Computational DNA motif discovery in plant promoters

François Fauteux

Doctor of Philosophy

Department of Plant Science

Faculty of Agricultural and Environmental Sciences

McGill University

Montreal, Quebec, Canada

January 2010

A thesis submitted to McGill University in partial fulfillments of the

requirements of the degree of Doctor of Philosophy

Abstract

The regulation of gene expression is driven primarily by transcription factors binding to short DNA sequences. Here three studies related to promoter cis-regulatory motif discovery in plant promoters are presented. In the first study, an exact discriminative seeding DNA motif discovery addressing key issues associated with popular DNA motif discovery algorithms is proposed. The Seeder algorithm outperforms popular motif discovery tools on biological benchmark data. In the second study, the algorithm is applied to the identification of *cis*-regulatory motifs in seed storage protein gene promoters. Known and new motifs are discovered. In the third study, groups of orthologous genes are identified among five dicotyledonous plant species, and DNA motif discovery is carried out in the proximal promoter sequence within each group. The presence of three large clusters of groups of orthologous promoters sharing similar motifs is revealed.

Résumé

L’expression des gènes est régulée, en grande partie, par la liaison des facteurs de transcription à de courtes séquences d’ADN. Trois études sont présentées, portant sur l’identification *in silico* de motifs régulateurs dans les séquences promotrices de gènes végétaux. Dans la première étude, un algorithme d’initiation discriminative exacte est présenté. L'algorithme surpasse plusieurs algorithmes populaires lorsque appliqué à des données biologiques de référence. Dans la deuxième étude, l'algorithme est utilisé pour l'identification de motifs *cis*-régulateurs conservés dans les promoteurs de gènes de protéines de réserve des graines chez diverses espèces végétales. Des motifs connus ainsi que de nouveaux motifs sont identifiés. Dans la troisième étude, des groupes de gènes orthologues sont identifiés chez cinq espèces dicotylédones, et une recherche de motifs *cis*-régulateurs est réalisée dans les séquences promotrices proximales pour chaque groupe. La présence de trois larges grappes de groupes d'orthologues partageant des motifs similaires est mise en évidence.

Acknowledgements

I first want to thank my supervisor Dr Martina Strömvik who has been a great mentor. Her interest in complex aspects of plant biology has inspired me throughout my research projects. She also knew how to keep me grounded and focused.

I also express my gratitude to Dr Mathieu Blanchette who has been an essential contributor, especially in the rigorous mathematical and statistical aspects of motif discovery. I also thank Dr Pierre Dutilleul for participating on my advisory committee and providing constructive criticism.

This experience has been enjoyable thanks to my lab mates Hanaa Saeed, Julie Livingstone and Annie Archambault, our lab technician Christine Ide and our systems administrator André Lessard. I also want to thank the secretaries of the Department of Plant Science, Roslyn James and Carolyn Bowes.

I thank many people for useful discussions, and ought to mention the invaluable participation of experts in online forums and communities. Among others, the Perl Monks community has been an inestimable

resource. I also thank many software developers for making their work available for the research community.

I acknowledge the National Science and Engineering Research Council of Canada for providing funding through a post-graduate scholarship, and the McGill Recruitment Excellence Fellowship and the Frederick Dimmock Memorial Fellowship for providing additional funding through excellence fellowships.

I am grateful to my family and friends, and most importantly my partner Julie, for encouragements and support giving me the motivation to go forward and always do my best.

Table of contents

List of tables

List of figures

List of abbreviations

ABA, abscisic acid

ABI3, ABA INSENSITIVE3

ABRE, ABA-responsive element

anr, any-number of repetitions

AM, algorithm Markov

AR, Algorithm Real

BLAST, Basic Local Alignment Search Tool

bp, base pair(s)

BRE, TFIIB recognition element

bZIP, basic leucine zipper

cDNA, complementary DNA

ChIP-chip, chromatin immunoprecipitation on microarray

CPAN, Comprehensive Perl Archive Network

CRE(s), *cis*-regulatory element(s)

CRM(s), *cis*-regulatory motif(s)

DDOPM, database of dicotyledonous orthologous motifs

DNA, deoxyribonucleic acid

DoOP, database of orthologous promoters

DPE, downstream promoter element

EM, expectation-maximization

EMSA, electrophoretic mobility shift assay

ENCODE, Encyclopedia of DNA elements

EST(s), expressed sequence tag(s)

FUS3, FUSCA3

GLM, GCN4-like

GO, gene ontology

HD(s), Hamming distance(s)

HMM(s), Hidden Markov Model(s)

Inr, Initiator

IUPAC, International Union of Pure and Applied Chemistry

LEC1, LEAFY COTYLEDON1

LEC2, LEAFY COTYLEDON2

MC(s), motifs cluster(s)

MM, Markov model

MR, Model Real

nCC, nucleotide-level Pearsons correlation coefficient

OP(s), group(s) of orthologous promoters

OG(s), group(s) of orthologous genes

oops, one occurrence per sequence

P-box, prolamin-box

PAL, phenylalanine ammonia-lyase

PC(s), ontology-based cluster(s) of promoters

PFM(s), position frequency matrix(ces)

PLACE, plant *cis*-acting regulatory DNA elements (database)

PWM(s), position weight matrix(ces)

RNA, ribonucleic acid

S/MAR(s), scaffold/matrix attachment region(s)

SELEX, systematic evolution of ligands by exponential enrichment

SMD(s), substring minimal distance(s)

SSP(s), seed storage protein(s)

TAIR, The Arabidopsis Information Resource

TFBS(s), transcription factor binding site(s)

TF(s), transcription factor(s)

TSS(s), transcription start site(s)

zoops, zero-or-one occurrence per sequence

Thesis format

This thesis is presented in manuscript-based format. Three manuscripts are included as chapters. Chapter 3, entitled “Seeder: discriminative seeding DNA motif discovery”, has been published in Bioinformatics in 2008. Chapter 4, entitled “Seed storage protein gene promoters contain conserved DNA motifs in Brassicaceae, Fabaceae and Poaceae”, has been published in BMC Plant Biology in 2009. Chapter 4, entitled “Promoters of dicotyledonous orthologous genes involved in fundamental cellular processes are enriched in highly conserved *cis*-regulatory motifs”, will be submitted for publication in 2009. The three manuscripts have been reformatted for thesis consistency.

Contributions of authors

Chapter 3 is co-authored by F. Fauteux, Dr. M. Blanchette and Dr. M. Strömvik. F. Fauteux designed the algorithm presented in the muscript, with the collaboration of Dr. M. Blanchette and under the supervision of Dr. M. Strömvik. F. Fauteux wrote the software implementation (Perl modules) of the algorithm and designed the data structure accelerating computations. F. Fauteux wrote the initial manuscript. All authors have participated in writing the final version of the manuscript.

Chapter 4 is co-authored by F. Fauteux and Dr. M. Strömvik. F. Fauteux and M. Strömvik designed the project. F. Fauteux performed the research under the supervision of Dr. M. Strömvik, and wrote the initial manuscript. Both authors have participated in writing the final version of the manuscript.

Chapter 5 is co-authored by F. Fauteux and Dr. M. Strömvik. F. Fauteux and M. Strömvik designed the project. F. Fauteux performed the research under the supervision of Dr. M. Strömvik, and wrote the initial manuscript. Both authors have participated in writing the final version of the manuscript.

1. Introduction

1.1 General introduction

Plant biotechnology applications, for example the engineering of crop plants for improved traits related to production and quality (Shewry*, et al.*, 2008), the use of plants as bioreactors for the production of heterologous proteins (Streatfield, 2007) and the recent engineering of plants for biofuel production (Sticklen, 2008), are increasingly important for different sectors of the economy. *Cis*-regulatory elements (CREs) are short, noncoding deoxyribonucleic acid (DNA) sequences required for the correct expression of genes, often containing transcription factor binding sites (TFBSs). Plant genetic engineers need access to a variety of CREs for designing expression cassettes allowing a precise control of where, when and at which level transcription occurs. Plant promoters can be exploited for isolating CREs driving a range of expression patterns including strong, constitutive expression and organ or tissue-specific expression (Potenza*, et al.*, 2004).

Plant CREs have commonly been delineated by the experimental manipulation of promoters and reporter gene expression assays (Guilfoyle, 1997). Such studies are labor-intensive, and this in turn has motivated the synergistic development of experimental and computational techniques for the identification of CREs (Elnitski*, et al.*, 2006). The computational prediction of CREs, referred to

as DNA motif discovery, consists of the identification of statistically overrepresented patterns in sets of sequences, typically promoter sequences upstream of co-regulated genes or orthologous genes. Difficulties associated with computational DNA motif discovery include the convergence towards patterns, such as low-complexity sequences, that are prevalent throughout the genome and not likely to represent regulatory elements, the computational requirements of enumerative approaches and the convergence towards local optima in sequence-driven approaches (GuhaThakurta, 2006; Stormo, 2000).

The identification of plant promoters and CREs driving tissue-specific expression in seed other tissues has been the objective of several studies (Guilfoyle, 1997; Potenza, et al., 2004). In this respect, seed storage protein (SSP) gene promoters provide an excellent model, since SSP genes are expressed, at high levels, specifically in seeds during seed maturation (Morton*, et al.*, 1995; Shewry*, et al.*, 1995). Many studies have characterized regulatory elements responsible for seed-specific gene expression (e.g. Chandrasekharan*, et al.*, 2003; Ellerstrom*, et al.*, 1996; Wu*, et al.*, 2000). However, to our knowledge, no large-scale computational analysis has been undertaken to identify motifs globally conserved in seed-specific promoters. Such knowledge will be beneficial for guiding the experimental characterization of CREs in newly sequenced or uncharacterized seed-specific gene promoters.

Plant *cis*-regulatory motifs (CRMs) are often reported as consensus sequences, a motif model of limited predictive power (Schneider, 2002). Collections of experimentally characterized plant CREs sequences such as the PLACE database (Higo*, et al.*, 1998) nevertheless remain an invaluable resource for plant promoter research. The position weight matrix (PWM) motif model (Stormo, 2000), based on the frequencies of nucleotides at each position in a collection of regulatory elements, is a better representation of CRMs that is used by a majority of contemporary computational approaches for the discovery of CREs (GuhaThakurta, 2006). A database of conserved CRMs in PWM format would therefore be a very useful resource for the computational identification of CREs in plant promoter sequences.

1.2 Definitions

Important terms are defined in this section to clarify concepts that are essential for the understanding of the thesis.

- **Gene**: functional region of genomic DNA, transcribed into ribonucleic acid (RNA).
- **Promoter**: regulatory region of DNA extending a few hundred base pairs (bp) upstream of gene's transcription start site (TSS).

- *Cis*-regulatory element (CRE): short segment of a promoter required for the correct expression of a gene.
- *Cis*-regulatory motif (CRM): recurring pattern in DNA, associated with a regulatory function.

1.3 Hypotheses

The research reported in this thesis is performed to test the following hypotheses.

- Computational methods for DNA motif discovery can be improved by the use of enumerative discriminative seeding.
- A data structure based on the geometry of the similarity matrix between combinations of nucleotide symbols will accelerate enumerative DNA motif discovery.
- Plant tissue-specific gene promoters contain combinations of conserved CRMs that are different in diverse plant families.
- The promoters of dicotyledonous orthologous genes contain conserved CRMs.

1.4 Objectives

My research objectives are:

- To design a DNA motif discovery algorithm using enumerative discriminative seeding.
- To compare the performance of the algorithm with that of other tools using biological benchmark data.
- To create a data structure based on the geometry of the similarity matrix between combinations of nucleotide symbols and to evaluate its performance in accelerating DNA motif discovery.
- To perform motif discovery in plant seed-specific promoters, and to compare conserved motifs in different plant families.
- To perform motif discovery in the promoters of dicotyledonous orthologous genes, and to identify major clusters of conserved motifs.

2. Literature review

2.1 “Gene” and “promoter” are evolving concepts

The notion and definition, in the greater biologist community, of what a gene is has evolved at a fast pace in recent years. From the early Mendelian concept of “discrete unit of heredity”, the gene became at the turn of the millennium and with the advent of genome sequencing and annotation, a “locatable region of genomic sequence” (Pearson, 2006). The ENCyclopedia Of DNA Elements (ENCODE) project currently defines the gene as “a union of genomic sequences encoding a coherent set of potentially overlapping functional products” (Gerstein, et al., 2007). Most importantly and in contrast to previous definitions, regulatory regions are excluded, the latter being “simply too complex to be folded into the definition” (Gerstein, et al., 2007).

The landmark work of Jacob and Monod (1961) suggested the existence of factors controlling gene expression by binding to regulatory DNA sequences. Two different *cis*-regulatory entities were originally defined: the promoter, bound by the RNA polymerase and at which transcription is initiated, and the operator, bound by regulatory proteins (Ullmann*, et al.*, 1965). The current view of the eukaryotic promoter organization (Heintzman and Ren, 2007) includes a core promoter extending approximately 50 bp upstream and downstream of the TSS,

which contains a minimal set of elements sufficient for transcription initiation (Figure 2.1). The proximal promoter, which contains additional CREs essential for transcriptional regulation, extends approximately 250 bp upstream of the TSS (Figure 2.1). Remote DNA elements that affect the rate of transcription were originally discovered in a virus (Benoist and Chambon, 1981; Gruss*, et al.*, 1981) and are termed enhancers (and silencers). Enhancers were originally defined as promoter elements affecting transcription "in a manner relatively independent of their position and orientation with respect to a nearby gene" (Khoury and Gruss, 1983). The current view of enhancer action involves DNA looping and direct interactions between enhancer-associated proteins and the target promoter (Blackwood and Kadonaga, 1998).

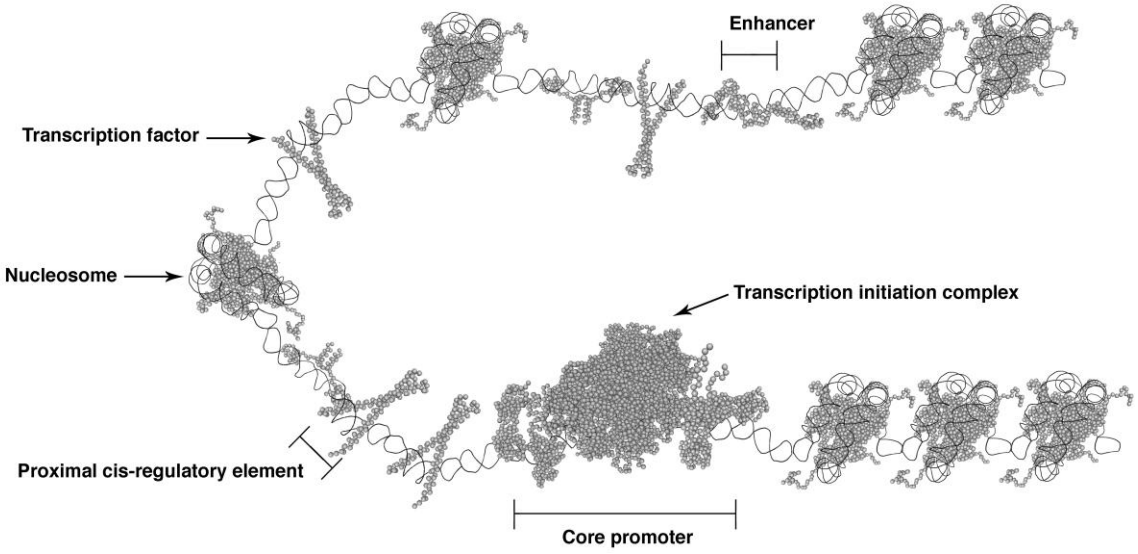


Figure 2.1: Overview of eukaryotic promoter organization.

