access repository for transcription factor binding site profiles. Nucleic Acids Res, 34(Database issue):D95–D97.

Washida H, Wu CY, Suzuki A, Yamanouchi U, Akihama T, Harada K, Takaiwa F. Identification of cis-regulatory elements required for endosperm expression of the rice storage protein glutelin gene GluB-1. *Plant Mol Biol*. 40:1-12 (1999)

Weiher H, Konig M, Gruss P. Multiple point mutations affecting the simian virus 40 enhancer. *Science*. 219:626-631 (1983)

Weinhold, A., Kallenbach, M., Baldwin, I.T. Progressive 35S promoter methylation increases rapidly during vegetative development in transgenic *Nicotiana attenuata* plants. *BMC Plant Bio*. 13:99 (2013)

Wingender E, Chen X, Hehl R, Karas H, Liebich I, Matys V, Meinhardt T, Pruss M, Reuter I, Schacherer F: TRANSFAC: an integrated system for gene expression regulation. *Nucleic Acids Res* 2000, 28(1):316-319

Woolfe, A., *et al*. Highly conserved non-coding sequences are associated with vertebrate development. *Plos Biol*. 3:e7 (2005)

Workman, C. T., Stormo, G. D. (2000) ANN-Spec: a method for discovering transcription factor binding sites with improved specificity. *Pac Symp Biocomput*, 467–478

Wray G.A., Hahn, M.W., Abouheif, E., Balhoff, J.P., Pizer, M., Rockman, M.V., Romano, L.A. The evolution on transcriptional regulation in eukaryotes. *Molecular Biology and Evolution:* 20(9):1377-1419 (2003)

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

Wyrick, J. J. & Young, R. A. Deciphering gene expression regulatory networks. *Curr. Opin. Genet. Dev*. 12, 130−136 (2002)

Xie, X., Lu, J., Kulbokas, E. J., et al. (2005) Systematic discovery of regulatory motifs in human promoters and 3 UTRs by comparison of several mammals. *Nature*. 434, 338-345.

Yamagata H, Yonesu K, Hirata A, Aizono Y. TGTCACA motif is a novel cis-regulatory enhancer element involved in fruit-specific expression of the cucumisin gene. *J Biol Chem*. 277:11582-11590 (2002)

Yamamoto, M., Wakatsuki, T., Hada, A., Ryo, A. Use of serial analysis of gene expression technology. *Journal of Immunology Methods* 250:45 – 66 (2001)

Yang, C., Bolotina, E., Jiang, T., Sladek, F.M., Martinez. E. Prevalence of the Initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA- less core promoters. *Gene*. March 1; 389(1): 52–65 (2007)

Yang, Y., Costa, A., Leonhardt, N., Siegel, R.S., Schroeder, J.I. Isolation of a strong Arabidopsis guard cell promoter and its potential as a research tool. *Plant Methods* 4:6 (2008)

Yin Y, Beachy RN. The regulatory regions of the rice tungro bacilliform virus promoter and interacting nuclear factors in rice (Oryza sativa L.). *Plant J*. 7:969-980 (1995)

Yin Y, Chen L, Beachy R. Promoter elements required for phloem-specific gene expression from the RTBV promoter in rice. *Plant J*. 12:1179-1188 (1997)

Yoo, S.K., Wu, X., Lee, J.S., Ahn, J.H. AGAMOUS-LIKE 6 is a floral promoter that negatively regulates the FLC/MAF clade genes and positively regulates FT in Arabidopsis. *Plant Journal*. Jan;65(1):62-76 (2011)

Zamorano P.L., Mahesh V.B., Brann D.W. Quantitative RT-PCR for neuroendocrine studies. A minireview. *Neuroendocrinology* 63 397–407 (1996)

Zavallo, D., Lopez Bilbao, M., Hopp H.E., Heinz, R. Isolation and functional characterization of two novel seed-specific promoters from sunflower (*Helianthus annuus* L.). *Plant Cell Rep*. Mar;29(3)239-248 (2010)

Zhang, H., Li, L. *SQUAMOSA promoter binding protein-like 7* regulated in microRNA408 is required for vegetative development in Arabidopsis. *Plant Journal*. Vol. 74(1):98-109 (2013)

Zhou, X., Ruan, J., Wang, G., Zhang, W. Characterization and Identification of MicroRNA Core Promoters in Four Model Species. *PLoS Comput Biol*, 3(3): e37 (2007)

Zhou DX, LI YF, Rocipon M, Mache R. Sequence specific interaction between S1F, a spinach nuclear factor, and a negative cis-element conserved in plastid-related genes. *J Biol Chem,* 267:23515-23519 (1992)

Appendix

Appendix 4.1.2.1

All-against-all Smith-Waterman peptide sequence alignment

Peptide sequences of 18 plant species were aligned using Smith-Waterman algorithm in DeCypher® TimeLogic® Biocomputing System (Active Motif Inc., Carlsbad, CA).

- All-against-all Smith-Waterman alignment script:

```
cat [QueueList.txt] | while read line;
do

query="[Species]"
database=$line"_aa";
peptide=$line"_aa.faa";

dc_run -p sw_aa_vs_aa -q $peptide -d $database -output_format tab fieldrecord -field
querylocus targetlocus score querylength targetlength querystart queryend targetstart
targetend -score 50 -max_scores 500 -max_alignments 500 -detach -description "SW
$line vs. $line for InParanoid";

dc_run -p sw_aa_vs_aa -q $peptide -d $query"_aa" -output_format tab fieldrecord -field
querylocus targetlocus score querylength targetlength querystart queryend targetstart
targetend -score 50 -max_scores 500 -max_alignments 500 -detach -description "SW
$line vs. $query for InParanoid";

dc_run -p sw_aa_vs_aa -q $query"_aa.faa" -d $database -output_format tab fieldrecord -
field querylocus targetlocus score querylength targetlength querystart queryend targetstart
targetend -score 50 -max_scores 500 -max_alignments 500 -detach -description "SW
$query vs. $line for InParanoid";

done;
```

- Smith-Waterman alignment results were parsed using a Python script:

```
def inparanoid_format(filename):
    from math import fabs

    first_orgn = filename[0:5]
    second_orgn = filename[6:11]
    inparanoid_out = first_orgn + "-" + second_orgn
    inp_out = open(inparanoid_out, "w")
    previous_are_same = 0
    try:
        with open(filename, "r") as file:
            prev_line = file.readline().split()

            for line in file:
                line = line.split()

                if line[0] == prev_line[0] and line[1] == prev_line[1]:
                    if previous_are_same == 0:

                        if int(line[5]) >= int(prev_line[5]):
                            longest_query = int(fabs(int(line[6])
- int(prev_line[5])) + 1)
                            longest_target = int(fabs(int(line[8])
- int(prev_line[7])) + 1)
                            total_query =
int(fabs(int(prev_line[6]) - int(prev_line[5])) + fabs(int(line[6])
-int(line[5])) + 2)
                            total_target =
int(fabs(int(prev_line[8]) - int(prev_line[7])) + fabs(int(line[8])
 -int(line[7])) + 2)
                            first_hsp =      "q:" + prev_line[5] +
"-" + prev_line[6] + " " + "h:" + prev_line[
7] + "-" + prev_line[8]
                            second_hsp = "q:" + line[5] + "-" +
line[6] + " " + "h:" + line[7] + "-" + line[8]
                        else:
                            longest_query =
int(fabs(int(prev_line[6]) - int(line[5])) + 1)
                            longest_target =
int(fabs(int(prev_line[8]) - int(line[7])) + 1)
                            total_query = int(fabs(int(line[6]) -
int(line[5])) + fabs(int(prev_line[6]) -int(
prev_line[5])) + 2)
                            total_target = int(fabs(int(line[8]) -
int(line[7])) + fabs(int(prev_line[8]) -int
(prev_line[7])) + 2)
                            first_hsp =      "q:" + line[5] + "-" +
line[6] + " " + "h:" + line[7] + "-" + line
[8]
                            second_hsp = "q:" + prev_line[5] +
"-" + prev_line[6] + " " + "h:" + prev_line[7]
+ "-" + prev_line[8]


                        line_to_write = prev_line[0] + "\t" +
prev_line[1] +"\t" + prev_line[2] + "\t" + prev_line
```

```
[3] + "\t" + prev_line[4] + "\t" + str(longest_query) + "\t" + str(longest_target) + "\t" +
str(total_query) + "\t" + str(total_target) +
"\t" + first_hsp + "\t" + second_hsp + "\n"
                        inp_out.write(line_to_write)



                    prev_line = line
                    previous_are_same = 1

                if line[1] != prev_line[1]:
                    if previous_are_same == 0:
                        longest_query = total_query =
int(fabs(int(prev_line[6]) - int(prev_line[5])) + 1)
                        longest_target = total_target =
int(fabs(int(prev_line[8]) - int(prev_line[7])) + 1)
                        line_to_write = prev_line[0] + "\t" +
prev_line[1] +"\t" + prev_line[2] + "\t" + prev_line
[3] + "\t" + prev_line[4] + "\t" + str(longest_query) + "\t" + str(longest_target) + "\t" +
str(total_query) + "\t" + str(total_target) +
"\t" + "q:" + prev_line[5] + "-" + prev_line[6] + " " + "h:" + prev_line[7] + "-" +
prev_line[8] + "\n"
                        inp_out.write(line_to_write)

                    prev_line = line
                    previous_are_same = 0
    except IndexError:
        print "Problems with last line"

from sys import argv

script, file = argv
inparanoid_format(file)
```

- Unix script was used to sort file:

```
ls *.tab | while read file;
do
echo "Sorting the file" $file
sed "1d" $file | sed '/^$/d' | sort -t $'\t' -k 1b,1 -k 2b,2 -k 3,3rn > temp.txt;
mv temp.txt $file;
echo "Formatting the file: $file"
python format_tab.py $file;
done
```

Appendix 4.1.2.2

Removal of all orthologous gene members that have confidence value lower than 1.0 in InParanoid results

Prior to running *qp* (QuickParanoid), any gene members of the COG that has InParanoid confidence value less than 1.0 are removed. Therefore, only the genes that have 1.0 InParanoid confidence value are included in further analysis.