2.2 Chromatin structure and gene expression

In the eukaryotic cell, DNA is enclosed in the nucleus and packaged into chromatin, a mixture of proteins and DNA. The double-helix is wound around histone octamers, forming nucleosome cores which are separated by 10-90 bp spacer DNA (Richmond and Davey, 2003). The fiber resulting from higher level packaging by linker histones is arranged in supercoiled loops attached by scaffold/matrix attachment regions (S/MARs) to a nuclear matrix consisting of proteins and RNA (Bode*, et al.*, 2003). Chromosomes are organized into heterochromatin (highly condensed, transcriptionally inert) and euchromatin (loosely condensed, transcriptionally active) regions that differ in respect to patterns of cytosine methylation and histone acetylation. Heterochromatin is characterized by methylation of residues Lys 9 and Lys 27 of the histone H3 tail and by low levels of core histone acetylation (Brzeski and Jerzmanowski, 2004). DNA Methylation decreases towards the promoter and 3' regions of genes (Zilberman*, et al.*, 2007). Furthermore, active promoter and enhancers are characterized by depletion in nucleosomes, and distinct patterns of histone modifications (Heintzman, et al., 2007).

2.3 Plant genomes and transcription factors

Different sets of genes are expressed through development, in response to stimuli and environmental conditions, and in different cell/tissue types. The highly organized and coordinated regulation of gene expression is primarily attributable to the binding of transcription factors (TFs) to specific *cis*-regulatory DNA elements, which activates or represses transcription (Dynan and Tjian, 1985; Lemon and Tjian, 2000). The understanding of plant transcriptional regulation has benefited from the recent sequencing and annotation of several plants, including Arabidopsis (*Arabidopsis thaliana* L. Heynh.) (AGI, 2000), rice (*Oryza sativa* L.) (Goff, et al., 2002; Yu, et al., 2002), poplar (*Populus trichocarpa* Torr. & A.Gray) (Tuskan, et al., 2006), grapevine (*Vitis vinifera* L.) (Jaillon, et al., 2007) and sorghum (*Sorghum bicolor* L. Moench) (Paterson, et al., 2009). The genome of the model plant Arabidopsis contains over 25,000 protein-coding genes, of which a large proportion (>1,500) codes for TFs. Of these, almost half (~45%) are plant-specific (Riechmann, et al., 2000). The picture is similar in other sequenced plant genomes, containing between 25,000 and 65,000 protein-coding genes and similar proportions of genes coding for TFs, grouped into some 60 families based on their DNA binding domains (Guo, et al., 2008). Plant TF-coding gene expansion rates are higher than that of other genes in plant genomes, and are higher in plants than in animals (Shiu*, et al.*, 2005). The latter

may be explained by the relatively high frequency of polyploidization in plants (Adams and Wendel, 2005). Examples of plant-specific TF families include the WRKY, involved in the plant response to stress, the NAC, involved in plant development, and the B3-domain, involved in signal transduction induced by plant hormones. Green plants have diverged from animals and fungi some 1.5 billion years ago (Wang*, et al.*, 1999) and they have indeed evolved remarkable features including photosynthesis, the biosynthesis of an extraordinary diversity of secondary metabolites, and adaptability to a range of environmental conditions. This may in part explain the relative high-complexity of plant transcriptional networks.

2.4 Plant promoters and transcription factor binding sites

Eukaryotic promoters generally contain conserved core elements (Hahn, 2004). Conserved core promoter elements are the binding targets of eukaryotic general TFs (TFIIA, TFIIB, TFIID, TFIIE and TFIIH) (Butler and Kadonaga, 2002; Smale and Kadonaga, 2003). The TATA-box is a well-known example of conserved core element. The TATA-box is bound by the TATA-binding protein (TBP), which is a subunit of the TFIID general TF (White and Jackson, 1992). Molina and Grotewold (2005) used a position frequency matrix (PFM) for the TATA-box derived from 305 experimentally characterized plant promoters (Shahmuradov*, et*

al., 2003) and, testing close to 13,000 core promoters in Arabidopsis, concluded that 30% of promoters contain a TATA-box located approximately 32 bp upstream of the TSS. Other eukaryotic promoter core elements include the initiator (Inr), the downstream promoter element (DPE), and the TFIIB recognition element (BRE) (Smale and Kadonaga, 2003). Plant core promoter elements identified by computational analysis (Shahmuradov, et al., 2003) include the CCAAT-box and the initiator (Inr) motif. The analysis of local distribution of short sequences in Arabidopsis and rice has also revealed that plant TSS are also associated with a "pyrimidine patch" (Y patch) (Yamamoto, et al., 2007).

A number of plant CREs have been characterized experimentally, and deposited in databases such as PLACE (Higo, et al., 1998), TRANSFAC® (Wingender*, et al.*, 1996), and JASPAR (Sandelin*, et al.*, 2004). In plants, studies have focused on promoters induced by environmental cues (e.g. light, temperature, dehydration, elicitors) and plant hormones, and on promoters responsible for tissue-specific expression (Guilfoyle, 1997). This focus is reflected in the importance of some classes of regulatory elements deposited in public databases *e.g* the PLACE database (Table 1.1).

Table 1.1: Ten most frequent keywords in the PLACE database.

Keyword	Count
Seed	92
Leaf	69
Shoot	67
Light	37
ABA	35
Root	30
bZIP	24
ABRE	21
Endosperm	21
Meristem	19

ABA, abscisic acid; bZIP, basic leucine zipper; ABRE, ABA-responsive element.

Like in other eukaryotes, plant promoters typically contain multiple elements allowing fine control of spatial and temporal gene expression. The degeneracy of binding sites can be compensated by the proximity of other sites, as a result of protein-protein interactions (Rombauts, et al., 2003). This modular organization, although adding an additional level of complexity, is exploitable for the computational identification of CRMs and modules.

2.5 Plant seed storage proteins and their promoters

Plant SSPs, because of their abundance and their economic importance, were among the first proteins to be characterized (Shewry*, et al.*, 1995). For example, a wheat glutenin was first isolated more than 250 years ago (Beccari, 1745). Seed storage proteins were classified by Osborne (1924) based on extraction

and solubility properties. Major SSPs include the albumins, soluble in water, and widely distributed in dicotyledonous plant species, the prolamins, soluble in water/alcool mixtures and found uniquely in *Poaceae*, and the globulins, soluble in dilute saline and widely distributed in both monocotyledonous and dicotyledonous plant species (Shewry*, et al.*, 1995). Seed storage protein promoters have a strong potential for biotechnology applications, because SSP genes are expressed at very high levels, and only in seeds (Morton*, et al.*, 1995; Shewry*, et al.*, 1995). Promoters conferring seed-specific expression are among the most studied plant promoters. In dicotyledons, the best-characterized CREs associated with seed-specific expression are the RY motif and the ACGT box, which are bound by B3 and bZIP TFs respectively (Vicente-Carbajosa and Carbonero, 2005). In monocotyledons, the best-characterized CREs associated with seed-specific expression are the GCN4-like (GLM) motif and the prolamin-box, which are bound by TFs of the bZIP and DOF families respectively (Vicente-Carbajosa and Carbonero, 2005). Promoters and CREs driving expression in other tissues and organs have also been identified, including green tissues, vascular tissues, roots, root nodules, pollen and flowers (Elliott and Shirsat, 1998; Eyal*, et al.*, 1995; Huang*, et al.*, 1995; Jensen*, et al.*, 1988; Keller and Heierli, 1994; Klinedinst*, et al.*, 2000; Mizukami*, et al.*, 1996).

2.6 Experimental characterization of *cis*-regulatory elements

Promoter-based experimental techniques for the identification of CREs typically involve the experimental manipulation of a promoter sequence (*e.g.* deletions) and visualization of resulting expression patterns (reporter gene expression assays). Heterologous systems are often used for such experiments, with the assumption that CREs and their cognate TFs are conserved in the plant species considered (Guilfoyle, 1997). Other methods are based on protein-binding assays (Elnitski, et al., 2006). Those include electrophoretic mobility shift assays (EMSA) (Garner and Revzin, 1981), systematic evolution of ligands by exponential enrichment (SELEX) (Tuerk and Gold, 1990) and chromatin immunoprecipitation assays on microarrays (ChIP-chip) (Ren, et al., 2000). This last approach has been used for the genome-scale identification of Arabidopsis TGA2 TFBSs (Thibaud-Nissen, et al., 2006). The ChIP-chip approach, however, needs to be complemented with computational methods for the precise identification of TFBSs.

2.7 Computational identification of *cis*-regulatory motifs

There may be small variations between functionally identical CRMs. This is because DNA-binding proteins have an affinity for a family of similar sequences rather than for a unique sequence (Stormo and Fields, 1998). However, TFs

must have a certain level of specificity in order to recognize a site in the midst of non-site sequences. Models must thus take into account the variability among *cis*-regulatory elements bound by a given TF in order to represent the specificity of the TF to its cognate binding site sequences. *Cis*-regulatory elements can be predicted by matching motif models in one or more DNA sequences. However, when the motifs sought are new or unknown, *de novo* DNA motif discovery is performed to identify patterns that are enriched and to predict potential functional elements.

2.7.1 *Cis*-regulatory motif models

Motifs are commonly represented either with consensus sequences, or with position frequency or weight matrices. In consensus sequences, the International Union of Pure and Applied Chemistry (IUPAC) degenerate symbols (Cornish-Bowden, 1985) are used to account for the variation at a given position in an alignment of sequences. Consensus sequences have the shortcoming of not providing the quantitative characteristics of binding specificity (Schneider, 2002). Consensus sequences and sequence logos are derived from an alignment of sequences corresponding to the motif *e.g.* a collection of TFBSs (Figure 2.2).

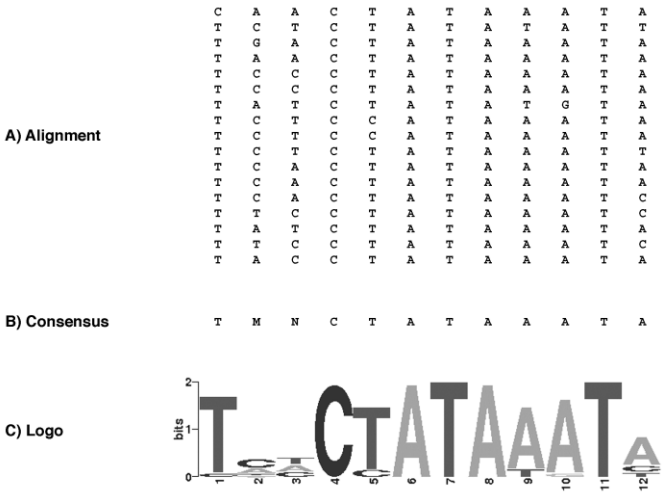


Figure 2.2: *Cis*-regulatory motif representation.

DNA motifs may be represented with alignments of *cis*-regulatory elements (A), consensus sequences (B) or sequence logos (C).

Position frequency and weight matrices are also built from the alignment of *cis*-regulatory element sequences. The position frequency matrix represents frequency of nucleotides at each position in the alignment. The PWM is computed from the PFM, using a base 2 logarithm of the frequencies, and adjusting for base frequencies in a background *e.g.* genomic set of sequences (Stormo, 2000). Figure 2.3 shows the frequency and weight matrices corresponding to the alignment of sequences in figure 2.2.

A) PFM

	A	C	G	T
P0	0	1	0	16
P1	5	9	1	2
P2	6	5	0	6
P3	0	17	0	0
P4	0	2	0	15
P5	17	0	0	0
P6	0	0	0	17
P7	17	0	0	0
P8	15	0	0	2
P9	16	0	1	0
P10	0	0	0	17
P11	12	3	0	2

B) PWM

P0	-2.3570186368606	-1.0084388232625	-2.3570186368606	1.26445781940235
P1	-0.1763545071375	1.54211040992947	-1.0084388232625	-1.0859223997540
P2	0.03264356589348	0.76946955608549	-2.3570186368606	0.03264356589348
P3	-2.3570186368606	2.41340995881423	-2.3570186368606	-2.3570186368606
P4	-2.3570186368606	-0.3237802141540	-2.3570186368606	1.17914198895243
P5	1.34500883948321	-2.3570186368606	-2.3570186368606	-2.3570186368606
P6	-2.3570186368606	-2.3570186368606	-2.3570186368606	1.34500883948321
P7	1.34500883948321	-2.3570186368606	-2.3570186368606	-2.3570186368606
P8	1.17914198895243	-2.3570186368606	-2.3570186368606	-1.0859223997540
P9	1.26445781940235	-2.3570186368606	-1.0084388232625	-2.3570186368606
P10	-2.3570186368606	-2.3570186368606	-2.3570186368606	1.34500883948321
P11	0.88797387522740	0.13863544849781	-2.3570186368606	-1.0859223997540

Figure 2.3: Position frequency and weight matrices.

Quantitative representations of DNA motifs include the position frequency matrix (A) and the position weight matrix (B).

Position weight matrices can be used directly to score DNA sequences and identify potential CREs. Probability values used to build PWMs can also be used to calculate the information content of a motif, in bits (Wasserman and Sandelin, 2004). This measure, introduced by Shannon (1948), can be used to estimate the statistical significance of a pattern. In the case of DNA sequences, it is defined for each position *i* in a motif, as (Stormo, 2000):

$$I_{seq}(i) = \sum_{b} f_{b,i} \log_2 \frac{f_{b,i}}{p_b}$$

where *b* refers to the four bases (A,C,G,T), $f_{b,i}$ is the frequency of each base and p_b is the frequency of base *b* in the genome. The minimum information content is

0 (where all bases are equally likely) and the maximum is 2 (where only one base is allowed). Adjacent position dependencies within motifs may be accounted for by the use of *n*-mer matrices or Hidden Markov Models (HMMs) of higher order (Bulyk*, et al.*, 2002).