❖ Removing any gene members in COGs that have confidence value lower than 1.0:

```
file_out = open('sqltable.volca-zeama_2', 'w')
with open('sqltable.volca-zeama', 'r') as file:
    for line in file:
        line = line.strip()
        if '1.000' in line:
            line_out = line + '\n'
            file_out.write(line_out)
file_out.close()
```

<u>**Appendix 4.1.2.3**</u>

Run QuickParanoid

A configuration file containing the list of species names in the dataset, separated by line, was created and stored in the same directory.

❖ Run QuickParanoid to merge COGs generated by InParanoid:

```
Dataset directory [default = "." (current directory)]: 18sp_quickparanoid
Data file prefix [default = "sqltable."]:
Data file separator [default = "-"]:
Configuration file [default = "18sp_quickparanoid/config"]:
Executable file prefix [default = "test"]:
```

QuickParanoid first preprocesses the entire dataset. An executable program dump was generated, and for each data file *File* in the dataset, a new file *File*_c was created in the working directory. The result of ortholog clustering was displayed in the MultiParanoid (Alexeyenko et al, 2006) output format and redirected to a text format as shown below.

❖ Converting QuickParanoid output file to text format:

```
./test > resultQP_araly-arath-bradi-carpa-chlre-glyma-linus-maldo-manes-medtr-orysa-
phypa-poptr-selmo-sorbi-vitvi-volca-zeama.txt
```

To extract information such as, clusterID, species, gene; the text output of QuickParanoid was then parsed using a simple Unix command line.

❖ Parse QuickParanoid output:

```
awk '{print $1, $2, $3}' resultQP_araly-arath-bradi-carpa-chlre-glyma-linus-maldo-
manes-medtr-orysa-phypa-poptr-selmo-sorbi-vitvi-volca-zeama.txt >
18sp_QPresult_Parsed.txt
```

The transcript names found in the QuickParanoid output file (COG file) as well as in the deflines of promoter sequence file need to be removed.

❖ Remove transcript name after gene name:

```
sed 's/|.*//' Filename1.txt > OutputFilename.txt
```

Appendix 4.1.3

Retrieve promoter sequence

Promoter sequences corresponding to the orthologous genes in the clusters were then retrieved using BioPython Bio.SeqIO (Cock, et al., 2009).

❖ Retrieving promoter sequences for each orthologous gene:

```
from Bio import SeqIO
import csv
testdict_all18species = SeqIO.index("18sp_prom_mod.fas", "fasta")
csvReader = csv.reader(open('18sp_cog_mod.txt', 'rb'), delimiter=" ")
csvWriter = csv.writer(open('18sp_cog_to_prom.txt', 'w'), delimiter="\t")
error_file = open("errors_18sp.txt", 'a')
for line in csvReader:
    try:
      Promoter = str(testdict_all18species[line[2]].seq)
      csvWriter.writerow([line[0], line[1], line[2], Promoter])
    except KeyError:
      #message = "Error with %s, not found in promoters file\n" % line[2]
      message = "%s\n" % line[2]
      error_file.write(message)
```

Appendix 4.1.4

De novo motif discovery using Seeder

Index file was generated to improve the performance of Hamming Distance (HD) calculation.

- Generation of index file:

```
use Seeder::Index;
my $index = Seeder::Index->new(
seed_width => "6",
out_file => "6.index",
);
$index->get_index;
```

Then, a background distribution file was generated for all 18 genomes.

- Generation of the background distribution file:

```
use Seeder::Background;
my $background = Seeder::Background->new(
seed_width => "6",
strand => "revcom",
hd_index_file => "6.index",
seq_file => "all18sp_promoters.fas",
out_file => "all18sp_prom_1K.bkgd",
);
$background->get_background;
```

In order to run the motif discovery, promoter sequences for each gene members in the COG in separate file was needed. This constituted as the “positive set”. The whole 18 species set of promoter sequences were used for the computation of background model only. Every promoter sequence of each gene members of each COGs were queried through *OrthoProMof* MySQL database, where a temporary FASTA file was written and the percentage of numbers of N in the sequence was calculated. The motif finder *finder.pl* script was also included in this query script and put in strings.

- Automatic query of motif discovery for each COG:

```
import MySQLdb as sql

import time
#import csv
from subprocess import call
from sys import argv

script, first_cog, last_cog = argv
print "Connecting to the MySQL database"
db = sql.connect(host="192.168.1.1", user="nadchai", db="OrthoProMofDB")

for cog_id in range(int(first_cog), int(last_cog)): ##put one more number on the last
number, in this case this  is 32158
        time_start = time.time()
        print decor
        print "Querying COG#%s" % cog_id
        c.execute("""SELECT Gene_ID, PromSeq FROM `COGtoProm_Trimmed`
WHERE COG_ID = %s""", (cog_id,))
        result = c.fetchall()

        print "Writing FASTA file"
        promoter_filename = "promoters_%s.fas" % cog_id
        fasta_output = open(promoter_filename, "w") ##creating test1.fas
        for line in result:
                defline = ">" + line[0] + "\n" ##takes the > sign, adds the gene name to it
and create a new line
                sequence = line[1].strip('\r') + "\n"
                npercent = (float(sequence.upper().count('N'))/len(sequence)) * 100 #
NOTE: modified npercent calculation to use .
count('N'), resulting in a simpler code

                if npercent <= 40:
                        fasta_output.write(defline)
                        fasta_output.write(sequence)
                else:
                        continue

        fasta_output.close()

        finder_wrapper = """use Seeder::Finder;
my $finder = Seeder::Finder->new(
seed_width => "6",
strand => "revcom",
motif_width => "12",
n_motif => "10",
hd_index_file => "6.index",
seq_file => "%s",
bkgd_file => "all18sp_prom_1K.bkgd",
out_file => "cog_%s.finder",
);
$finder->find_motifs;
```

```
    """ % (promoter_filename, cog_id)

    finder_filename = "finder_%s.pl" % cog_id
    finder_file = open(finder_filename, "w")
    finder_file.write(finder_wrapper)
    finder_file.close()

    print "Number of orthologs in cluster:"
    call(["egrep", "-c", ">", promoter_filename])
    print "Running Seeder"
    call(["perl", "%s" % finder_filename])
    print "Seeder results output in cog_%s.finder" % cog_id
    call(["rm", "%s" % finder_filename])
    call(["rm", "%s" % promoter_filename])

    run_time_minutes = (time.time() - time_start)//60
    run_time_seconds = (time.time() - time_start)%60
    print "It took %d:%02d minutes to analyze COG# %s" % (run_time_minutes,
run_time_seconds, cog_id)
    print décor
```

To expedite the motif finding step, a Q submission system was used where the number of processors to be used and what range of COG to be analyzed were indicated in the script (Note: Skogul computer has 32 processors).

- Bash script that submits one query:

```
#!/bin/bash

ARGV0=$0 # First argument is shell command, the script itself
COG1=$1
COG2=$2

python query_cogs.py $COG1 $COG2
```

- Submit queries to Q submission system:

```
from sys import argv
from subprocess import call

script, cog1, cog2 = argv

cog1 = int(cog1)
cog2 = int(cog2)
```

```
processors = 32

cogs_per_processor = (cog2 - cog1)/ processors #half of the processors on compute
nodes
print cogs_per_processor

cogs_per_processor_remainder = (cog2 - cog1)%processors
print cogs_per_processor_remainder

cog_pairs = [[i, i + cogs_per_processor] for i in range(cog1, cog2, cogs_per_processor)]

for cog_pair in cog_pairs[0:len(cog_pairs)-1]:
        print "Submiting COGs %d to %d" % (cog_pair[0], cog_pair[1])
        call('qsub -cwd -S /bin/bash -pe mpi 1 -R y submit_one_query.sh %d %d' %
(cog_pair[0], cog_pair[1]), shell=True)

print "Submiting COGs %d to %d" % (cog_pairs[-1][0], cog2 + 1)
call('qsub -cwd -S /bin/bash -pe mpi 1 -R y submit_one_query.sh %d %d'% (cog_pairs[-
1][0], cog2 + 1), shell=True)
```

In order to check the process after submission, *qstat* command was used. It described how many processors were occupied at the moment; therefore more jobs were submitted when there were processors available. Since we had a large number of COGs, it was very important to keep track of the job submission logs so there would be no overlaps or COG left behind.

Only motifs that had *Q*-value lower than 0.01 (*Q*-value < 1.0e-02) were considered to be significant (Fauteux et al., 2008). Therefore, a simple Unix command line was used to extract only the ones with *Q-values* higher than the threshold.

- Filter significant motifs using Unix command:

```
awk -F= '$2 < 1.0e-02' Q-valuefile.txt > Significant_Qvalue.txt
```

Appendix 4.2.1.1

Analyze differential expression RNA-Seq data using DESeq

Raw counts files from RNA-Seq were analyzed using DESeq to see differential expression between each tissue (shoot epidermis *vs*. shoot apical meristem) in each Soybean cv. (Clark standard, 5-Leaflet mutant and Glabrous mutant).