2.7.2 Searching sequences with known DNA motifs

The vast majority of plant TFBS models deposited in public databases consists of consensus sequences. With such input, the search for putative new TFBS locations involves scanning DNA sequences with regular expressions. This approach has been applied at the genome-scale in Arabidopsis (O'Connor*, et al.*, 2005; Palaniswamy*, et al.*, 2006). Because consensus sequences do not capture the quantitative variability of TFBSs, the sensitivity and specificity of this methodology is not optimal. More accurate predictions may be achieved by using PWMs to scan DNA sequences. The advantage of PWM-based methods is that they are probabilistic, and the most significant matches may be identified as potential binding sites (Claverie and Audic, 1996; Staden, 1989). A PWM-based approach has been used to map a limited number of TFBS models (23) in Arabidopsis (Steffens*, et al.*, 2004).

Searching sequences with a known motif in either consensus or PWM format is relatively straightforward. However, distinguishing true sites from the

background noise is difficult. More accurate predictions can be achieved by matching multiple motifs to identify potential *cis*-regulatory modules (GuhaThakurta, 2006). Indeed, most module searching software require as input a set of known, single motif models (Klepper*, et al.*, 2008). However, since plant-specific databases contain consensus sequence motif models, of known limited predictive power, and since PWM databases have little covering of plant CREs, plant computational scientists often have to rely on *de novo* motif discovery for identifying putative binding sites in plant promoter sequences.

2.7.3 Computational DNA motif discovery

Computational DNA motif discovery involves searching sequences for imperfect copies of an unknown pattern and it is a notoriously difficult problem in bioinformatics (D'Haeseleer, 2006).

Low-complexity sequences can be masked before the input set is subjected to motif discovery using *e.g.* the RepeatMasker software (Pavesi*, et al.*, 2004) but there is no guarantee that real motifs will be avoided, and it also adds an additional level of complexity in statistical analysis since different proportions are masked in different sequences. The discriminative motif discovery strategy deals with this problem by identifying motifs that are over-represented (taking into account the distribution of motifs and not only their count) in a positive set of

sequences as compared to a background *e.g.* genomic set of sequences (Sinha, 2003).

Sequence sampling-based strategies for predicting motifs proceed by iteratively sampling subsequences of length *l* in N sequences of length L until convergence (local optimum), or a maximum of iteration, is reached. In such strategies, there is $(L-l+1)^N$ different combinations in the full search space, the number of possibilities thus increases exponentially with the number of sequences in the input set. Although relatively fast, heuristic PWM-based methods are not guaranteed to find a global optimum (GuhaThakurta, 2006; Stormo, 2000). Scanning a set of regulatory sequences with all possible PWMs would guarantee to find the global optimum, and perhaps the best TFBS predictions. The number of different PWMs that can be produced from the alignment of N sequences of length L, given an alphabet of A letters, is (adapted from Hertz and Stormo, 1999):

$$\left(\frac{(N+A-1)!}{N!(A-1)!}\right)^{L}$$

The infinite PWMs search space may thus be limited by using a finite number of sequences and further restricted based on the matrix length. A discriminative approach to motif discovery based on enumerating a discrete space of matrices has been developed recently (Smith*, et al.*, 2005).

Enumerative approaches for finding putative regulatory elements are guaranteed to yield a global optimum but are more expensive in terms of computational requirements. Algorithms using such strategy typically involve enumerating all words (nucleotide combinations), and then counting matches (with or without substitutions) in a positive set of sequences to identify potential candidates based on their overrepresentation. Yamamoto *et al*. (2007) have used an approach based on evaluating the distribution of hexamers and octamers in Arabidopsis and rice promoters, and have found several examples of words with high localized distribution close to the TSS. Nevertheless, counting exact occurrences of words may be too rigid an approach for identifying degenerate motifs.

2.7.4 DNA motif discovery algorithms

The history of algorithms designed to identify motifs in nucleic sequences, related in an excellent review by Stormo (2000), goes back some *ca.* 30 years ago. The first algorithm using weight matrices (Perceptron) was designed by Stormo *et al.* (1982) and was used to identify translation initiation sites in *E. coli*. The current method of calculating weight matrices, using the logarithm of base frequencies at each position, was introduced by Staden (1984). Since then, numerous algorithms have been designed to find motifs in unaligned sets of nucleic acid

sequences. A major landmark in DNA motif discovery was the introduction of the expectation-maximization (EM) algorithm by Lawrence and Reilly (1990). In this approach, a weight matrix is built from an initial set of sites, then all possible sites (subsequences) are scored and a new matrix is built from those maximizing the scores to the weight matrix; the process is iterated until convergence. This method has been implemented in the Gibbs Sampler (Lawrence*, et al.*, 1993) and in MEME (Bailey and Elkan, 1994), two algorithms that are still popular and in use today. The major advantage of this approach is its speed, but the disadvantage is that it may fail to identify the best solution (global optimum). A benchmark suite based on sets of TFBSs from the TRANSFAC® database (Wingender*, et al.*, 1996) was recently designed by Tompa *et al.* (2005), and 13 popular DNA motif discovery algorithms, including MEME (Bailey and Elkan, 1994) and Weeder (Pavesi*, et al.*, 2004) were benchmarked using this suite. The Weeder algorithm generally outperformed the other tools; however no single tool performed consistently better than all other tools in all datasets (Tompa*, et al.*, 2005). The Weeder algorithm uses an exhaustive enumeration of words but allows matches with a defined number of substitutions; overlapping words among over-represented words are combined to produce longer motifs (Pavesi, et al., 2004). Another recent enumerative algorithm, the Seeder algorithm (Fauteux*, et al.*, 2008), uses a different strategy: sums of distances between words and best

matching subsequence (with any number of substitutions) in a positive set and a background set of sequences are used to score words; and high-scoring words are used to seed motifs. The process called “seeding” is a heuristic commonly used in bioinformatics. For example, seeding is used as the initial step of the Basic Local Alignment Search Tool (BLAST) algorithm (Altschul*, et al.*, 1990). Seeding involves scoring words in a set of sequences, and then extending high-scoring words in both directions, in an attempt to find locally optimal alignments (in sequence alignment) or full width conserved patterns (in motif discovery). The Seeder algorithm is guaranteed to find a global optimum, it automatically avoids frequent patterns such as low-complexity sequences; another important advantage is the use of a data structure that makes exhaustive computations extremely fast at moderate seed lengths (Fauteux*, et al.*, 2008).

Preface to Chapter 3

As presented in Chapter 2, key issues for accurate and efficient computational DNA motif discovery include the convergence towards local optima and towards low-complexity patterns, and the computational requirements of enumerative approaches. In Chapter 3, we present Seeder, an algorithm addressing these difficulties. The algorithm is designed for fast and reliable DNA motif discovery in the promoter sequences of eukaryotic genes. The Seeder algorithm uses an enumerative approach and an objective function based on the probability of the sum of Hamming distances (HDs) between words and best matching subsequences, given a word-specific background probability distribution. This computation is accelerated by using the SMD index, a data structure allowing an efficient lookup, in a given sequence, for a subsequence minimally distant to a given word. An application of the algorithm to the identification of motifs significantly enriched in the promoters of Arabidopsis seed-specific genes is presented.

Reference: Fauteux, F., Blanchette, M., & Stromvik, M. V. (2008). Seeder: Discriminative Seeding DNA Motif Discovery. Bioinformatics, 24(20), 2303-2307.

3. Seeder: discriminative seeding DNA motif discovery

François Fauteux, Mathieu Blanchette and Martina V. Strömvik.

3.1 Abstract

3.1.1 Motivation

The computational identification of transcription factor binding sites is a major challenge in bioinformatics and an important complement to experimental approaches.

3.1.2 Results

We describe a novel, exact discriminative seeding DNA motif discovery algorithm designed for fast and reliable prediction of *cis*-regulatory elements in eukaryotic promoters. The algorithm is tested on biological benchmark data and shown to perform equally or better than other motif discovery tools. The algorithm is applied to the analysis of plant tissue-specific promoter sequences and successfully identifies key regulatory elements.

3.1.3 Availability

The Seeder Perl distribution includes four modules. It is available for download on the Comprehensive Perl Archive Network (CPAN) at http://www.cpan.org.

3.1.4 Supplementary information

Supplementary information is available at Bioinformatics online.

3.2 Introduction

The binding of TFs to relatively short and variably degenerate regulatory DNA sequences (*cis*-regulatory elements) is central to the regulation of gene expression (Orphanides and Reinberg, 2002). While several sequenced genomes are nearly deciphered in terms of the protein-coding gene repertoire, the inventory and comprehensive characterization of *cis*-regulatory elements remains elusive.

Motif discovery has motivated the development of numerous tools and algorithms, and the use of various motif models and statistical approaches (reviewed in GuhaThakurta, 2006). Motif discovery can be broadly divided into “sequence-driven” and “pattern-driven” methods. The former methods typically involve building a position-weight matrix from sequence data, and local search techniques such as expectation-maximization or Gibbs sampling are used to

optimize the log likelihood ratio until convergence or a maximum number of iterations is reached. Though routinely fast, those methods are not guaranteed to yield the best solution, or global optimum (Stormo, 2000). Enumerative methods, on the other hand, are guaranteed to find a global optimum but have the drawback of being computationally expensive and limited to short motifs.

Searching a set of sequences for patterns that are over-represented relative to a given background model may converge towards motifs that are prevalent in the genome thus not likely to represent regulatory elements. Sinha (2003) introduced the notion of "discriminative" motif discovery in which a motif is treated as a feature that leads to good classification between positive sequences deemed to contain common *cis*-regulatory elements and a set of background sequences.

In this work, we present the Seeder algorithm – a novel, exact discriminative seeding DNA motif discovery algorithm inspired by (Keich and Pevzner, 2002; Pizzi*, et al.*, 2005). The major benefits of the Seeder algorithm are (*i*) the use of intuitive and reliable statistics for the choice of motif seeds and (*ii*) a data structure that significantly accelerate the computation of motifs and background models. The algorithm is benchmarked against popular motif finding tools and demonstrates greater performance. The algorithm is applied to the analysis of Arabidopsis seed-specific (the plant structure seed, not to be

confused with motif seed) promoters and identifies motifs with high similarity to seed-specific *cis*-regulatory elements experimentally characterized in *Brassica napus*, a closely related species.

3.3 Methods

3.3.1 The Seeder algorithm

Our algorithm starts by enumerating all nucleotide combinations (words) of a given length, usually 6. For each word, it calculates the HD between the word and its best matching subsequence (we call this distance the substring minimal distance, SMD) in each sequence of a background set. This data is used to produce a word-specific background probability distribution for the SMD. For each word, it then calculates the sum of SMDs to sequences in a positive set. The *p*-value for this sum is calculated using the word-specific background probability distribution. The word for which the *p*-value is minimal is retained, and a seed PWM is built from the closest matches to this word found in every positive sequence. The seed PWM is extended to full motif width and sites maximizing the score to the extended PWM are selected, one in each positive sequence. A new PWM is built from those sites and the process is iterated until convergence, or a maximum number of iterations is reached.

3.3.1.1 Input data and parameters

Our algorithm takes as input a set $B=\{B_1,\ldots,B_m\}$ of m background sequences of length L, a set $P=\{P_1,\ldots,P_n\}$ of n positive sequences of length L, the length k of the motif seed and the length l of the full motif to discover.

3.3.1.2 Substring minimal distance

The HD between two strings of equal lengths is the number of positions at which symbols differ (Hamming, 1950). We define the SMD $d(w,w')$ between a short nucleotide sequence w and a longer sequence w' as the minimal HD between w and a $|w|$-length substring of w'.

3.3.1.3 Background model

A discrete random variable $Y(w)$ is associated with each word w of seed length k, corresponding to the SMD between w and a randomly selected background sequence from B. This w-specific distribution function is obtained empirically from B; for each word w, we set $g_w(y) = \Pr[\ Y(w) = y\ ] = |\ \{B_i:\ d(w,B_i) = y\}\ |\ /\ m$, for $y=0,\ldots,k$.

3.3.1.4 Seed position weight matrix

For each word w, the sum of SMDs to the positive sequences $S(w)=\sum_j d(w,P_j)$ is computed. Under the background model, the distribution function of this sum of n i.i.d. random variables is $g_w^{n*}(y)$, the n-fold self-convolution of $g_w(y)$ (Grinstead and Snell, 1997). The p-value (p) for word w with sum $S(w)$, which is the probability of obtaining a sum lower or equal to $S(w)$ under the assumption that P_j's are random in respect to w, is

$$p(S(w)) = \sum_{y=0}^{S(w)} g_w^{n*}(y)$$

The word w^* for which the p-value $p(S(w))$ is minimal is retained. For each positive sequence in P, the set of one or more subsequences of length k having the SMD to w^* are retained. A PWM P_0 is built from this set of selected subsequences using standard procedures and pseudocounts proportional to $\sqrt{n}$ (reviewed in Wasserman and Sandelin, 2004), with the modification that when a sequence contains more than one match, each match (subsequence) weight is reduced proportionally. The subsequence associated with the highest score to P_0 is retained in each sequence, and the seed PWM P_s is built from this optimal set of n subsequences, as described above.

3.3.1.5 Full length motifs

The seed PWM P_s is of width k, smaller than the full motif width. It is extended to full motif width l by adding null weights at $(l-k)/2$ positions upstream and downstream. The full length PWM is then refined by iterating the following process. (1) Sites (one per sequence in P) maximizing the score to the extended weight matrix are selected. (2) A revised full length PWM is built from those sites. This process is repeated until convergence (*i.e.* the sites maximizing the PWM score are fixed in all sequences) or for at most a default number of 10 iterations, which we observed to often be sufficient for the convergence of significant seeded motifs.

3.3.1.6 N-fold self-convolution

Our implementation of the n-fold self-convolution uses the binary expansion of n (Sundt and Dickson, 2000), and is an adaptation of the "square and multiply" algorithm (reviewed in Gordon, 1998) while convolutions *per se* are computed using the "input side algorithm" (Smith, 1997).

3.3.1.7 Multiple hypothesis testing correction

For each motif predicted, a list of 4^k p-values is generated thus prompting for a multiple testing correction. This is carried out by generating a list of q-values from

the list of *p*-values associated with words of seed length *k*, using the general algorithm for estimating *q*-values described in (Storey and Tibshirani, 2003). The statistical significance of a motif is evaluated with the *q*-value of the sum $S(w^*)$, which is the expected proportion of false positives incurred when calling the sum significant (i.e. not likely to have occurred if the positive sequences were randomly selected).

3.3.1.8 Searching both strands

Because TFBSs can be located either on the forward or the reverse strand, motifs are typically searched for on both strands. This is easily achieved with Seeder: one simply redefines the SMD so as to consider matches one both strands (for both the background and positive sequences) and perform PWM matching similarly.

3.3.1.9 Multiple motifs

When the user asks to retrieve more than one motif, the sites identified in the preceding run(s) are masked and the motif-finding process is repeated. The positions of the sites are obtained by scanning each sequence (plus strand first) until the highest scoring subsequence is found.