- DESeq script used for differential expression analysis:

```
# Differential gene expression using DESeq
# To use this script : Rscript deseq.R -d designfile -c rawcountfile -o output_dir

library(DESeq)

# Usage

usage=function(errM) {
        cat("\nUsage : Rscript deseq.R [option] <Value>\n")
        cat("       -d        : design file\n")
        cat("       -c        : raw count file\n")
        cat("       -o        : output directory\n")
        cat("       -h        : this help\n\n")
        stop(errM)
}
set.seed(123456789)
perform_dge=function(counts, groups, count_limit, path) {

# Retain row which have > count_limit

counts<-counts[rowSums(counts) > count_limit,]

# Normalize and do test

cds<-newCountDataSet(counts, groups)
cds<-estimateSizeFactors(cds)
sizeFactors(cds)
if(length(groups)==2) {
cds<-estimateDispersions(cds, method="blind", sharingMode="fit-only")
}
else {
cds<-estimateDispersions(cds, method="pooled")
}

res<-nbinomTest(cds, "1", "2" )
```

```
res[,c(5,6)] = round(res[,c(5,6)], digits=3)
res[,7] = as.numeric(format(res[,7], digits=2))
res[,8] = as.numeric(format(res[,8], digits=2))
colnames(res)[c(1,7,8)] = c("id", "deseq.p-value", "deseq.adj.pvalue")
write.table(res[order(res[,8]), c(1,7,8)], paste(path,"deseq_results.csv",sep="/"), quote =
FALSE, sep = "\t",  eol = "\n", na = "NA", dec = ".", row.names = FALSE, col.names =
TRUE)
fileOpen=paste(path,"edger_results.csv",sep="/")
d1<-read.table(fileOpen, header=T, sep="\t", quote="")
d2<-merge(d1, res[, c(1,7,8)], by.x=1, by.y=1, sep="\t")
d2<-d2[order(d2[,(ncol(d2)-1)]),]
vecWrite<-c(1:4, (ncol(d2)-1), ncol(d2), 5:6, 7:(ncol(d2)-2))
write.table(d2[,vecWrite], paste(path,"dge_results.csv",sep="/"), quote = FALSE, sep =
"\t",  eol = "\n", na = "NA", dec = ".", row.names = FALSE, col.names = TRUE)
}

##################################

ARG = commandArgs(trailingOnly = T)
## default arg values
count_limit=9
fpath="."
design_file=""
rawcount_file=""
out_path=""
## get arg variables
for (i in 1:length(ARG)) {
        if (ARG[i] == "-d") {
          design_file=ARG[i+1]
        } else if (ARG[i] == "-c") {
          rawcount_file=ARG[i+1]
        } else if (ARG[i] == "-o") {
          out_path=ARG[i+1]
        } else if (ARG[i] == "-h") {
          usage("")
        }
}
## check arg consitency
if (!(file.exists(design_file))) {
        usage("Error : Design file not found")
}
if (!(file.exists(rawcount_file))) {
        usage("Error : Raw count file not found")
}
if (out_path == "") {
        usage("Error : Output directory not specified")
```

```
}
tmpFP=strsplit(fpath,"")
if (tmpFP[[1]][length(tmpFP[[1]])] == "/" ) {
        bckS=""
} else {
        bckS="/"
}
tmpOP=strsplit(out_path,"")
if (tmpOP[[1]][length(tmpOP[[1]])] == "/" ) {
        out_path=paste(tmpOP[[1]][1:(length(tmpOP[[1]]-1))],collapse="")
}

design = read.csv2(design_file, header=T, sep = "\t", na.strings = "0", check.names=F)
rawcount = read.csv(rawcount_file, header=T, sep ="\t", check.names=F)

print(design)

name_sample= as.character(as.vector(design[,1]))
countMatrix = rawcount[,3:ncol(rawcount)]

# Iterate over each design

for (i in 2:ncol(design)) {

        name_folder = paste(out_path,names(design[i]),sep="/")

         # Create output directory

        if (!file.exists(name_folder)) {
          system(paste("mkdir",name_folder,sep=" "))
        }

       current_design=design[,i]
       subsampleN=name_sample[!(is.na(current_design))]
       group = as.character(current_design)[!(is.na(current_design))]
       groupN = unique(group)
       current_countMatrix = NULL
       for (j in 1:length(subsampleN)) {

current_countMatrix=cbind(current_countMatrix,countMatrix[,is.element(colnames(coun
tMatrix),subsampleN[j])])
       }
       colnames(current_countMatrix)=subsampleN
       rownames(current_countMatrix)=rawcount[,1]
       libSize <- colSums(current_countMatrix)
       geneSymbol=rawcount[,2]
```

```
        cat("Processing for the design\n")
        cat(paste("Name folder: ",name_folder,"\n",sep=""))
        cat(paste("Design : ",paste(subsampleN, group,sep="=",collapse=" ; "),"\n",spe=""))

         # Perform gene differential expression

         perform_dge(current_countMatrix, group, count_limit, name_folder)
}
```

Appendix 4.2.1.2

Retrieve promoter sequence of each co-regulated genes

Promoter sequence of each of the genes was retrieved using BioPython (Cock, et al., 2009).

- Retrieve promoter sequences of co-regulated genes:

```
from Bio import SeqIO
import csv
prom_dict_18sp = SeqIO.index("18sp_prom_mod.fas", "fasta")
csvReader = csv.reader(open('upreg_genes.csv', 'rb'), delimiter=" ")
csvWriter = csv.writer(open('upreg_genes_prom.txt', 'w'), delimiter="\t")
error_file = open("errors_upreg_genes.txt", 'a')
for line in csvReader:
    try:
        Promoter = str(prom_dict_18sp[line[0]].seq)
        csvWriter.writerow([line[0], Promoter])
    except KeyError:
        #message = "Error with %s, not found in promoters file\n" % line[0]
        message = "%s\n" % line[0]
        error_file.write(message)
```

Appendix 4.2.2

De novo motif discovery in promoters of co-regulated genes using Seeder

Motifs in promoters of co-regulated genes were discovered using Seeder (Fauteux *et al.,* 2008).

❖ Seeder *finder* script:

```
use Seeder::Finder;
my $finder = Seeder::Finder->new(
seed_width => "6",
strand => "revcom",
motif_width => "12",
n_motif => "10",
hd_index_file => "6.index",
seq_file => "input_seqfile",
bkgd_file => "glyma_prom_1K.bkgd",
out_file => "outputfile.finder",
);
$finder->find_motifs;
```

Appendix 4.2.3

De novo motif discovery in promoters of co-regulated genes using MEME

Motifs in promoters of co-regulated genes were discovered using MEME (Bailey *et al.*, 2009).

❖ Unix script to run MEME:

```
for file in clark_epi_prom.txt;
do meme $file -text -dna -mod anr -nmotifs 10 -w 6 -revcomp -maxsites 250 –maxsize
1000000 -bfile glyma.MEMEbkg > $file.MEME;
done
```

Appendix 4.3.1

Build MySQL relational database for genome-promoter exploration in 18 plant species

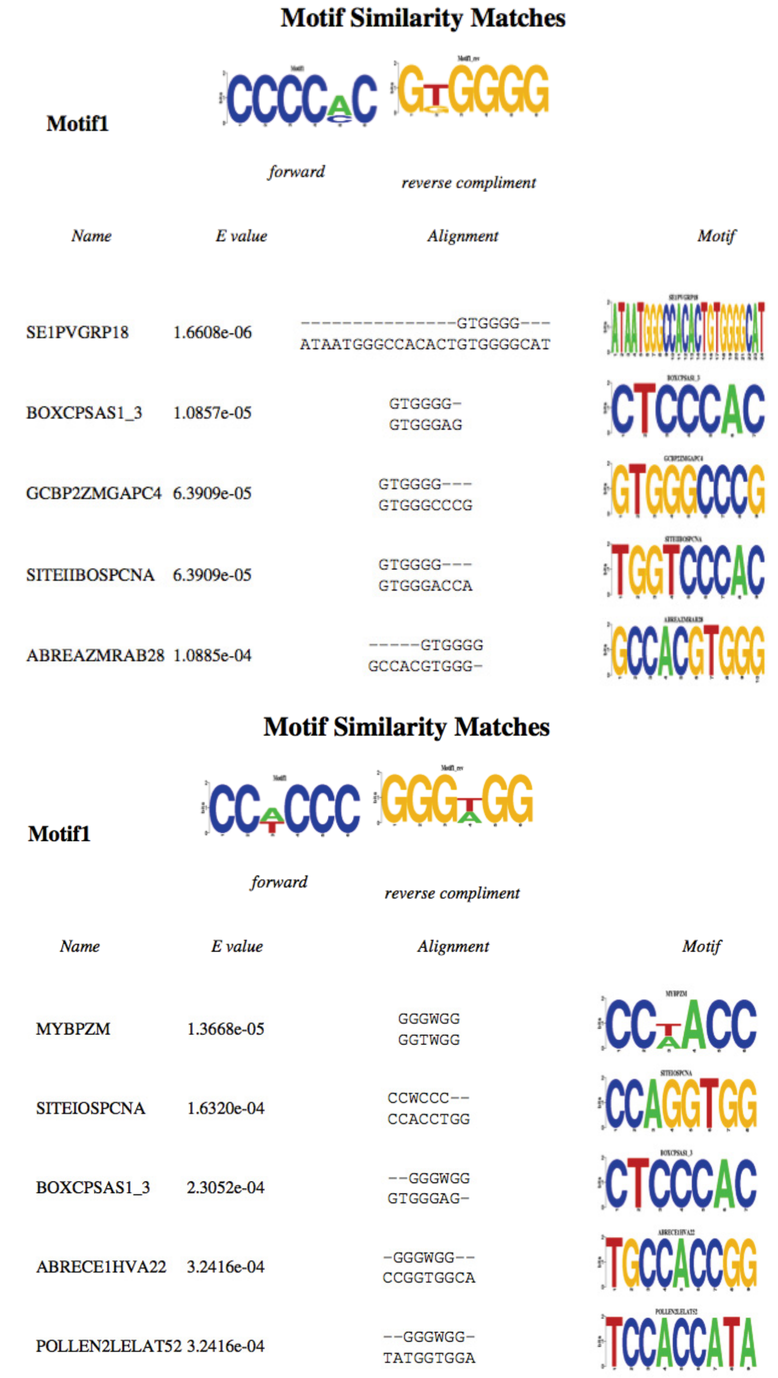
The relational database called Promologous contains pre-computed whole-genome comparative data of promoter sequences across 18 plant species, as well as other genome information associated with them.

❖ Create MySQL relational database:

```
CREATE DATABASE PROMOLOGOUS;

USE PROMOLOGOUS;

CREATE TABLE cog_info (cog_id varchar(20), version varchar(20), species_abr
varchar(10), gene_id varchar(50)) ENGINE=InnoDB;

CREATE TABLE gene_info (gene_id varchar(50), transcript_id varchar(50),
chrom_scaff varchar(100), strand varchar(10), gene_start int, gene_end int)
ENGINE=InnoDB;

CREATE TABLE seeder_motif (cog_id varchar(20), seeder_seq varchar (20), q_value
varchar(30), motif_ref varchar(100), desc1 varchar(100), desc2 varchar(100), desc3
varchar(100)) ENGINE=InnoDB;

CREATE TABLE go_info (gene_id varchar(50), transcript_id varchar(50), go_id
varchar(50), go_description varchar(2000)) ENGINE=InnoDB;

CREATE TABLE place_info (unknown varchar(50), motif_name varchar(50), place_seq
varchar(100), motif_length varchar(50), place_id varchar (50)) ENGINE=InnoDB;

LOAD DATA LOCAL INFILE ‘cog_info_tab.txt’ INTO TABLE cog_info;
LOAD DATA LOCAL INFILE ‘seeder_motif_tab.txt’ INTO TABLE seeder_motif;
LOAD DATA LOCAL INFILE ‘gene_info_tab.txt’ INTO TABLE gene_info;
LOAD DATA LOCAL INFILE ‘go_info_tab.txt’ INTO TABLE go_info;
LOAD DATA LOCAL INFILE ‘place_info_tab.txt’ INTO TABLE place_info;
```

Supplementary Figures

Supplementary Figure 6.2.2.1

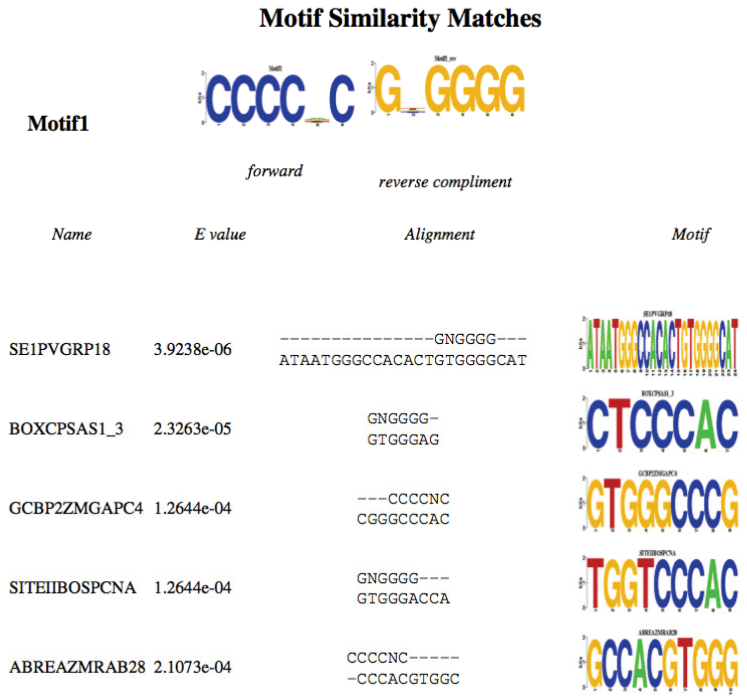


Name	*E value*	*Alignment*
SE1PVGRP18	1.6608e-06	---------------GTGGGG--- ATAATGGGCCACACTGTGGGGCAT
BOXCPSAS1_3	1.0857e-05	GTGGGG- GTGGGAG
GCBP2ZMGAPC4	6.3909e-05	GTGGGG--- GTGGGCCCG
SITEIIBOSPCNA	6.3909e-05	GTGGGG--- GTGGGACCA
ABREAZMRAB28	1.0885e-04	-----GTGGGG GCCACGTGGG-

Name	*E value*	*Alignment*
MYBPZM	1.3668e-05	GGGWGG GGTWGG
SITEIOSPCNA	1.6320e-04	CCWCCC-- CCACCTGG
BOXCPSAS1_3	2.3052e-04	--GGGWGG GTGGGAG-
ABRECE1HVA22	3.2416e-04	-GGGWGG-- CCGGTGGCA
POLLEN2LELAT52	3.2416e-04	--GGGWGG- TATGGTGGA

Sequence logos generated by STAMP (Mahony and Benos, 2007) showing significant motifs enriched in gene promoters of soybean cv. Clark standard in shoot epidermis tissue. Motifs were detected by MEME (Bailey *et al*., 2009) and matched to 9 known motifs in PLACE database (Higo *et al*., 1999).

Supplementary Figure 6.2.2.2



Motif Similarity Matches

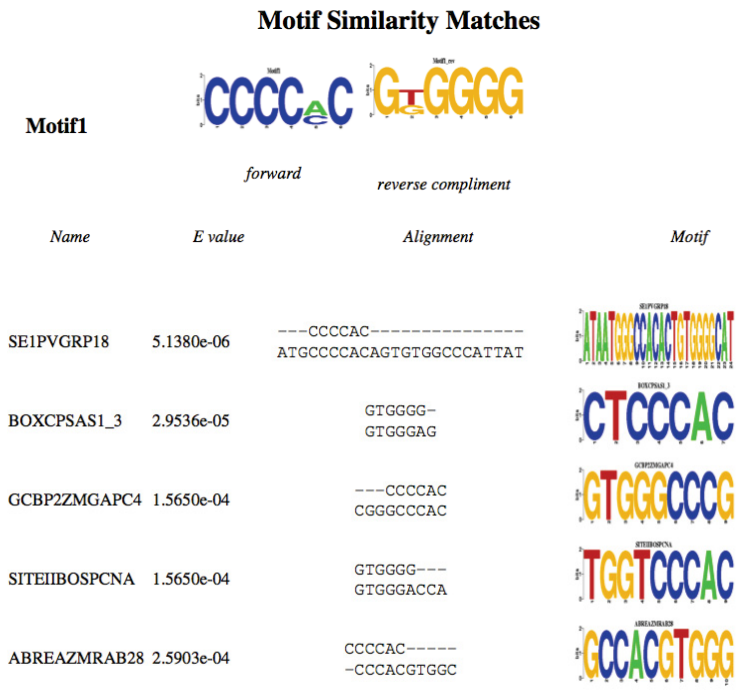
Motif1

forward *reverse compliment*

Name	*E value*	*Alignment*	*Motif*
SE1PVGRP18	3.9238e-06	---------------GNGGGG--- ATAATGGGCCACACTGTGGGGCAT	ATAATGGGCCACACTGTGGGGCAT
BOXCPSAS1_3	2.3263e-05	GNGGGG- GTGGGAG	CTCCCAC
GCBP2ZMGAPC4	1.2644e-04	---CCCCNC CGGGCCCAC	GTGGGCCCG
SITEIIBOSPCNA	1.2644e-04	GNGGGG--- GTGGGACCA	TGGTCCCAC
ABREAZMRAB28	2.1073e-04	CCCCNC----- -CCCACGTGGC	GCCACGTGGG

Sequence logos generated by STAMP (Mahony and Benos, 2007) showing significant motifs enriched in gene promoters of soybean 5-Leaflet mutants in shoot epidermis tissue. Motifs were detected by MEME (Bailey *et al*., 2009) and matched to five known motifs in PLACE database (Higo *et al.,* 1999).