3.3.2 Data structures

The calculation of SMDs using direct string comparison approaches requires a considerable amount of operations and this probably explains in part why this quantity has not been more often exploited for DNA motif discovery. We have designed a data structure based on the organization of the matrix of HDs between words of length 6 (Supplementary Figure 1, Appendix 1). This structure, called the SMD index (Figure 3.1), allows very efficient lookup, in a given sequence, for a subsequence minimally distant to a given word, hence improving the efficiency of the SMD computation.

3.3.2.1 SMD index generation

Each nucleotide N is mapped to a numerical value (A,C,G,T -> 0,1,2,3). For a given word $w=w_1w_2\ldots w_k$ of length k, a list of indices is generated equivalent to a tree structure with levels 0,…,k-1. At each new level d of the tree, each node is expanded into four nodes, one for each possible nucleotide N$\in${0,1,2,3} at that position. An index i_d=N+(4 $\times$ i_{d-1}) is assigned to each new node, where i_{d-1} is the index of the parent node. At the final level, the tree has nodes and indices corresponding to all possible nucleotide sequences of length k. For a given node at a given level d, the HD is one more than that of the parent, except for the node corresponding to nucleotide w_{d+1}, where the HD is unchanged (Figure 3.1). The

SMD index is pre-computed for every word *w* of seed length *k* and HDs between 0 and 3, which requires a marginal amount of memory and appreciably accelerates the process.

3.3.2.2 SMD calculation

The number of occurrences of every word of length *k* in each sequence in *P* is stored using base 4 indexing (word count array). The SMD between *w* and sequence P_j is obtained by looking up elements in word count array of P_j, in order of increasing HD to w, until a non-zero count is found.

3.3.3 Benchmarking of motif discovery tools

The performance of the Seeder algorithm was compared with that of popular motif discovery tools using benchmarks designed for robust assessment of motif discovery algorithms (Sandve*, et al.*, 2007). In the benchmark suites, binding site sequences from the TRANSFAC® database (Wingender, et al., 1996) are represented either in their original genomic context sequences ("Model Real", MR; "Algorithm Real", AR) or in sequences generated with a third-order Markov model (MM) ("Algorithm Markov", AM). The reverse complement of sequences is used in cases where the original binding site appears on the negative strand, so all sites within the benchmark suites appear in the forward sequence. The MR

suite contains motifs that, according to Sandve *et al*. (2007), are harder to distinguish from the local background using common motif models (consensus, PWM and mismatch). The AM and AR suites each contain 50 datasets and a total of 810 sequences of mean length ~1300 nucleotides, and the MR suite contains 25 datasets and a total of 410 sequences of mean length of ~1250 nucleotides.

3.3.3.1 Parameter settings

In order to be representative of common usage where parameter adjustment is nominal while providing homogeneous instructions to different software, sequences were scanned in the forward orientation, searching for one motif of width 12 with one occurrence (site) per sequence. Other parameters were left to default values. We ran Seeder v. 0.01 (this paper), 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), the Gibbs Motif Sampler v. 3.03.003 (Lawrence, et al., 1993), and Motif Sampler v. 3.2 (Thijs, et al., 2001) on each dataset. The DIPS algorithm (Sinha, 2006) was not included in the benchmark study because it was associated with prohibitive runtime requirements under our computational conditions. Background models were generated separately for each suite using all sequences within the suite. Background distributions for words of length 6 were

generated using the Seeder::Background module. Frequency files (expected values for 6-mers and 8-mers) used by Weeder were generated using a custom Perl script. A sixth-order MM was generated for MEME using a custom Perl script, and for Motif Sampler using the INCLUSive CreateBackgroundModel program (Thijs, et al., 2002). The default (third-order) MM was generated for BioProspector using the genomebg program provided with the software.

3.3.3.2 Evaluation of motifs versus known binding sites

The predictions were evaluated using the suite of tools described in (Sandve, et al., 2007) (http://tare.medisin.ntnu.no). The predictions were scored using the nucleotide-level Pearsons correlation coefficient (nCC) (see Tompa, et al., 2005). Differences between scores were assessed using paired t-tests ($\alpha = 0.05$).

3.3.4 Motif discovery in the promoters of Arabidopsis seed-specific genes

A background set of 22,032 nuclear protein-coding gene promoters (500 bp upstream of the TSS) was generated using the TAIR (release 7) “loci upstream sequences” dataset (sequences preceding the 5' end of each transcription unit) and the “protein-coding with transcript support” listing (loci with supporting complementary DNA (cDNA) or expressed sequence tags (ESTs) deposited in Genbank), downloaded from the TAIR ftp server (ftp://ftp.arabidopsis.org).

Tissue-specific promoter sequence sets were assembled according to marker gene data from Schmid *et al.* (2005). The Seeder algorithm was used to perform motif prediction in seed-specific promoters using a seed length of six and a motif length of 12, and the "protein-coding with transcript support" gene promoters as a background.

3.5 Results

3.5.1 Performance of motif discovery tools

Figure 3.2 shows the differences between scores of different motif discovery tools on the benchmark suites of Sandve *et al*. (2007). On the AM suite, the performance of each tool was statistically equivalent. Interestingly, the tool that performed the best (though by a non-significant margin), BioProspector, models background sequences using a third-order MM, the same type as that used by Sandve *et al.* (2007) to generate the AM background sequences. Seeder, BioProspector, Weeder, MEME and the Gibbs Sampler scored equally on the AR suite, which contains binding sites in their original sequence. The MR suite also contains binding sites in their original sequence, but in this case the binding sites have a composition that is more similar to that of the surrounding background sequence. This suite was assembled for the purpose of testing novel motif

models (Sandve, et al., 2007). Seeder scored significantly higher on the MR suite than any other algorithm tested.

At first glance, it may seem surprising that the performance of some tools is actually higher on the MR suite than on AR suite. However, although the similarity of motifs to their local background does complicate the task of motif-finding approaches using local background models, this does not overly affect those based on global background models. It nonetheless appears that our discriminative approach to seed selection yields a non-negligible advantage to Seeder. Having said that, it should be noted that for a number of individual datasets the scores obtained by other tools are higher than that of Seeder, which highlights the complementary of these programs.

3.5.2 Arabidopsis seed-specific motifs

The Seeder algorithm was used to discover motifs (on both strands) in a set of 57 promoter sequences of *A. thaliana* seed-specific marker genes identified by expression data analysis (Schmid, et al., 2005). The computation of the background distributions (motif seed length of 6) took 35 minutes using a single Intel® x86 processor, and motif computation took ~3.5 minutes per motif reported. This example shows that most of the computing time is used to compute the background model, particularly when using genome-scale

background datasets. The Seeder::Background module was therefore designed to pre-compute background models which can be reused for any number of motif finding operations.

The top two predictions (*q*-value < 0.01) were compared to known plant motifs in the PLACE database (Higo, et al., 1998) using the STAMP web server (Mahony and Benos, 2007). The first motif (Fig. 3.3, m1) (*q*-value=4.4×10^{-9}, information content=7.4) and the second motif (Fig. 3.3, m2) (*q*-value=1.1×10^{-3}, information content=7.6) are similar to two experimentally characterized *cis*-regulatory elements found in the *napA* promoter in *Brassica napus*, the RY repeat (CATGCA) (E value=6.32×10^{-8}) and the G-box (CACGTG) (E value=2.92×10^{-5}) (Ezcurra*, et al.*, 1999). The function of these regulatory elements was shown by substitution mutation analysis using promoter-reporter gene fusions, leading to a strong reduction of the *napA* promoter activity in seeds (Ezcurra, et al., 1999). The second motif is also highly similar to a sequence (ACGTGTC) (E value=4.70×10^{-11}) overrepresented in the promoters of *A. thaliana* genes downregulated during seed germination (Ogawa, et al., 2003).

3.6 Conclusion

We have described a novel algorithm for DNA motif discovery and demonstrated its capacity to discover motifs in real biological datasets. Advantages of the

algorithm over other approaches include *i*) the enumerative-guaranteed optimality of seed selection, *ii*) a background model based on empirical distribution of SMDs and *iii*) efficient data structures that make background and motif computations relatively fast at moderate seed lengths.

We have benchmarked the algorithm against popular motif finding tools and demonstrated its performance to be equal or better than that of other tools on biological datasets. We note however that, although the Sandve *et al.* (2007) benchmarks proved extremely useful for our performance analysis, it would be ideal to have suites designed specifically for discriminative motif-finding algorithms.

Tompa *et al*. (2005) recommend biologists to use a few complementary tools, and to consider the top few predicted motifs of each tool. Based on the benchmarks results presented in this study, we recommend the inclusion of Seeder in the biologist's DNA motif discovery toolbox.

The present implementation of Seeder allows for motif searches in the mode "one occurrence per sequence" (oops). This assumption is deeply engrained in the algorithm and statistics for the selection of the motif seed and the construction of the seed PWM. Of course, once a good seed PWM has been selected, other search modes (e.g. "zero-or-one occurrence per sequence"

(zoops) or “any-number of repetitions” (anr)) could be implemented using the type of frameworks previously implemented in tools like MEME or BioProspector.

We have applied the algorithm to the analysis of *A. thaliana* seed-specific promoters and found that the top two motifs were similar to experimentally characterized *cis*-regulatory elements found in the promoters of *B. napus* seed-storage protein genes. This was unanticipated, considering the array of gene families and functions found in the seed-specific gene set from (Schmid, et al., 2005).

3.7 Funding

This work was funded by a Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant to M.V. Strömvik and an NSERC Postgraduate Scholarship to F. Fauteux. The authors also acknowledge support from Fonds québécois de recherche sur la nature et les technologies (FQRNT) and Centre SÈVE.

3.8 Aknowledgements

We thank G.K. Sandve (Norwegian University of Science and Technology, Trondheim, Norway) for helpful comments, and the Perl Monks

(http://perlmonks.org) for support in the development of the Perl modules. We also thank the reviewers for their constructive comments.

d	N	HD0	HD1	HD2	HD3
0	0 1 2 3	1	0 2 3		
1	0 1 2 3	4	0 5 6 7 8 12	1 2 3 9 10 11 13 14 15	
2	0 1 2 3	18	2 16 17 19 22 26 30 34 50	0 1 3 6 10 14 20 21 23 24 25 27 28 29 31 32 33 35 38 42 46 48 49 51 54 58 62	4 5 7 8 9 11 12 13 15 36 37 39 40 41 43 44 45 47 52 53 55 56 57 59 60 61 63

	Index	Representation	Sequence
HD0	18	102	CAG
HD1	2 16 (...) 50	002 100 302	AAG CAA TAG
HD2	0 1 (...) 62	000 001 332	AAA AAC TTG
HD3	4 5 (...) 63	010 011 333	ACA ACC TTT

Figure 3.1: SMD index generation.

SMD index generation for the word “CAG”. Each nucleotide (N) is mapped to a numerical value (A,C,G,T -> 0,1,2,3). For a given word $w=w_1w_2\ldots w_k$ of length k, a list of indices is generated equivalent to a tree structure with levels 0,…,k-1. At each new level (d) of the tree, each node is expanded into four nodes. An index i_d=N+(4 × i_{d-1}) is assigned to each new node, where i_{d-1} is the index of the parent node. For a given node at a given level d, the HD is one more than that of the parent, except for the node corresponding to nucleotide w_{d+1}, where the HD is unchanged.

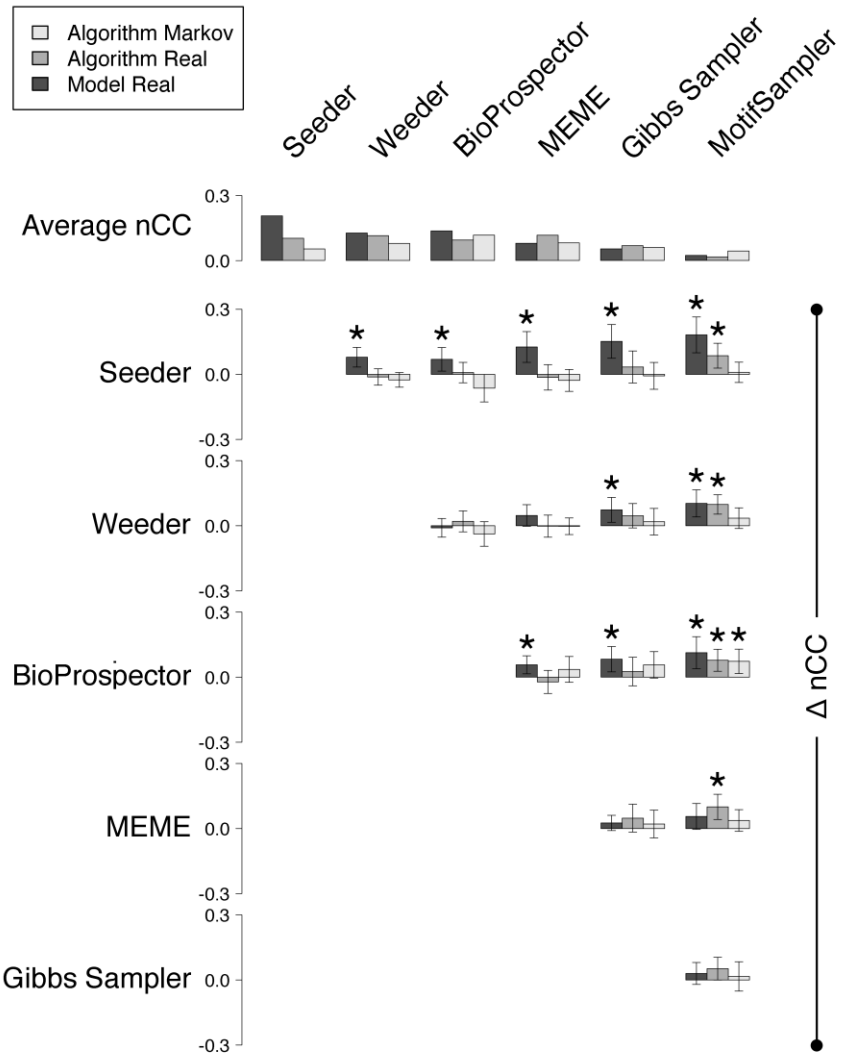


Figure 3.2: Average benchmarking scores and pairwise differences between motif discovery tools.

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.

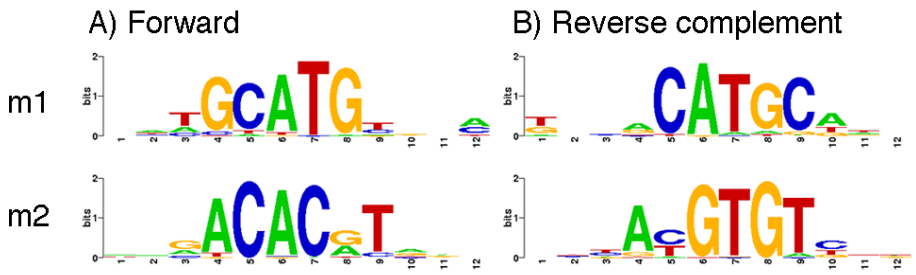


Figure 3.3: Arabidopsis seed-specific motifs.