Supplementary Figure 6.2.2.3



Motif Similarity Matches

Motif1

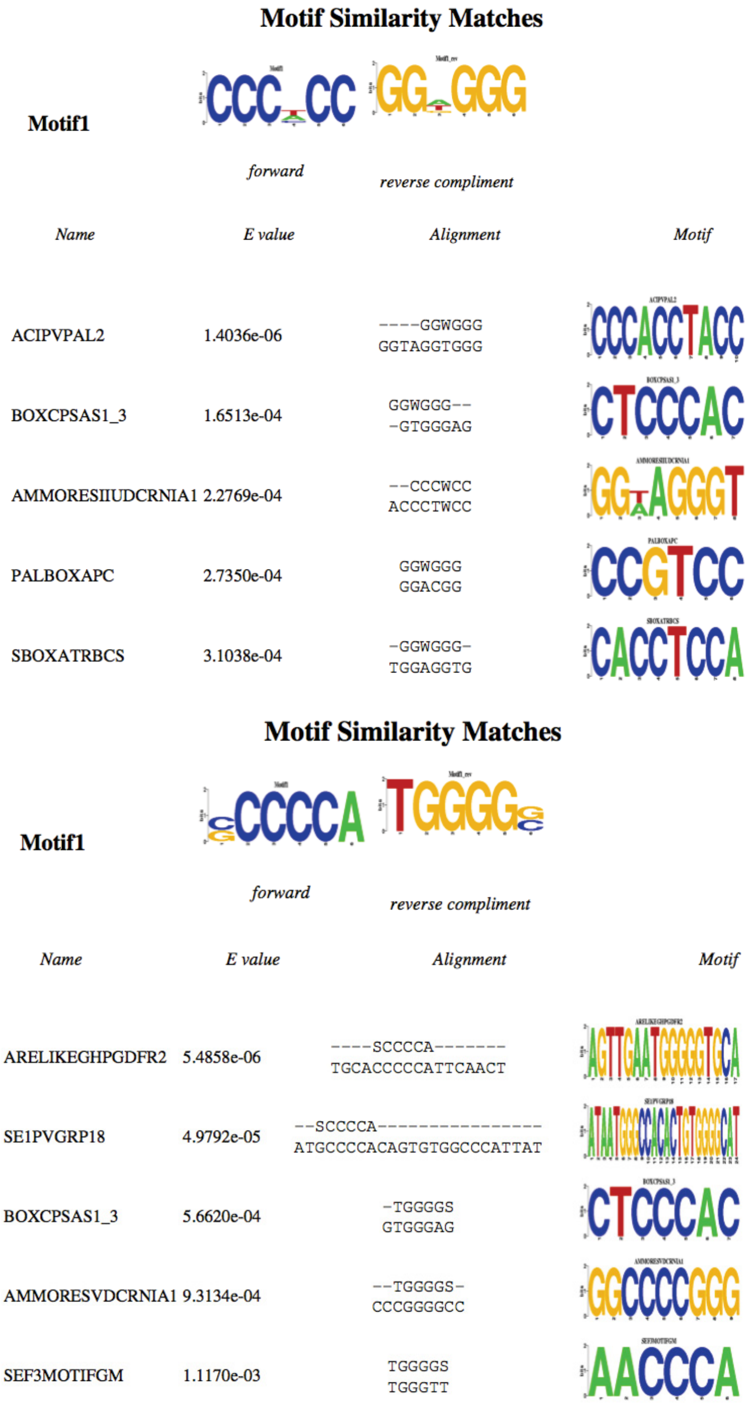
forward

reverse compliment

Name	*E value*	*Alignment*	*Motif*
SE1PVGRP18	5.1380e-06	`---CCCCAC---------------` `ATGCCCCACAGTGTGGCCCATTAT`	ATAATGGGCCACACTGTGGGGCAT
BOXCPSAS1_3	2.9536e-05	`GTGGGG-` `GTGGGAG`	CTCCCAC
GCBP2ZMGAPC4	1.5650e-04	`---CCCCAC` `CGGGCCCAC`	GTGGGCCCG
SITEIIBOSPCNA	1.5650e-04	`GTGGGG---` `GTGGGACCA`	TGGTCCCAC
ABREAZMRAB28	2.5903e-04	`CCCCAC-----` `-CCCACGTGGC`	GCCACGTGGG

Sequence logos generated by STAMP (Mahony and Benos, 2007) showing significant motifs enriched in gene promoters of soybean Glabrous mutants in shoot epidermis tissue. Motifs were detected by MEME (Bailey *et al*., 2009) and matched to five known motifs in PLACE database (Higo *et al*., 1999).

Supplementary Figure 6.2.2.4



Motif Similarity Matches

Motif1

forward *reverse compliment*

Name	*E value*	*Alignment*	*Motif*
ACIPVPAL2	1.4036e-06	----GGWGGG GGTAGGTGGG	CCCACCTACC
BOXCPSAS1_3	1.6513e-04	GGWGGG-- -GTGGGAG	CTCCCAC
AMMORESIIUDCRNIA1	2.2769e-04	--CCCWCC ACCCTWCC	GGTAGGGT
PALBOXAPC	2.7350e-04	GGWGGG GGACGG	CCGTCC
SBOXATRBCS	3.1038e-04	-GGWGGG- TGGAGGTG	CACCTCCA

Motif Similarity Matches

Motif1

forward *reverse compliment*

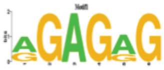
Name	*E value*	*Alignment*	*Motif*
ARELIKEGHPGDFR2	5.4858e-06	----SCCCCA------- TGCACCCCCATTCAACT	AGTTGAATGGGGGTGCA
SE1PVGRP18	4.9792e-05	--SCCCCA---------------- ATGCCCCACAGTGTGGCCCATTAT	ATAATGGGCCACACTGTGGGGCAT
BOXCPSAS1_3	5.6620e-04	-TGGGGS GTGGGAG	CTCCCAC
AMMORESVDCRNIA1	9.3134e-04	--TGGGGS- CCCGGGGCC	GGCCCCGGG
SEF3MOTIFGM	1.1170e-03	TGGGGS TGGGTT	AACCCA

Sequence logos generated by STAMP (Mahony and Benos, 2007) showing significant motifs enriched in gene promoters of soybean cv. Clark standard in shoot apical meristem tissue. Motifs were detected by MEME (Bailey *et al.,* 2009) and matched to 9 known motifs in PLACE database (Higo *et al.,* 1999).

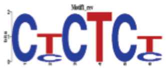
Supplementary Figure 6.2.2.5

Motif Similarity Matches

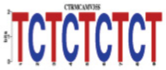
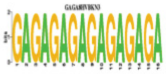
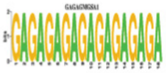
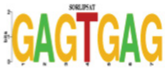
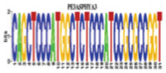
Motif1



forward

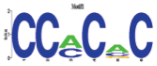


reverse compliment

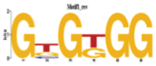
Name	*E value*	*Alignment*	*Motif*
CTRMCAMV35S	1.1765e-07	-CTCTCT-- TCTCTCTCT	
GAGA8HVBKN3	1.9337e-06	-CTCTCT--------- TCTCTCTCTCTCTCTC	
GAGAGMGSA1	2.8132e-06	-CTCTCT----------- TCTCTCTCTCTCTCTCTC	
SORLIP5AT	3.4401e-04	-AGAGAG GAGTGAG	
PE3ASPHYA3	3.8756e-04	------------CTCTCT------------- CAGCTCCCATGGCTCTCCCATCCGCGCCGGT	

Motif Similarity Matches

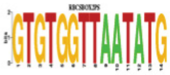
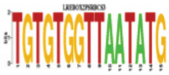
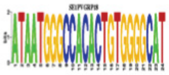
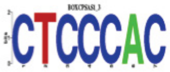
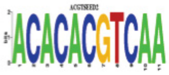
Motif1



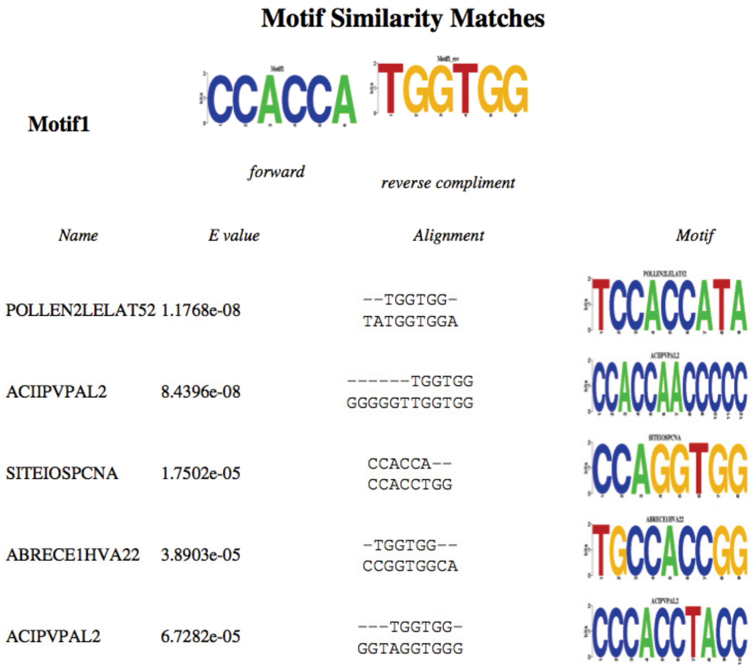
forward



reverse compliment

Name	*E value*	*Alignment*	*Motif*
RBCSBOX2PS	3.6447e-06	--------CCMCMC CATATTAACCACAC	
LREBOX2PSRBCS3	5.1960e-06	--------CCMCMC- CATATTAACCACACA	
SE1PVGRP18	1.9674e-05	----------GKGKGG-------- ATGCCCCACAGTGTGGCCCATTAT	
BOXCPSAS1_3	6.5420e-04	GKGKGG- GTGGGAG	
ACGTSEED2	1.0823e-03	-----GKGKGG TTGACGTGTGT	

Supplementary Figure 6.2.2.5 continued



Motif Similarity Matches

Motif1

forward

reverse compliment

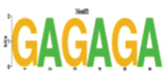
Name	*E value*	*Alignment*	*Motif*
POLLEN2LELAT52	1.1768e-08	`--TGGTGG-` `TATGGTGGA`	TCCACCATA
ACIIPVPAL2	8.4396e-08	`------TGGTGG` `GGGGGTTGGTGG`	CCACCAACCCCC
SITEIOSPCNA	1.7502e-05	`CCACCA--` `CCACCTGG`	CCAGGTGG
ABRECE1HVA22	3.8903e-05	`-TGGTGG--` `CCGGTGGCA`	TGCCACCGG
ACIPVPAL2	6.7282e-05	`---TGGTGG-` `GGTAGGTGGG`	CCCACCTACC

Sequence logos generated by STAMP (Mahony and Benos, 2007) showing significant motifs enriched in gene promoters of soybean 5-Leaflet mutants in shoot apical meristem tissue. Motifs were detected by MEME (Bailey *et al*., 2009) and matched to 15 known motifs in PLACE database (Higo *et al*., 1999).

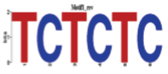
Supplementary Figure 6.2.2.6

Motif Similarity Matches

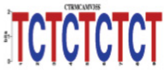
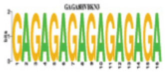
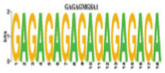
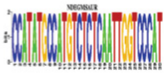
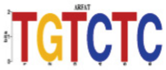
Motif1



forward

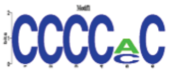


reverse compliment

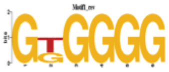
Name	*E value*	*Alignment*	*Motif*
CTRMCAMV35S	1.1768e-08	-GAGAGA-- AGAGAGAGA	
GAGA8HVBKN3	2.3090e-07	TCTCTC---------- TCTCTCTCTCTCTCTC	
GAGAGMGSA1	3.3845e-07	TCTCTC------------ TCTCTCTCTCTCTCTCTC	
NDEGMSAUR	8.9045e-07	------------GAGAGA------------ ATGGGACCAATTGAGAGACATGGCATATGG	
ARFAT	1.2971e-05	TCTCTC TGTCTC	

Motif Similarity Matches

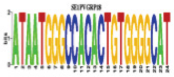
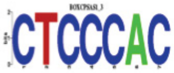
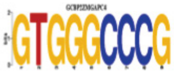
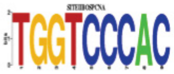
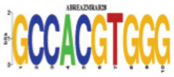
Motif1



forward



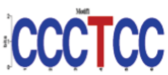
reverse compliment

Name	*E value*	*Alignment*	*Motif*
SE1PVGRP18	6.7289e-06	---------------GKGGGG--- ATAATGGGCCACACTGTGGGGCAT	
BOXCPSAS1_3	3.7501e-05	GKGGGG- GTGGGAG	
GCBP2ZMGAPC4	1.9368e-04	---CCCCMC CGGGCCCAC	
SITEIIBOSPCNA	1.9368e-04	GKGGGG--- GTGGGACCA	
ABREAZMRAB28	3.1832e-04	CCCCMC----- -CCCACGTGGC	

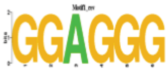
Supplementary Figure 6.2.2.6 continued

Motif Similarity Matches

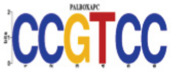
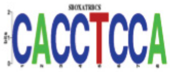
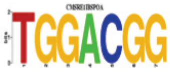
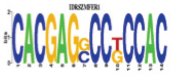
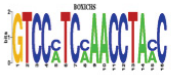
Motif1



forward

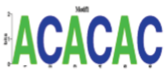


reverse compliment

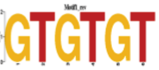
Name	*E value*	*Alignment*	*Motif*
PALBOXAPC	1.2971e-05	GGAGGG GGACGG	
SBOXATRBCS	1.7502e-05	-GGAGGG- TGGAGGTG	
CMSRE1IBSPOA	6.7873e-05	CCCTCC- CCGTCCA	
IDRSZMFER1	9.7372e-05	--GGAGGG------ GTGGMGGSCTCGTG	
BOXICHS	1.3989e-04	--CCCTCC-------- GTCCMTCMAACCTAMC	

Motif Similarity Matches

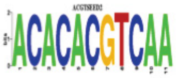
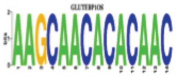
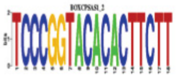
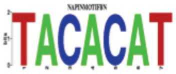
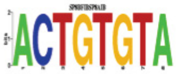
Motif1



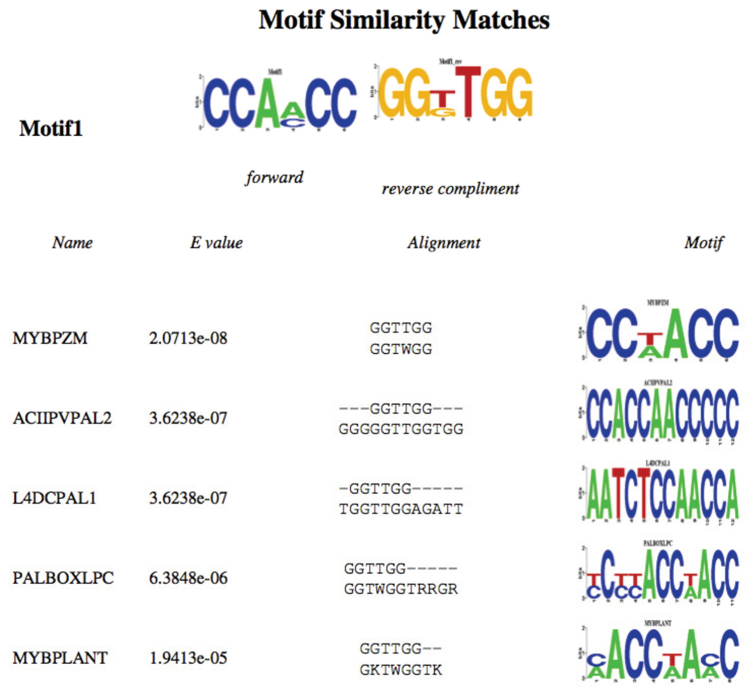
forward



reverse compliment

Name	*E value*	*Alignment*	*Motif*
ACGTSEED2	4.9110e-08	-----GTGTGT TTGACGTGTGT	
GLUTEBP1OS	1.4850e-07	---GTGTGT----- GTTGTGTGTTGCTT	
BOXCPSAS1_2	3.3845e-07	-----GTGTGT------- AAGAAGTGTGTACCGGGA	
NAPINMOTIFBN	6.2435e-06	GTGTGT- ATGTGTA	
SP8BFIBSP8AIB	1.7502e-05	-ACACAC- TACACAGT	

Supplementary Figure 6.2.2.6 continued



Motif Similarity Matches

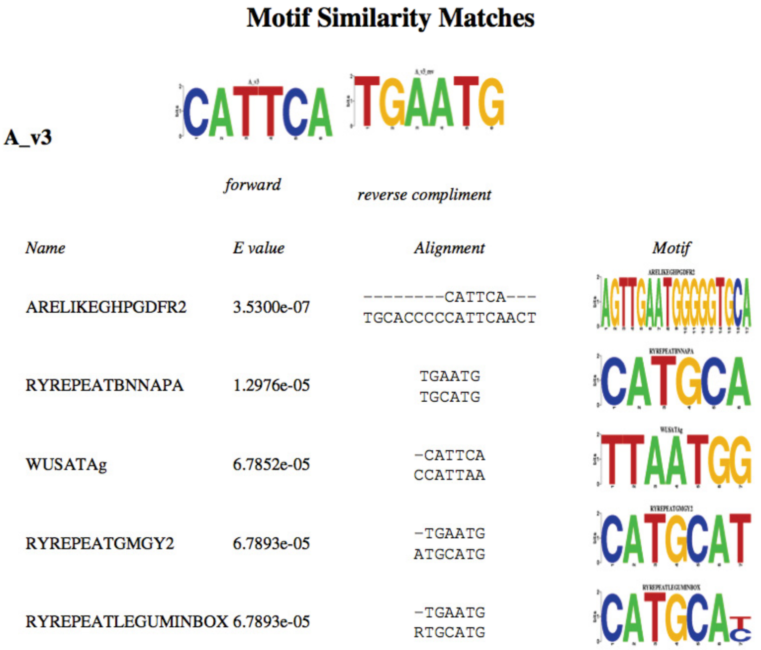
Motif1

forward *reverse compliment*

Name	*E value*	*Alignment*	*Motif*
MYBPZM	2.0713e-08	GGTTGG GGTWGG	
ACIIPVPAL2	3.6238e-07	---GGTTGG--- GGGGGTTGGTGG	
L4DCPAL1	3.6238e-07	-GGTTGG----- TGGTTGGAGATT	
PALBOXLPC	6.3848e-06	GGTTGG----- GGTWGGTRRGR	
MYBPLANT	1.9413e-05	GGTTGG-- GKTWGGTK	

Sequence logos generated by STAMP (Mahony and Benos, 2007) showing significant motifs enriched in gene promoters of soybean Glabrous mutants in shoot apical meristem tissue. Motifs were detected by MEME (Bailey *et al*., 2009) and matched to 24 known motifs in PLACE database (Higo *et al*., 1999).