Sequence logos of motifs overrepresented in the promoters of *A. thaliana* seed-specific marker genes. Motif 1 (m1) is an RY motif. Motif 2 (m2) is an ACGT motif. A) Full-length forward motifs. B) Reverse complement of motifs.

Preface to Chapter 4

In Chapter 3, we have described a novel exact discriminative seeding DNA motif discovery algorithm. The algorithm outperforms popular motif discovery tools on biological benchmark data. The algorithm is also applied to the analysis of the promoter of Arabidopsis seed-specific marker genes. Two motifs significantly enriched in seed-specific promoters are identified. In Chapter 4, we present an in-depth analysis of 54 SSP gene promoters from 14 plant species in three plant families. Plant SSP gene promoters have a strong potential for plant biotechnology applications because they are driving high-levels of expression specifically in the seed, which organ is ideal as a bioreactor *e.g.* for the production of recombinant proteins. Conserved motifs are identified in the promoters of *Brassicaceae* and *Fabaceae* SSP gene promoters, and distinct motifs are identified in the promoters of *Poaceae* SSP genes. Those results are of importance for the understanding of, and further experimental characterization of plant promoters driving seed-specific gene expression.

Reference: Fauteux, F., and Stromvik, M. V. (2009). Seed storage protein gene promoters contain conserved DNA motifs in *Brassicaceae*, *Fabaceae* and *Poaceae*. BMC Plant Biol. 9:126.

4. Seed storage protein gene promoters contain conserved DNA motifs in *Brassicaceae*, *Fabaceae* and *Poaceae*

François Fauteux and Martina V. Strömvik.

4.1 Abstract

4.1.1 Background

Accurate computational identification of CRMs is difficult, particularly in eukaryotic promoters, which typically contain multiple short and degenerate DNA sequences bound by several interacting factors. Enrichment in combinations of rare motifs in the promoter sequence of functionally or evolutionarily related genes among several species is an indicator of conserved transcriptional regulatory mechanisms. This provides a basis for the computational identification of CRMs.

4.1.2 Results

We have used a discriminative seeding DNA motif discovery algorithm for an in-depth analysis of 54 seed storage protein (SSP) gene promoters from three plant families, namely *Brassicaceae* (mustards), *Fabaceae* (legumes) and *Poaceae* (grasses) using backgrounds based on complete sets of promoters from a

representative species in each family, namely Arabidopsis, soybean (*Glycine max* (L.) Merr.) and rice respectively. We have identified three conserved motifs (two RY-like and one ACGT-like) in *Brassicaceae* and *Fabaceae* SSP gene promoters that are similar to experimentally characterized seed-specific *cis*-regulatory elements. *Fabaceae* SSP gene promoter sequences are also enriched in a novel, seed-specific E2Fb-like motif. Conserved motifs identified in *Poaceae* SSP gene promoters include a GCN4-like motif, two prolamin-box-like motifs and an Skn-1-like motif. Evidence of the presence of a variant of the TATA-box is found in the SSP gene promoters from the three plant families. Motifs discovered in SSP gene promoters were used to score whole-genome sets of promoters from Arabidopsis, soybean and rice. The highest-scoring promoters are associated with genes coding for different subunits or precursors of seed storage proteins.

4.1.3 Conclusions

Seed storage protein gene promoter motifs are conserved in diverse species, and different plant families are characterized by a distinct combination of conserved motifs. The majority of discovered motifs match experimentally characterized *cis*-regulatory elements. These results provide a good starting point for further experimental analysis of plant seed-specific promoters and our

methodology can be used to unravel more transcriptional regulatory mechanisms in plants and other eukaryotes.

4.2 Background

Designing expression cassettes allowing a precise control of where, when and at which level transcription should occur may ultimately be achieved through synthetic promoter engineering (Venter, 2007). The basic building blocks for such promoters are regions of *cis*-regulatory DNA, which in eukaryotes often comprise clusters of *cis*-regulatory elements (CREs) (called composite motifs, or modules) bound by a combination of TFs. The unraveling of eukaryotic transcriptional regulation is a challenging area of research driving the synergetic development of experimental and computational techniques (Elnitski*, et al.*, 2006). *Cis*-regulatory motifs of plant promoters have commonly been delineated by the experimental manipulation of DNA segments and reporter gene expression assays (Guilfoyle, 1997). Plant CRMs are often reported as consensus sequences, a motif model of limited predictive power (Schneider, 2002). Collections of experimentally characterized plant *cis*-regulatory elements sequences such as the PLACE database (Higo*, et al.*, 1998) nevertheless remain an invaluable resource *e.g.* for annotating motifs discovered in sequences that have not been characterized experimentally. The majority of contemporary computational approaches for the

discovery of *cis*-regulatory elements (GuhaThakurta, 2006) use the PWM motif model, based on the frequencies of nucleotides at each position in a collection of regulatory elements. The Seeder DNA motif discovery algorithm, designed for fast and reliable prediction of *cis*-regulatory elements in eukaryotic promoters, uses a string-based approach to identify motifs that are statistically significant (enriched) in a set of positive sequences as compared to a background set of sequences and it was recently shown to outperform some popular motif discovery tools on biological benchmark data (Fauteux*, et al.*, 2008).

The maturation of plant seeds, and more specifically protein storage in seeds, is regulated by a combination of hormonal, genetic and metabolic controls (Gutierrez*, et al.*, 2007). In Arabidopsis, four master regulators of seed maturation have been identified including three TFs of the B3 DNA-binding domain family, namely ABA INSENSITIVE3 (ABI3), FUSCA3 (FUS3) and LEAFY COTYLEDON2 (LEC2), and a HAP3 subunit of the CCAAT-box binding TF (LEC1) (Baud*, et al.*, 2008; Gutierrez*, et al.*, 2007; Santos-Mendoza*, et al.*, 2008). Known dicotyledonous seed maturation regulatory motifs include the RY motif and the ACGT motif, which are targets of B3 and bZIP TFs respectively (Vicente-Carbajosa and Carbonero, 2005). In rapeseed (*Brassica napus* L.), a comprehensive analysis of the *napA* promoter revealed the presence of two regulatory element complexes, the B-box which contains the distB element

(GCCACTTGTC) together with the proxB element (TCAAACACC), and the RY/G complex which contains two RY repeats (CATGCA) and one G-box (CACGTG) (Ellerstrom*, et al.*, 1996; Ezcurra*, et al.*, 1999; Stalberg*, et al.*, 1996). In bean (*Phaseolus vulgaris* L.), a comprehensive promoter analysis was performed on the *phas* promoter by Chandrasekharan *et al.* (2003). The site-directed substitution mutations analysis within the -295 region of the *phas* promoter revealed that the G-box, the CCAAAT box, the E-box (CACCGT) and RY elements mediate levels of expression in embryos (2003). Several studies have shown that motifs conferring seed-specific expression reside in the proximal region of the promoter, often within 500 bp upstream of the transcriptional start (e.g. Chamberland*, et al.*, 1992; Chandrasekharan*, et al.*, 2003; Lindstrom*, et al.*, 1990; Wu*, et al.*, 2000). The analysis of prolamin gene promoters from barley (*Hordeum vulgare* L.), wheat (*Triticum aestivum* L.) and maize (*Zea mays* L.) uncovered a conserved ~30 bp conserved sequence containing two CREs, the GCN4-like (GLM) element (GRTGAGTCAT) (see (Cornish-Bowden, 1985) for the nomenclature of incompletely specified bases), and the prolamin-box (also referred to as the endosperm element) (TGTAAAGT) (Forde*, et al.*, 1985). An additional element called AACA (AACAAACTCTATC) was further found to be involved in the seed-specific regulation of rice glutelin genes (Takaiwa*, et al.*, 1996). These three CREs (GLM, P-box and AACA) are frequently found in

monocotyledonous SSP gene promoters and are bound by TFs of the bZIP, DOF and MYB families, respectively (Vicente-Carbajosa and Carbonero, 2005).

In this work, we performed *de novo* motif discovery in 54 SSP gene promoters from *Brassicaceae*, *Fabaceae* and *Poaceae* using discriminative seeding DNA motif discovery, and uncovered the presence of family-specific conserved motifs, the validity of which was corroborated by matching to experimentally characterized plant seed-specific CREs. Furthermore, we show that the discovered motifs constitute signatures of SSP gene promoters in the different species.

4.3 Methods

4.3.1 Sequence data collection

The Uniprot database (Apweiler*, et al.*, 2004) release 14.6 was parsed using Bioperl (Stajich*, et al.*, 2002) and a total of 233 plant SSP were retrieved (annotated as SSP in description or keywords). Those records were matched to 230 UniRef100 entries (Suzek*, et al.*, 2007). Database references (EMBL) were used to retrieve a maximum of one promoter (500 bp upstream of the transcriptional start) per UniRef100 cluster using the BioPerl toolkit (Stajich*, et al.*, 2002). Transcriptional start positions were retrieved from The Arabidopsis Information Resource website (http://www.arabidopsis.org) and the Rice Genome

Annotation Project (http://rice.plantbiology.msu.edu) website for Arabidopsis and rice respectively. In other species, the transcriptional start positions were retrieved in the literature (Baumlein*, et al.*, 1986; Bown*, et al.*, 1988; Depigny-This*, et al.*, 1992; Derose*, et al.*, 1989; Doyle*, et al.*, 1986; Forde*, et al.*, 1985; Gatehouse*, et al.*, 1988; Josefsson*, et al.*, 1987; Kitamura*, et al.*, 1990; Newbigin*, et al.*, 1990; Pedersen*, et al.*, 1982; Rafalski, 1986; Rerie*, et al.*, 1990; Rodin*, et al.*, 1992; Ryan*, et al.*, 1989; Scheets and Hedgcoth, 1988; Sims and Goldberg, 1989; Sumner-Smith*, et al.*, 1985; Takei*, et al.*, 1989; Weschke*, et al.*, 1987). The TSSs were predicted in 13 promoters for which transcriptional start data was unavailable in GenBank or literature, using the TSSP software from Softberry Inc. (http://www.softberry.ru). One representative sequence among sequences with percentage identity > 0.90 over clustalw alignment (Larkin*, et al.*, 2007) was selected for further analysis. This process returned 15 *Brassicaceae* SSP gene promoter sequences, 17 *Fabaceae* SSP gene promoter sequences and 22 *Poaceae* SSP gene promoter sequences (listed in Supplementary Table 2, Appendix 4). Background sets of promoter sequences (500 bp upstream of annotated mRNAs) from Arabidopsis, soybean and rice sequences were retrieved using BioPerl and genome annotation data available for each species in generic feature format (GFF). A set of 27,234 promoters Arabidopsis protein-coding gene promoters were retrieved using The Arabidopsis Information

Resource release 8 (TAIR8) (http://www.arabidopsis.org). A set of 66,155 predicted soybean promoters were retrieved using the Glyma1.0 chromosome-scale assembly and genome annotation (Soybean Genome Project, DoE Joint Genome Institute) (http://www.phytozome.net/soybean). A set of 41,019 rice promoters was retrieved using the rice genome assembly and annotation release 5.0 (http://rice.plantbiology.msu.edu).

4.3.2 Computation of background distributions and motifs

For all sequence species, background SMD distributions were computed using a seed length of six and matches on both strands (Fauteux*, et al.*, 2008). For motif discovery in *Brassicaceae*, we used a background model based on Arabidopsis promoters, for *Fabaceae* we used a background model based on soybean promoters, and for motif discovery in *Poaceae* we used a background model based on rice promoters. Background models were computed using the Seeder::Background perl module (Fauteux*, et al.*, 2008). The Seeder algorithm was used to perform motif discovery in SSP gene promoters using a seed-length of six and a motif length of 12. The top-five motifs were compared to known plant motifs in the PLACE database (Higo, et al., 1998) using the STAMP web server (Mahony and Benos, 2007). For each group of promoters, quartiles and deciles

for the motif positions were computed using a custom perl script implementing the median-unbiased estimator algorithm (Hyndman and Fan, 1996).

4.3.3 Scoring of soybean promoter sequences

Scoring of the three promoter sets from soybean, Arabidopsis and rice was performed using PWMs as follow: for each given promoter, for a given PWM (in descending order of significance), each (unmasked) position is scored (Wasserman and Sandelin, 2004), and the position at which the score is maximum is masked; the process is repeated for each motif. Individual scores (for each motif) and the total score (for all motifs) are reported for each promoter sequence.

4.3.4 Annotation of soybean genes

Smith-Waterman alignments of the soybean predicted peptides corresponding to the top-ten scoring promoters was performed against the Uniprot release 14.6 (plant sequences) using a TimeLogic DeCypher system (Active Motif, Inc., 1914 Palomar Oaks Way, Suite 150, Carlsbad, CA. 92008) with BLOSUM62 scoring matrix, gap opening penalty -12, gap extension penalty -2 and an *E* value threshold of 1e-5. The top-scoring protein from Uniprot was reported for each soybean predicted peptide. For retrieving soybean genes corresponding to a

reference set of soybean SSP [Swiss-Prot:P04776, Swiss-Prot:P04405, Swiss-Prot:P11828, Swiss-Prot:P02858, Swiss-Prot:P04347, Swiss-Prot:P11827, Swiss-Prot:P13916, Swiss-Prot:P13916, Swiss-Prot:P25974, Swiss-Prot:P25974, Swiss-Prot:P13917, Swiss-Prot:P13917, Swiss-Prot:Q8RVH5, Swiss-Prot:Q8RVH5], alignment against all soybean predicted peptides (66,210 sequences) was performed. For each reference sequence, the soybean predicted peptide among hits with significance < 1e-100 and percent identities > 90% over the alignment maximizing the alignment score was attributed as best match.

4.4 Results

4.4.1 Seed storage protein gene promoters contain conserved motifs

Seed storage protein gene promoter sequences (the 500 bp upstream region of the transcriptional start) from *Brassicaceae* (15 promoters), *Fabaceae* (17 promoters) and *Poaceae* (22 promoters) were retrieved from public sequence databases. Discriminative seeding DNA motif discovery (Fauteux*, et al.*, 2008) was performed separately in each of the three plant families using a background model based on the complete set of promoters from a representative species, namely Arabidopsis (27,234 sequences), soybean (66,155 sequences) and rice (41,019 sequences). Statistically significant conserved CRMs (q-value < 0.05)

were identified in SSP gene promoter sequences within each plant family. Discovered motifs were matched to consensus sequences of experimentally characterized plant *cis*-regulatory elements from the PLACE database (Higo*, et al.*, 1998) using the STAMP suite of tools (Mahony and Benos, 2007) (Table 4.1).

Figure 4.1 (A) shows sequence logos of the significant motifs enriched in SSP gene promoters from *Brassicaceae* (B1-B3), *Fabaceae* (F1-F5), and *Poaceae* (P1-P7). Three motifs were statistically significant (q-value ≤ 0.05) in the *Brassicaceae* SSP gene promoters, corresponding to two RY-like motifs and one ACGT-like motif (motifs B1-B3).

Five significant motifs were found in the *Fabaceae* SSP gene promoters, including two RY-like motifs and one ACGT-like motif (motifs F1, F2, F5). Motif F3 is a TATA-box motif and is discussed below. The fourth motif discovered (motif F4) is possibly related to the E2Fb motif (GCGGCAAA) found in the tobacco (*Nicotiana tabacum* L.) ribonucleotide reductase 2 (*RNR2*) gene promoter (Chaboute*, et al.*, 2000). The *Fabaceae* E2Fb-like motif (motif F4) does not have similarity to any known plant seed-specific *cis*-regulatory elements; it is thus a novel putative SSP gene promoter CRMs.

Motifs enriched in the promoters of *Poaceae* SSP genes (seven significant motifs) are distinct from those observed in the two other plant families. The first motif discovered (motif P1) is most similar to the GCN4-like motif (GLM). The

second motif (motif P2) is similar to a variant of the prolamin-box motif (TGCAAAG) found in a rice glutelin promoter (Wu*, et al.*, 2000). This sequence has also been suggested to act as a prolamin-box variant in a wheat glutenin promoter (Thomas and Flavell, 1990). The third motif (motif P3) is a strong match to the typical prolamin-box (TGTAAAGT). Motif P4 is a TATA-box motif and is discussed below. The fifth motif (motif P5) has some core similarity with a rice BELL homeodomain TFBS (Luo*, et al.*, 2005). It is also similar to an Skn-1-like motif identified in a rice glutelin gene promoter (Washida*, et al.*, 1999). Motif P6 is related to the GCAA motif found in a maize zein promoter (So and Larkins, 1991). Motif P7 does not have similarity to any known monocotyledonous seed promoter motif but is weakly related to an opaque-2 recognition site (Vincentz*, et al.*, 1997).

4.4.2 Seed storage protein gene promoters contain TATA-box motifs

The third motif discovered in *Fabaceae* (motif F3), and the fourth motif discovered in *Poaceae* SSP gene promoters (motif P4), are highly similar to a TATA-box motif (CTATAAATA). In *Fabaceae* SSP gene promoters, the best matching subsequences to the TATA-box motif (motif F3) are localized between positions -20 to -30 upstream of the TSS (interquartile range of 7.0 bp). No TATA-box motif was initially discovered in *Brassicacea* SSP gene promoters. To

investigate whether *Brassicaceae* SSP gene promoters also contain a TATA-box motif, we searched the *Brassicaceae* promoter sequences with the TATA-box motif found in *Fabaceae* (motif F4). Scoring promoter sequences with the F4 motif's PWM returned a highly similar TATA-box motif (Figure 4.1 (B), motif BT). In both *Brassicaceae* and *Fabaceae*, most best matching subsequences to the TATA-box motif are also localized approximately -20 to -30 upstream of the transcriptional start (Figure 4.2).

4.4.3 Some seed storage regulatory motifs are highly localized

The position of the best matching subsequences to discovered motifs (putative CREs) in promoter sequences, identified by the Seeder algorithm (Fauteux*, et al.*, 2008), is illustrated in Figure 4.2. The distribution of best matching subsequence positions (deciles) is represented in Supplementary Figure 2, Appendix 2. Several patterns emerge from this map: (*i*) the TATA-box motif is highly localized to positions approx. between -20 to -30 upstream of the transcriptional start in *Brassicaceae*, *Fabaceae* and *Poaceae* SSP promoters; (*ii*) *Brassicaceae* and *Fabaceae* SSP promoters have one RY motif localized in close proximity upstream of the TATA-box, and one additional RY motif and one ACGT motif at variable position upstream of the TATA-box; (*iii*) *Poaceae* SSP

promoters are characterized by one GLM, two P-box, one Skn-1 and one GCAA motifs scattered at variable positions upstream of the transcriptional start.

4.4.4 The combination of *Fabaceae* seed storage motifs is a signature of seed storage protein gene promoters in the soybean genome

The recently sequenced soybean genome is predicted to contain over 65,000 protein-coding genes (Soybean Genome Project, DoE Joint Genome Institute http://www.phytozome.net/soybean). This publicly available genome sequence set was used to retrieve 66,155 promoter sequences. We used the *Fabaceae* PWMs (F1-5) to identify the best matching promoter sequences from the soybean genome by a PWM scoring and sequence matching strategy. In order to assign a function to the genes whose promoters were enriched in these six motifs, we manually annotated the top-ten matching gene sequences from the genome. The translated gene sequences corresponding to the top-ten scoring promoters were aligned with the Swiss-Prot database (plant sequences) using the Smith-Waterman algorithm. All of the top-scoring promoters are associated with soybean genes coding for different subunits of glycinin, β-conglycinin or 7S globulin (Table 4.2). Similar results were obtained in Arabidopsis and rice (Supplementary Table 1, Appendix 3), where eight out of the top-ten scoring

Arabidopsis promoters are associated with SSP genes and the top-ten scoring rice promoters are all associated with SSP genes.

4.4.5 The promoters of soybean genes coding for different seed storage protein subunits vary in motif composition

Although genes coding for different soybean SSP subunits have been shown to be expressed specifically in seeds during maturation, some subunits are differentially expressed (in cotyledons *vs*. embryonic axes; at different time points) (Meinke*, et al.*, 1981). We investigated whether there were also differences in promoter motif composition. Soybean major SSP sequences from the Swiss-Prot database, namely glycinin subunits (Gy1-Gy5) (Nielsen*, et al.*, 1989), β-conglycinin subunits (α,α', β) (Harada*, et al.*, 1989) and basic 7S globulins (Watanabe and Hirano, 1994), were aligned against all soybean predicted peptides (from the genome sequence). We identified 12 soybean peptide sequences with high similarity (percent identity over the alignment > 0.90, expected value < 1.0e-250), and an additional two sequences with moderate similarity (percent identity over the alignment > 0.50, expected value < 1.0e-50). Figure 4.3 shows the PWM scores for each *Fabaceae* SSP promoter motif in soybean SSP gene promoters compared with a baseline (the mean score of all 66,155 soybean promoters). The promoters of genes coding respectively

for glycinins Gy1 (Glyma03g32030.1), Gy2 (Glyma03g32020.2), Gy3 (Glyma19g34780.1), Gy4 (Glyma10g04280.1) and Gy5 (Glyma13g18450.1) scored relatively high (ranks 1, 4, 2, 5 and 6) for the presence of *Fabaceae* SSP gene promoter motifs. The promoters of all genes coding for the β-conglycinin subunits, namely α' (Glyma10g39150.1), α (Glyma20g28660.1, Glyma20g28650.2) and β (Glyma20g28640.1, Glyma20g28460.2) were among the top-15 scoring promoters (ranks 7, 13, 3, 8, 9) out of the 66,155 soybean promoters. The promoters of the gene coding for the basic 7S globulin 1 (Glyma03g39940.1) was also among the top-ten promoters (rank 10), while that of the gene coding for the basic 7S globulin 2 (Glyma19g42490.1) scored lower (rank 177). The products of two genes flanking gene Glyma10g39150.1 on chromosome 10 (Glyma10g39160.1, Glyma10g39170.2) are equivalently good matches to the three β-conglycinin subunits (α, α' and β) (percent identity > 0.50, expected value < 1.0e-50), making a precise annotation difficult for those two genes. Interestingly, the promoter of Glyma10g39160.1 scored very low (rank 3,252) while that of Glyma10g39170.2 was among the top-15 scoring promoters (rank 12).

4.5 Discussion

We have applied the Seeder discriminative DNA motif discovery algorithm to an in-depth analysis of SSP gene promoters from *Brassicaceae*, *Fabaceae* and *Poaceae*. Most discovered motifs match experimentally characterized *cis*-regulatory element consensus sequences, which strongly supports the validity of the discovered motifs.

The analysis of *Brassicaceae* SSP gene promoters highlighted the presence of three significant motifs corresponding to two RY motifs and one ACGT motif. It is interesting to contrast this result with that obtained from the analysis of promoters of Arabidopsis seed-specific marker genes where one RY motif and one ACGT motif were significantly enriched (Fauteux*, et al.*, 2008). The three motifs match components of the RY/G complex experimentally characterized in the rapeseed *napA* promoter (Ellerstrom*, et al.*, 1996). The analysis of *Brassicaceae* SSP gene promoter sequences using the Seeder algorithm did not initially reveal enrichment in a TATA-box motif. This could be explained by the proportion of promoters containing a TATA-box in the background set of sequences, or by the relatively low complexity of TATA-box motifs which makes them hard to discriminate from background, particularly if we take into account the fact that promoter sequences are generally A/T rich (Pandey and Krishnamachari, 2006). We used a PWM corresponding to a

putative *Fabaceae* TATA-box motif to retrieve, in *Brassicaceae* SSP gene promoter sequences, a motif highly localized around position -20 to -30 relative to the transcriptional start site. The localization, the information content of the motif and the fact that it is very similar to TATA-box motifs found in *Fabaceae* and *Poaceae* SSP gene promoters suggest that this motif indeed corresponds to a *Brassicaceae* SSP gene promoter TATA-box motif, in accordance with reported occurrences of TATA-box motifs in the promoters of *e.g*. *napA* and *napB* (Ericson*, et al.*, 1991; Josefsson*, et al.*, 1987).

Fabaceae SSP gene promoters have also revealed enrichment in two RY motifs. The RY motif has long been known to be conserved in legume seed-protein gene promoters (Dickinson*, et al.*, 1988) and RY CREs have been proven to be functional *e.g*. in soybean (Fujiwara and Beachy, 1994; Lelievre*, et al.*, 1992) and broad bean (*Vicia faba* L.) (Baumlein*, et al.*, 1992). A novel, E2Fb-like motif was discovered in *Fabaceae* SSP gene promoters. E2F TFs are involved in the control of cell cycle (Inze and Veylder, 2006). The role of this E2Fb-like motif in seed-specific gene expression will require further experimental verification.

Position weight matrices corresponding to motifs discovered in *Fabaceae*, Arabidopsis and rice SSP gene promoters were used to score the respective whole genome sets of promoter sequences. The top-ten scoring promoters are associated with SSP-coding genes in soybean and rice, as are eight out of the

top-ten scoring promoters in Arabidopsis. This combination of few motifs is thus sufficient to constitute a signature of SSP gene promoters. The fact that the promoter of some soybean genes coding for SSP protein subunits did score relatively low to the combination of *Fabaceae* SSP gene promoter motifs may indicate alternative regulatory mechanisms for those genes. Furthermore, the promoters of other soybean SSP protein genes such as those coding for albumin-1 (Glyma13g26330.1, Glyma13g26340.1) and 2S albumin (Glyma12g34160.1, Glyma13g36400.1) did also score relatively low (data not shown) and could be regulated by a different set of TFs.

In soybean, experimental SSP gene promoter analyses have focused on the α' and β subunits of β-conglycinin (Allen*, et al.*, 1989; Chamberland*, et al.*, 1992; Chen*, et al.*, 1988; Chen*, et al.*, 1986; Lessard*, et al.*, 1991; Lessard*, et al.*, 1993). Experimental analyses have revealed the importance of the proximal region (~ 250 bp upstream of the TSS) and the presence of several factors binding the promoters (soybean embryo factors SEF) and the presence of a RY *cis*-regulatory element. The study by Fujiwara and Beachy (1994) disproved a *cis*-regulatory role for the binding sites of SEF3 and SEF4 located within the proximal promoter and confirmed the role of the RY element in seed-specific gene regulation. The work by Yoshino *et al.* (2001; 2006) on the promoter of the α subunit of β-conglycinin also suggests a role for RY elements in seed-specific

gene regulation. The promoters of genes coding for glycinin subunits Gy2 and Gy3 have also been analyzed experimentally (Itoh*, et al.*, 1993; Itoh*, et al.*, 1994; Lelievre*, et al.*, 1992) yet although an A/T-rich SEF-binding sequence has been identified, the only clearly confirmed *cis*-regulatory element therein is a RY element. Our results suggest that soybean SSP promoters may be characterized by four CRMs, in addition to a TATA-box motif.

Motifs enriched in the promoters of *Poaceae* SSP genes were all good matches to experimentally characterized plant seed-specific *cis*-regulatory elements including a GLM motif, two prolamin-box-like motifs, a Skn-1-like motif and a TATA-box motif. A recent study (Moreno-Risueno*, et al.*, 2008) has identified a barley protein homologous to the Arabidopsis FUSCA3 that regulates SSP genes and binds RY boxes; this was the first report of a possible implication of the RY motif in seed-specific gene regulation in a monocotyledonous plant species. Our computational analysis did not reveal significant enrichment in RY motifs among *Poaceae* SSP gene promoters. This however does not necessarily refute a possible role for B3-type TFs and RY-like elements in the transcriptional regulation of some *Poaceae* SSP genes, which could be an attribute of a limited number of genes only, and not a general feature of *Poaceae* SSP gene promoters.

In counterpart, motifs containing the AAAG core of Dof TFBSs (Yanagisawa and Schmidt, 1999) were found only in *Poaceae* SSP gene promoters. Soybean Dof-type TFs have been reported to be involved in the regulation of the lipid content in soybean seeds (Wang*, et al.*, 2007), and a prolamin-box motif has been reported in pea (*Pisum sativum* L.) (Shirsat*, et al.*, 1989). However, prolamin-box motifs have been reported mostly in *Poaceae* promoters (e.g. Forde*, et al.*, 1985; Mena*, et al.*, 1998; Muller and Knudsen, 1993; Thomas and Flavell, 1990; VicenteCarbajosa*, et al.*, 1997; Wang*, et al.*, 2007; Wu*, et al.*, 2000). Indeed, our results suggest that prolamin-box-like motifs are conserved in *Poaceae* SSP gene promoters, but are not featured in *Brassicaceae* or *Fabaceae* SSP gene promoters.

4.6 Conclusions

Presented results highlight motifs that are conserved in SSP gene promoters within three plant families. Promoter/motif combinations generated in this analysis can be further validated experimentally, *e.g.* in a framework such as that used by (Chandrasekharan*, et al.*, 2003). Most motifs conserved in SSP gene promoters have a high degree of similarity with experimentally characterized *cis*-regulatory elements; this is an indicator that they are indeed functional in seed-specific gene regulation. The same methodology can be applied to analyze

various data sets and decipher transcriptional regulation mechanisms in plants and other eukaryotes.

4.7 Authors' contributions

FF and MVS designed the study. FF performed programming and data analysis. MVS supervised the project. Both authors have participated in writing the manuscript and have read and approved the final version.