Supplementary Figure 6.2.3.1



Motif Similarity Matches

A_v3

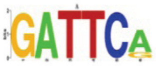
Name	E value	Alignment	Motif
ARELIKEGHPGDFR2	3.5300e-07	--------CATTCA--- TGCACCCCCATTCAACT	AGTTGAATGGGGGTGCA
RYREPEATBNNAPA	1.2976e-05	TGAATG TGCATG	CATGCA
WUSATAg	6.7852e-05	-CATTCA CCATTAA	TTAATGG
RYREPEATGMGY2	6.7893e-05	-TGAATG ATGCATG	CATGCAT
RYREPEATLEGUMINBOX	6.7893e-05	-TGAATG RTGCATG	CATGCA

Sequence logos generated by STAMP (Mahony and Benos, 2007) showing significant motifs enriched in gene promoters of soybean cv. Clark standard in shoot epidermis tissue. Motifs were detected by BioProspector software (Liu *et al*., 2001) and matched to five known motifs in PLACE database (Higo *et al*., 1999).

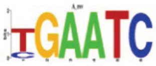
Supplementary Figure 6.2.3.2

Motif Similarity Matches

A



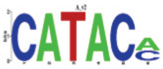
forward



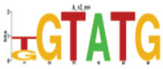
reverse compliment

Name	*E value*	*Alignment*	*Motif*
EIN3ATERF1	1.2360e-06	-GATTCA--------------------- GGATTCAAGATACATGCCCCTTGAATCC	
GCN4OSGLUB1	8.7819e-05	-GATTCA TGACTCA	
L1DCPAL1	2.0352e-04	GATTCA------- -ATTCACCTACCC	
AGMOTIFNTMYB2	2.1088e-04	-TGAATC- TTGGATCT	
ARR1AT	3.5358e-04	-GATTCA NGATT--	

A_v2



forward

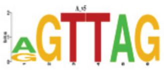


reverse compliment

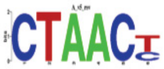
Name	*E value*	*Alignment*	*Motif*
S2FSORPL21	2.1938e-07	--KGTATG- AATGTATGG	
SORLIP4AT	3.8903e-05	KGTATG---- -GTATGATGG	
RYREPEATBNNAPA	1.3412e-04	KGTATG TGCATG	
JASE2ATOPR1	2.0428e-04	-------KGTATG TTGACGACGTATG	
SORLREP3AT	3.1470e-04	---CATACM ATATATACA	

Supplementary Figure 6.2.3.2 continued

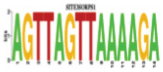
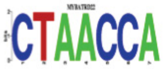
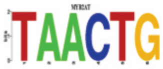
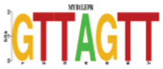
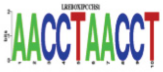
A_v5



forward

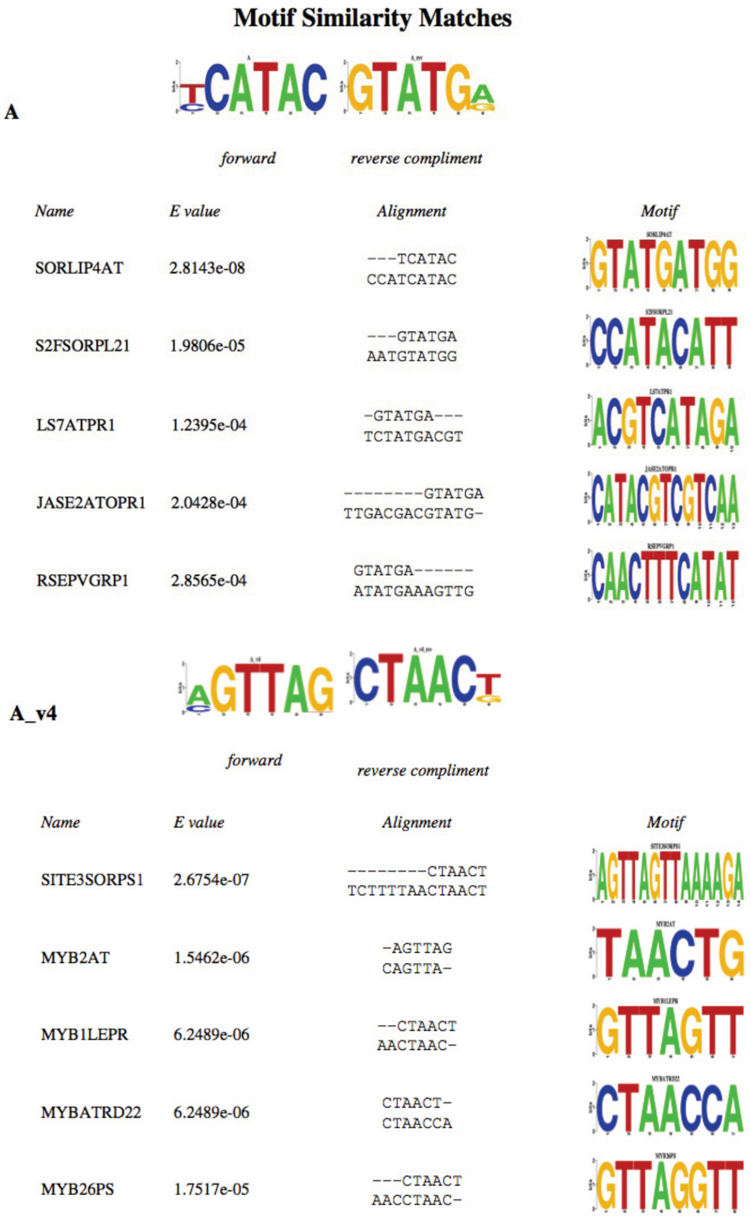


reverse compliment

Name	*E value*	*Alignment*	*Motif*
SITE3SORPS1	6.0270e-07	AGTTAG-------- AGTTAGTTAAAAGA	
MYBATRD22	9.5234e-07	-AGTTAG TGGTTAG	
MYB2AT	3.4085e-06	-AGTTAG CAGTTA-	
MYB1LEPR	6.2435e-06	AGTTAG-- -GTTAGTT	
LREBOXIPCCHS1	1.3000e-05	-AGTTAG--- AGGTTAGGTT	

Sequence logos generated by STAMP (Mahony and Benos, 2007) showing significant motifs enriched in gene promoters of soybean 5-Leaflet mutants in shoot epidermis tissue. Motifs were detected by BioProspector software (Liu *et al*., 2001) and matched to 15 known motifs in PLACE database (Higo *et al*., 1999).

Supplementary Figure 6.2.3.3



Motif Similarity Matches

A

Name	E value	Alignment	Motif
SORLIP4AT	2.8143e-08	---TCATAC CCATCATAC	GTATGATGG
S2FSORPL21	1.9806e-05	---GTATGA AATGTATGG	CCATACATT
LS7ATPR1	1.2395e-04	-GTATGA--- TCTATGACGT	ACGTCATAGA
JASE2ATOPR1	2.0428e-04	--------GTATGA TTGACGACGTATG-	CATACGTCGTCAA
RSEPVGRP1	2.8565e-04	GTATGA------ ATATGAAAGTTG	CAACTTTCATAT

A_v4

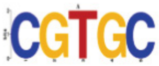
Name	E value	Alignment	Motif
SITE3SORPS1	2.6754e-07	--------CTAACT TCTTTTAACTAACT	AGTTAGTTAAAAGA
MYB2AT	1.5462e-06	-AGTTAG CAGTTA-	TAACTG
MYB1LEPR	6.2489e-06	--CTAACT AACTAAC-	GTTAGTT
MYBATRD22	6.2489e-06	CTAACT- CTAACCA	CTAACCA
MYB26PS	1.7517e-05	---CTAACT AACCTAAC-	GTTAGGTT

Sequence logos generated by STAMP (Mahony and Benos, 2007) showing significant motifs enriched in gene promoters of soybean Glabrous mutants in shoot epidermis tissue. Motifs were detected by BioProspector software (Liu *et al.,* 2001) and matched to 10 known motifs in PLACE database (Higo *et al.,* 1999).

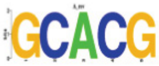
Supplementary Figure 6.2.3.4

Motif Similarity Matches

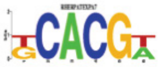
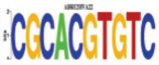
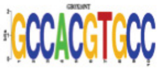
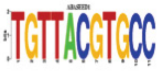
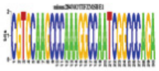
A



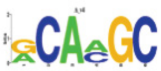
forward



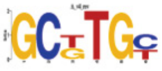
reverse compliment

Name	*E value*	*Alignment*	*Motif*
RHERPATEXPA7	5.0501e-06	-CGTGC WCGTGM	
ABRE2HVA22	5.8084e-06	----CGTGC- GACACGTGCG	
GBOX10NT	5.8084e-06	-GCACG---- GGCACGTGGC	
ABASEED1	9.4711e-06	-GCACG----- GGCACGTAACA	
minus284MOTIFZMSBE1	8.3614e-05	-------------------------GCACG TCTGGGCCGATTGGCCTTTGGGCTTGCACG	

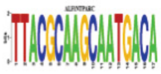
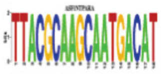
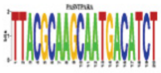
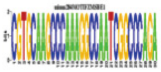
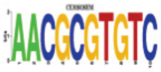
A_v4



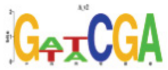
forward



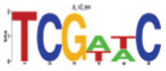
reverse compliment

Name	*E value*	*Alignment*	*Motif*
ALF1NTPARC	2.9440e-05	-------GCKTGC---- TGTCATTGCTTGCGTAA	
ASF1NTPARA	2.9746e-05	--------GCKTGC---- ATGTCATTGCTTGCGTAA	
PASNTPARA	4.8634e-05	----------GCKTGC---- AGATGTCATTGCTTGCGTAA	
minus284MOTIFZMSBE1	7.1435e-05	---------------------GCKTGC--- TCTGGGCCGATTGGCCTTTGGGCTTGCACG	
CE3OSOSEM	2.1253e-04	-GCAMGC--- GACACGCGTT	

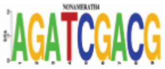
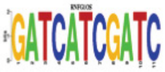
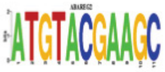
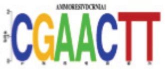
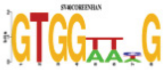
Supplementary Figure 6.2.3.4 continued



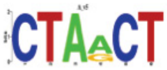
forward



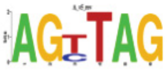
reverse compliment

Name	*E value*	*Alignment*	*Motif*
NONAMERATH4	2.4413e-06	-GWWCGA-- AGATCGACG	
RNFG1OS	7.9322e-06	GWWCGA----- GATCGATGATC	
ABAREG2	2.1965e-04	---TCGWWC-- GCTTCGTACAT	
AMMORESIVDCRNIA1	5.8749e-04	--GWWCGA AAGTTCG-	
SV40COREENHAN	1.4368e-03	TCGWWC--- -CNWWCCAC	

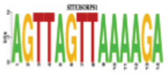
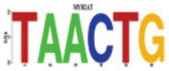
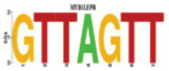
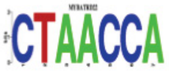
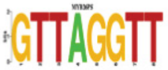
A_v5



forward



reverse compliment

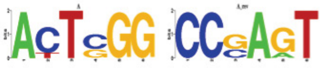
Name	*E value*	*Alignment*	*Motif*
SITE3SORPS1	3.7780e-07	--------CTAACT TCTTTTAACTAACT	
MYB2AT	2.1653e-06	-AGTTAG CAGTTA-	
MYB1LEPR	1.3968e-05	--CTAACT AACTAAC-	
MYBATRD22	1.3968e-05	-AGTTAG TGGTTAG	
MYB26PS	3.7407e-05	---CTAACT AACCTAAC-	

Sequence logos generated by STAMP (Mahony and Benos, 2007) showing significant motifs enriched in gene promoters of soybean cv. Clark standard in shoot apical meristem tissue. Motifs were detected by BioProspector software (Liu *et al*., 2001) and matched to 19 known motifs in PLACE database (Higo *et al*., 1999).

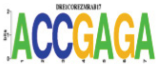
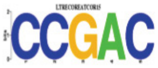
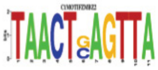
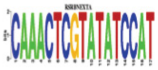
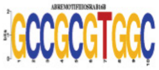
Supplementary Figure 6.2.3.5

Motif Similarity Matches

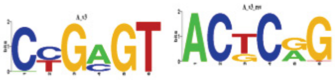
A



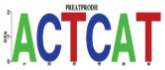
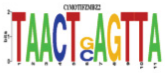
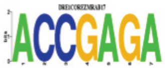
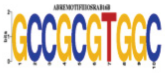
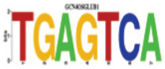
forward *reverse compliment*

Name	*E value*	*Alignment*	*Motif*
DRE1COREZMRAB17	2.8198e-05	ACTCGG- TCTCGGT	
LTRECOREATCOR15	1.1056e-03	ACTCGG -GTCGG	
C1MOTIFZMBZ2	1.2473e-03	---CCGAGT-- TAACTSAGTTA	
RSRBNEXTA	1.5839e-03	--------CCGAGT--- ATGGATATACGAGTTTG	
ABREMOTIFIIIOSRAB16B	1.6542e-03	---ACTCGG- GCCACGCGGC	

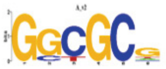
A_v3

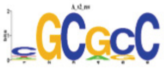


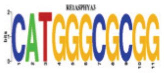
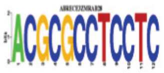
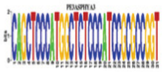
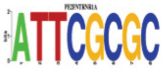
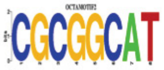
forward *reverse compliment*

Name	*E value*	*Alignment*	*Motif*
PREATPRODH	8.8413e-05	CYGAGT ATGAGT	
C1MOTIFZMBZ2	1.2465e-04	---CYGAGT-- TAACTSAGTTA	
DRE1COREZMRAB17	1.2928e-04	ACTCRG- TCTCGGT	
ABREMOTIFIIIOSRAB16B	3.4043e-04	---ACTCRG- GCCACGCGGC	
GCN4OSGLUB1	3.6714e-04	CYGAGT-- -TGAGTCA	

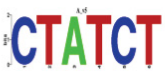
Supplementary Figure 6.2.3.5 continued

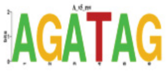


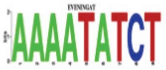
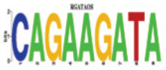
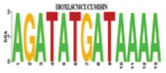
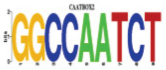
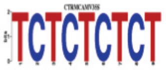


	forward	*reverse compliment*	
Name	*E value*	*Alignment*	*Motif*
RE1ASPHYA3	6.8776e-08	`-CGCGCC----` `CCGCGCCCATG`	
ABRECE3ZMRAB28	1.1750e-07	`-CGCGCC-----` `ACGCGCCTCCTC`	
PE3ASPHYA3	1.2267e-06	`----------------------CGCGCC---` `CAGCTCCCATGGCTCTCCCATCCGCGCCGGT`	
PE2FNTRNR1A	2.2890e-05	`GGCGCG---` `-GCGCGAAT`	
OCTAMOTIF2	5.2820e-05	`--GGCGCG` `ATGCCGCG`	

A_v5

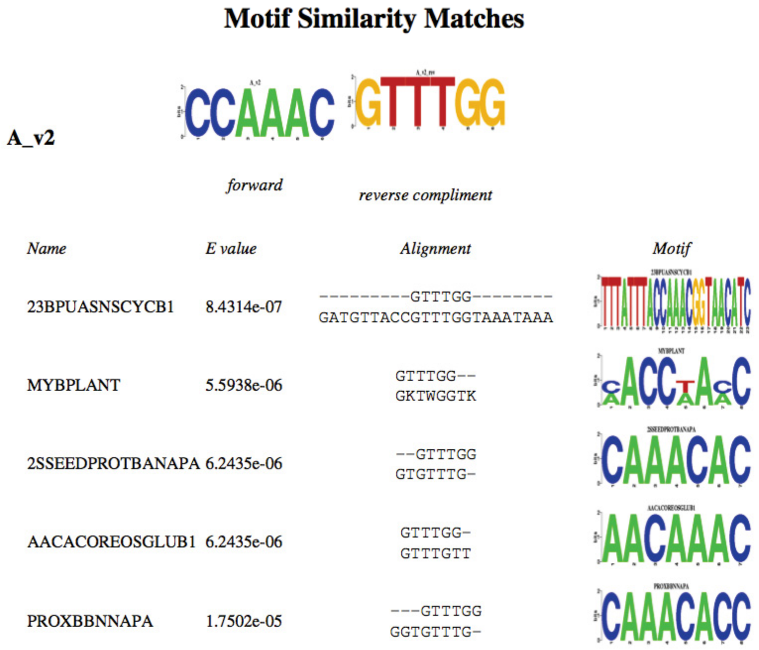




	forward	*reverse compliment*	
Name	*E value*	*Alignment*	*Motif*
EVENINGAT	3.8903e-05	`AGATAG---` `AGATATTTT`	
RGATAOS	3.8903e-05	`CTATCT----` `-TATCTTCTG`	
IBOXLSCMCUCUMISIN	2.0428e-04	`-------CTATCT` `TTTTATCATATCT`	
CAATBOX2	3.2792e-04	`AGATAG---` `AGATTGGCC`	
CTRMCAMV35S	3.2897e-04	`AGATAG---` `AGAGAGAGA`	

Sequence logos generated by STAMP (Mahony and Benos, 2007) showing significant motifs enriched in gene promoters of soybean 5-Leaflet mutants in shoot apical meristem tissue. Motifs were detected by BioProspector software (Liu *et al*., 2001) and matched to 18 known motifs in PLACE database (Higo *et al*., 1999).

Supplementary Figure 6.2.3.6



Motif Similarity Matches

A_v2

Name	*E value*	*Alignment*	*Motif*
23BPUASNSCYCB1	8.4314e-07	`---------GTTTGG--------` `GATGTTACCGTTTGGTAAATAAA`	
MYBPLANT	5.5938e-06	`GTTTGG--` `GKTWGGTK`	
2SSEEDPROTBANAPA	6.2435e-06	`--GTTTGG` `GTGTTTG-`	
AACACOREOSGLUB1	6.2435e-06	`GTTTGG-` `GTTTGTT`	
PROXBBNNAPA	1.7502e-05	`---GTTTGG` `GGTGTTTG-`	

Sequence logos generated by STAMP (Mahony and Benos, 2007) showing significant motifs enriched in gene promoters of soybean Glabrous mutants shoot apical meristem tissue. Motifs were detected by BioProspector software (Liu *et al*., 2001) and matched to five known motifs in PLACE database (Higo *et al*., 1999).