4.8 Acknowledgements

The authors thank Mathieu Blanchette for critical reading of the manuscript, and acknowledge the Natural Sciences and Engineering Research Council of Canada (NSERC) for a Discovery grant to M.V.S. and an NSERC Postgraduate Scholarship (PGS D) to F.F.. We also acknowledge le Fonds de recherche sur la nature et les technologies (FQRNT) and the Centre Sève for financial support.

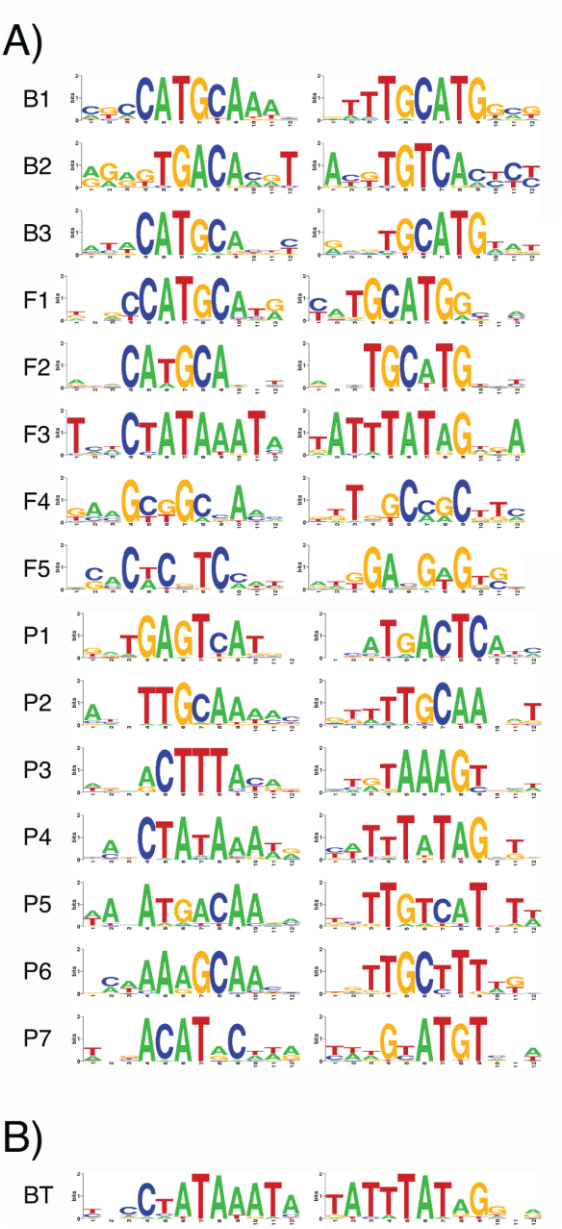


Figure 4.1: Sequence logos of motifs enriched in seed storage protein gene promoter sequences

A) Sequence logos of significant DNA motifs discovered in SSP gene promoter sequence from *Brassicaceae* (B1-3), *Fabaceae* (F1-5) and *Poaceae* (P1-P7). B) Sequence logos of the TATA-box motif identified in *Brassicaceae* SSP gene promoter sequences. Left, forward motif; right, reverse complement of motif.

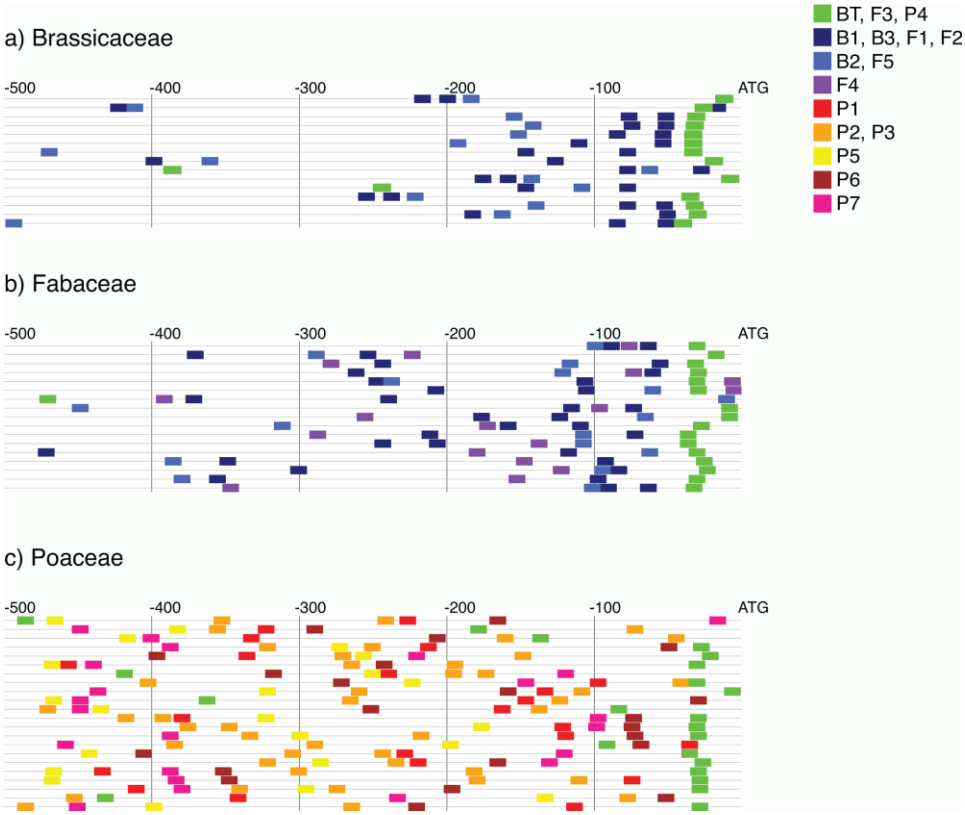


Figure 4.2: Position of cis-regulatory motifs on seed storage protein gene promoter sequences

The positions of the best matching subsequence to motifs discovered in SSP gene promoters from (a) *Brassicaceae*, (b) *Fabaceae* and (c) *Poaceae* are mapped on promoter sequences. (BT, F3, P4) TATA-box motifs; (B1, B3, F1, F2) RY motifs; (B2, F5) ACGT motifs, (F4) E2Fb-like motif; (P1) GLM motif; (P2, P3) prolamin box motifs; (P5) Skn-1-like motif; (P6) GCAA motif; (P7) opaque-2 recognition site-like motif.

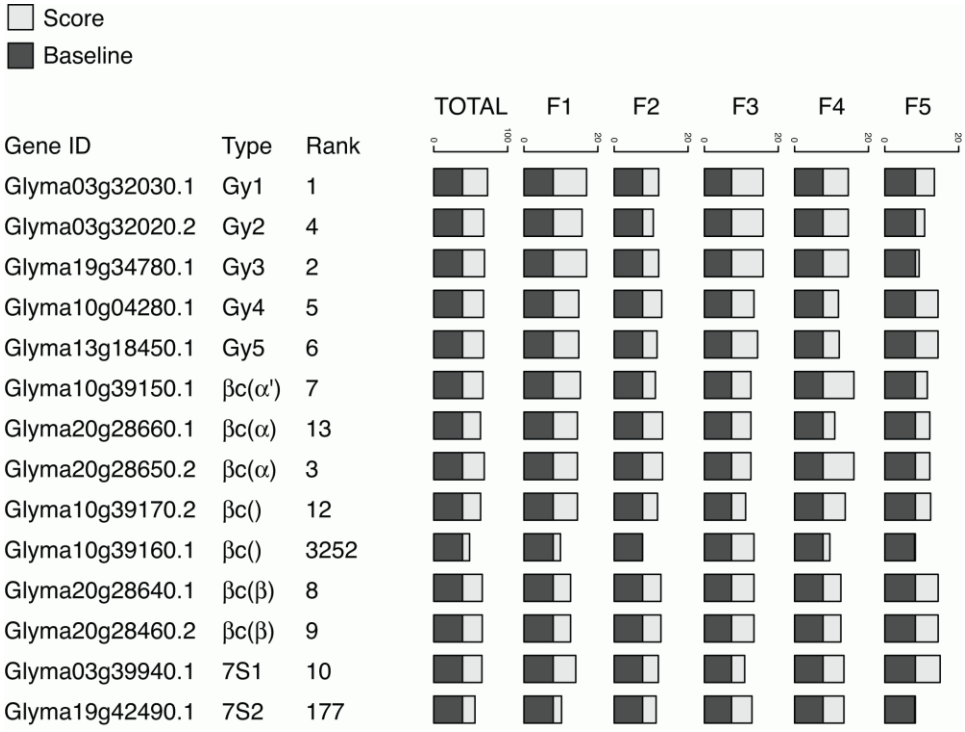


Figure 4.3: PWM score and rank of Fabaceae SSP gene promoter motifs in 14 soybean SSP gene promoters

The PWM matrix score associated to *Fabaceae* SSP gene promoter motifs in 14 soybean SSP gene promoters is compared to the average score obtained in 66,155 soybean promoters (baseline). Gy (1-5), glycinin subunit (1-5); βc (α', α, β), β-conglycinin subunit (α', α, β).

Table 4.1: DNA motifs discovered in the promoters of plant seed-storage protein genes

Plant family	Motif ID	*q*-value	PLACE ID	STAMP alignment	*E* value
Brassicaceae	B1	1.60e-07	RYREPEATBNNAPA	MKCCATGCAAAN ---CATGCA---	5.02-08
	B2	5.45e-04	GADOWNAT	AYKTGTCACYCY ACGTGTC-----	6.86e-08
	B3	1.84e-02	RYREPEATBNNAPA	NYWCATGCANNY ---CATGCA---	9.68e-08
Fabaceae	F1	1.43e-07	LEGUMINBOXLEGA5	NNRCCATGCATR TAGCCATGCAWR	4.73e-12
	F2	2.06-03	RYREPEATLEGUMINBOX	RNNCATGCANNN ---CATGCAY--	1.05e-09
	F3	7.24e-03	TATABOX1	TMNCTATAAATA ---CTATAAATA	1.58e-12
	F4	9.05e-03	E2FBNTRNR	KAMGCGGCNAMN ---GCGGCAAA-	9.03e-05
	F5	4.53e-02	ACGTSEED2	NSACWCNTCMWY ACACACGTCAA-	1.32e-08
Poaceae	P1	6.84e-08	GLMHVCHORD	KRTGAGTCATNN -RTGASTCAT--	1.52e-13
	P2	2.17e-05	PROLAMINBOXOSGLUB1	ANNTTGCAAAMN ----TGCAAAG-	4.41e-06
	P3	1.85e-04	EMHVCHORD	NYRTAAAGTNNW -TGTAAAGT---	6.45e-11
	P4	2.94e-03	TATABOX1	NANCTATAAAWR ---CTATAAATA	6.12e-10
	P5	9.17e-03	BIHD1OS	KNTTGTCATNTW ---TGTCA----	6.65e-06
	P6	1.14e-02	GCAACREPEATZMZEIN	NMWAAAGCAANN -GCAACGCAAC-	5.47e-03
	P7	2.82e-02	O2F3BE2S1	WNNACATRCWWR TCCACGTACT--	1.55e-05

q-value, statistical significance of motif

PLACE ID, identifier of PLACE consensus sequence matching motif

STAMP alignment, alignment of motif consensus sequence (top) with PLACE consensus sequence (bottom)

E value, expectation value of the STAMP alignment

Table 4.2: Top-ten scoring soybean promoters for the presence of Fabaceae seed-storage protein gene promoter motifs

Gene ID	PWM rank	Hit id	Hit description	*E* value
Glyma03g32030.1	1	P04776	Glycinin Gy1	0.0
Glyma19g34780.1	2	P11828	Glycinin Gy3	0.0
Glyma20g28650.2	3	P13916	β-conglycinin, alpha chain	0.0
Glyma03g32020.2	4	P04405	Glycinin Gy2	2.0e-251
Glyma10g04280.1	5	P02858	Glycinin Gy4	0.0
Glyma13g18450.1	6	P04347	Glycinin Gy5	3.0e-285
Glyma10g39150.1	7	P11827	Beta-conglycinin, alpha' chain	3.0e-105
Glyma20g28640.1	8	P25974	β-conglycinin, beta chain	7.0e-298
Glyma20g28460.2	9	P25974	β-conglycinin, beta chain	1.0e-269
Glyma03g39940.1	10	P13917	Basic 7S globulin 1	7.0e-302

Gene ID, gene identifier (soybean genome assembly and annotation Glyma1)

PWM rank, total PWM matching score rank

C-B rank, total Cluster-Buster score rank

Hit ID, hit identifier (Uniprot/Swiss-Prot)

Description, hit description

E value, hit alignment expectation value

Preface to Chapter 5

In Chapter 4, we presented an analysis of SSP gene promoters in 14 plant species, using the Seeder discriminative seeding DNA motif discovery algorithm. The analysis revealed conserved motifs in the promoters of SSP genes within plant families, and different sets of motifs conserved in different plant families.

In Chapter 5, we present a large-scale analysis of conserved motifs in orthologous gene promoters in five dicotyledonous plant species. A total of 1,773 significant motifs are discovered in 1,335 groups of orthologous promoters. Three major clusters of motifs associated with the promoters of genes coding for proteins involved in fundamental cellular processes are highlighted. This data is structured in a public database that can be downloaded or queried over the web.

5. Promoters of dicotyledonous orthologous genes involved in fundamental cellular processes are enriched in highly conserved *cis*-regulatory motifs

François Fauteux and Martina V. Strömvik.

5.1 Abstract

The repertoire of plant *cis*-regulatory elements is largely undetermined. *Cis*-regulatory elements are conserved in promoters of some orthologous genes, as a result of regulation by TFs containing conserved DNA-binding domains. In this study, groups of orthologous genes were assembled computationally from five dicotyledonous plant genomes: Arabidopsis, Medicago (*Medicago truncatula*), soybean, grapevine and poplar. DNA motif discovery was carried out in the proximal promoter sequence within each orthologous group. In total, 1,773 significant motifs were discovered in 1,335 groups of orthologous promoters. Co-clustering of DNA motif similarity and gene ontology semantic similarity revealed the presence of three important clusters of groups of orthologous promoters sharing similar motifs: (*i*) promoters of genes involved in translation, (*ii*) promoters of genes involved in DNA metabolism and replication and (*iii*)

promoters of histone genes. The annotated motif data is structured in a database that can be downloaded or queried over the web (http://ddopm.agrenv.mcgill.ca).

5.2 Introduction

The genome of the model plant Arabidopsis was the first to be sequenced and is at present the best annotated plant genome. Arabidopsis *cis*-regulatory components nevertheless remain mostly undefined. The Arabidopsis genome contains over 25,000 protein-coding genes. A large number of Arabidopsis genes (>1,500) codes for TFs, of which 45% are specific to plants (Riechmann*, et al.*, 2000). The fact that the number of TFBSs in any given eukaryotic genome could be over an order of magnitude higher than the number of coding genes (*i.e.* each promoter may contain several, up to 10 or more TFBSs) (GuhaThakurta, 2006) and the fact that several factors typically interact at composite *cis*-regulatory regions (Istrail and Davidson, 2005) make the broad understanding and comprehensive experimental characterization of *cis*-regulatory elements and modules a complex task and a long term objective.

Cis-regulatory elements (CREs) are regions of DNA containing TFBSs and regulating the expression of genes. Transcription factors display varying levels of binding specificity. One given TF typically binds a range of (slightly) different DNA sequences. *Cis*-regulatory motifs are patterns associated with a

group of sequences bound by a given DNA-binding domain. *Cis*-regulatory motifs can be represented with degenerate consensus sequences, with position frequency and weight matrices, or with sequence logos.

A multitude of plant promoters have been characterized experimentally. *Cis*-regulatory motifs, often in the form of consensus sequences, have been collected and deposited in databases such as TRANSFAC® (Wingender*, et al.*, 1996) or plant-specific databases such as PLACE (Higo*, et al.*, 1998). Such knowledge resources are important for plant biotechnology research and development (Potenza*, et al.*, 2004). Solutions for accelerating the discovery and characterization of *cis*-regulatory elements and for reaching the ultimate goal of deciphering *cis*-regulomes include the synergistic development of high-throughput experimental and computational methods (Elnitski*, et al.*, 2006).

Datasets used as positive input for computational identification of *cis*-regulatory elements commonly include promoters of co-regulated genes, or promoters of orthologous genes. The ortholog-based approach has motivated the development of a database of orthologous promoters from *Viridiplantae* and *Chordata* species, DoOP (Barta*, et al.*, 2005).

In this work, we have identified groups of orthologous genes (OGs) across five dicotyledonous plant species using the MultiParanoid algorithm (Alexeyenko*, et al.*, 2006) and performed DNA motif discovery in groups of orthologous gene

promoters (OPs) using the Seeder DNA motif discovery algorithm (Fauteux*, et al.*, 2008). We report a total of 1,773 significant motifs ($q \leq 0.05$) discovered in 1,335 OPs. We also present the results of a co-clustering analysis of DNA motif similarity and gene ontology (GO) semantic similarity, and demonstrate that similar motifs are found in OGs of similar function. Finally, we present an online database resource, the Database of Dicotyledonous Orthologous Promoter Motifs (DDOPM), accessible at (http://ddopm.agrenv.mcgill.ca). The user can query the database, using motif identification (ID), GO term ID, or keywords, and access individual OG annotations and motif data, including sequence logo and PFM. Alternatively, the user can match motifs in the database using STAMP (Mahony and Benos, 2007). Motifs can also be downloaded in TRANSFAC®-like format.

5.3 Methods

5.3.1 Sequence data collection

Genomic sequences, protein sequences and annotation data was downloaded for each of five plant species (Arabidopsis, Medicago, soybean, poplar, grapevine) from The Arabidopsis Information Resource (http://www.arabidopsis.org) (TAIR8), The International Medicago Genome Annotation Group (IMGAG) (http://www.medicago.org) (genome release 2.0), the

Glyma1.0 chromosome-scale assembly and genome annotation (Soybean Genome Project, DoE Joint Genome Institute) (http://www.phytozome.net/soybean), the Genoscope genome release 1 (http://www.genoscope.cns.fr) and the Joint Genome Institute (JGI) poplar genome release 1.1 (http://genome.jgi-psf.org). For each species, we collected unique gene and protein models, and corresponding promoter sequences (500 bp upstream of annotated mRNA). We retrieved 27,235, 38,728, 66,210, 45,555 and 30,434 protein sequences, and 27,234, 38,737, 66,155, 42,177 and 30,432 promoter sequences for Arabidopsis, Medicago, soybean, poplar and grapevine respectively. Protein sequence data was used as input for sequence alignment and for finding orthologs using the MultiParanoid algorithm (Alexeyenko*, et al.*, 2006).

5.3.2 Generation of groups of orthologous genes

Protein sequences for each plant species (Arabidopsis, Medicago, soybean, poplar, grapevine) was 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® DeCypher® system (Active Motif Inc., Carlsbad CA). This data was used as input for the InParanoid software (Remm*, et al.*, 2001) (version 3.0) with default settings. The InParanoid output (ortholog

tables) was used as input for the MultiParanoid software (Alexeyenko*, et al.*, 2006). For each retrieved OG, corresponding promoter sequences were retrieved using BioPerl (Stajich*, et al.*, 2002). Promoter sequence data corresponding to each OG was used as input for motif discovery using the Seeder algorithm (Fauteux*, et al.*, 2008).

5.3.3 Motif discovery in groups of orthologous promoters

Each of the 7,156 OPs with at least one promoter from each of the five dicotyledonous species (Arabidopsis, Medicago, soybean, poplar, grapevine) was submitted to motif discovery using the Seeder algorithm (Fauteux*, et al.*, 2008) using a *q*-value threshold of 0.05. The background sequence set was composed of the combined, whole set of promoters from the five plant species (204,735 promoters). Discovered motifs were clustered using the STAMP software (version 1.1, standalone) (Mahony and Benos, 2007). The average score for each motif in each OP was computed using standard procedures (Wasserman and Sandelin, 2004) implemented in custom Perl scripts.

5.3.4 Gene ontology annotation of groups of orthologous genes

Gene ontology annotations for Arabidopsis, poplar and grapevine were used to generate annotations for each OG. This was performed for each OG by counting

the number of occurrences of each ontology term, and then assigning the final annotation with the following precedence: (1) most abundant (if any) biological process, (2) most abundant (if any) molecular function and (3) most abundant (if any) cellular component. Groups of orthologous genes with no GO annotations were tagged with “unknown function”. Ontology terms were clustered by similarity using the Functional Similarity Matrix (Schlicker and Albrecht, 2008) and the functional similarity measure of Schlicker et al. (Schlicker*, et al.*, 2006). Hierarchical clustering of semantic similarities was performed with agglomerative nesting hierarchical clustering (agnes) (Kaufman and Rousseeuw, 1990) using the R statistical software (http://www.r-project.org).

5.3.5 Co-clustering meta-analysis of motifs and gene ontologies

Motifs discovered in OPs using the Seeder algorithm (Fauteux*, et al.*, 2008) were clustered using the STAMP software (Mahony and Benos, 2007). The average score for each motif in each OP was also computed. Three major groups of OGs with similar function, in which clusters of motifs were enriched, were identified using a two-dimensional kernel density estimation plot (R statistical software, library MASS, kde2d) and represented on a hybrid image-contour plot.

5.4 Results

5.4.1 Five dicotyledonous plant genomes share over 7,000 groups of orthologous genes

Groups of orthologous genes were identified among five dicotyledonous species, namely Arabidopsis, Medicago, soybean, poplar and grapevine. We retrieved 27,235, 38,728, 66,210, 45,555 and 30,434 protein sequences from Arabidopsis, Medicago, soybean, poplar and grapevine, respectively. Groups of orthologous genes (OGs) were identified using the MultiParanoid software (Alexeyenko*, et al.*, 2006) and an all-against-all Smith-Waterman alignment (Smith and Waterman, 1981) of protein sequences from each species. The number of OGs in pairs of plant species varies between 8,000 and 14,500, as shown in Table 5.1. A total of 7,156 OGs with at least one gene from each of the five plant species were identified, and this set was used for further analysis. Each OG contains in average 7.06 genes. Gene names and GO terms were assigned to each OG using annotation data from the five plant species.

5.4.2 Conserved DNA motifs are prevalent within groups of orthologous gene promoters

To explore conserved CRMs within promoters of dicotyledonous orthologous genes, collections of 27,234, 38,737, 66,155, 42,177 and 30,432 promoter

sequences were retrieved for Arabidopsis, Medicago, soybean, poplar and grapevine, respectively. From these sets, the corresponding promoter sequences (500 bp upstream of annotated mRNAs) were retrieved for each gene in the 7,156 OGs. The Seeder algorithm (Fauteux*, et al.*, 2008) was used to discover motifs within each of 7,156 Ops, using as a background the combined, whole set of promoters from the five species. A total of 1,773 significant motifs were discovered in 1,335 of the OPs. Motifs associated with the remaining 5,821 OPs were not statistically significant (q-value < 0.05).

5.4.3 Clusters of highly conserved motifs are discovered in groups of orthologous promoters

To investigate whether promoters of orthologous gene groups of similar function also contain conserved elements, a co-clustering of DNA motif similarity and GO semantic similarity was performed. DNA motifs discovered in OPs were clustered by similarity using the STAMP software (Mahony and Benos, 2007). The GO annotations corresponding to each OG were clustered by semantic similarity using the Functional Similarity Matrix (Schlicker and Albrecht, 2008) and the functional similarity measure of Schlicker et al. (2006). In addition, each group of orthologous gene promoters (OP) was scored with the PWM corresponding to each DNA motif discovered in OPs.

Figure 5.1 illustrates the results of this analysis. We have focused on three large ontology-based clusters of promoters (PCs) (1, 2 and 3) in which we found four enriched motif clusters (MCs) (A, B, C and D). Those co-clusters correspond to high-density regions identified using two-dimensional kernel density estimation (Supplementary Figure 3, Appendix 5). The first ontology-based cluster of promoters (PC 1) consists of 51 OPs (508 promoters) associated with genes involved in translation and translational elongation (Table 5.2). The corresponding OPs are enriched in two motif clusters (MC A and B) (Figure 5.1). The second ontology-based cluster of promoters (PC 2) consists of 23 OPs (169 promoters) associated with genes involved in DNA replication (Table 5.2). Those OPs are enriched in one motif cluster (MC D) (Figure 5.1). The third ontology-based cluster of promoters (PC 3) consists of 9 OPs (154 promoters) associated with genes involved in nucleosome assembly (Table 5.2) and their corresponding OPs are enriched in one motif cluster (MC C) (Figure 5.1).

Familial binding profiles (FBPs) were generated for each MC (Figure 5.1, MCs A-D) and were matched against the PLACE database (Higo*, et al.*, 1998) using the STAMP software (Mahony and Benos, 2007). Figure 5.2 shows the sequence logos corresponding to each familial binding profile (FBP). The FBP of MC A (enriched in promoter cluster 1, PC1) is highly similar (E=7.6e-15) to the telo-box motif (AACCCTAA) originally observed in the promoters of Arabidopsis

translation elongation factor genes (Axelos*, et al.*, 1989). The FBP of MC B (also enriched in promoter cluster 1, PC1) is similar (E=1.37e-10) to the site II element (TGGGCY), which is found in the promoters of Arabidopsis cytochrome c genes (Welchen and Gonzalez, 2005). Interestingly, this site is also associated with a telo-box motif in the Cytc-1 promoter (Welchen and Gonzalez, 2005). The FBP of MC D (enriched in promoter cluster 2, PC2) is highly similar (E=1.18e-10) to an E2F binding site (TTTCCCGC) (Chaboute*, et al.*, 2000). Finally, the FBP of MC (C) (enriched in promoter cluster 3, PC3) is highly similar (E= 5.93e-12) to the octamer motif, originally found in a wheat (*Triticum aestivum*) histone promoter (Nakayama*, et al.*, 1992).

5.4.4 Discovered motifs are available at DDOPM, a plant *cis*-regulatory motif database

Because the large dataset of orthologous genes and motifs is impractical to search in file format, we present all the data through a searchable interface, publicly available on the web (http://ddopm.agrenv.mcgill.ca), as shown in Figure 5.3. Motifs discovered in 1,335 OPs from the five dicotyledonous species constitute the core data of the DDOPM database. The user can query the database by DDOPM motif ID, which returns a page containing the annotations of the motif's parent OG, the motif's sequence logos (forward and reverse

complement) and PFM. The database can also be queried by keywords (such as 'transcription') or by GO ID, which returns a table with all OGs associated with the given keyword or ontology, with links to the motif page for each OG. Alternatively, a user-defined PFM can be matched against motifs in the database using STAMP (Mahony and Benos, 2007). The motifs can be downloaded in TRANSFAC®-like format, and OG annotations can be downloaded in tab-delimited format.

5.5 Discussion

In this work, the objective was to discover DNA motifs in promoters of orthologous genes across several dicotyledonous plant species. We have identified a total of 1,773 significant motifs in 1,335 OPs consisting of promoters from Arabidopsis, Medicago, soybean, poplar and grapevine. This data set covers a wide range of dicotyledonous TF binding profiles and presents a great complement to existing databases of plant *cis*-regulatory elements where consensus sequences still predominate and where a limited number of TF binding profiles are represented.

The largest clusters of promoters with conserved motifs discovered in the dicotyledonous OPs in this study are associated with promoters of genes coding for proteins involved in translation, DNA metabolism and nucleosome assembly.

This suggests that the regulation of such processes is highly conserved in dicotyledonous species. These motifs may in fact even be conserved beyond plant species. For example, the E2F TF is known to be conserved in animal and plant species, and the motif reported here as being enriched in promoters of genes involved in DNA replication (MC D) is highly similar to consensus sequences previously reported in both plant (Vandepoele*, et al.*, 2005) and animal species (Tao*, et al.*, 1997). In contrast, the telo-box motif (MC A), which we found in plant promoters from genes involved in translation, is also related to sequences found in the promoters of ribosomal protein coding genes in fungi, but in animal species this sequence is associated with the organization of telomeres (Hogues*, et al.*, 2008).

A number of motifs within the database show no close similarity with known, experimentally characterized *cis*-regulatory elements. This data and related cluster annotations is useful to researchers aiming to annotate and characterize novel *cis*-regulatory elements or motifs with no close similarity to known *cis*-regulatory elements. Large-scale discovery of motifs inherently implies some database redundancy, and several motifs may be associated with a single TFBS or a single familial binding profile (FBP). In the case where reducing redundancy is necessary, users are encouraged to use tools such as the scluster

software (Pape, *et al.*, 2008) to generate clusters of similar motifs within the DDOPM database.

The DDOPM database complements widely used resources such as the PLACE database (Higo, *et al.*, 1998). The PLACE database is a high-quality resource containing exclusively experimentally characterized *cis*-regulatory elements. However, because the motifs therein are consensus sequences and the coverage of plant *cis*-regulomes is currently limited, other resources for motif discovery and annotations are needed. The TRANSFAC® database (Wingender, *et al.*, 1996), on the other hand, contains both plant motif consensus sequences and matrices, but the number of plant motif matrices is rather small, less than 100 (http://www.gene-regulation.com/info/plant.html). The DDOPM database and online tools (http://ddopm.agrenv.mcgill.ca) are thus an essential complement to existing databases.

5.6 Conclusions

We have performed motif discovery in orthologous gene promoters across five dicotyledonous plant species. We have highlighted four clusters of motifs (MCs) associated with three clusters of promoters (PCs) associated with genes involved in fundamental cellular processes. Finally, we have generated a publicly

accessible database of 1,773 DNA motifs representing putative, conserved plant *cis*-regulatory elements.

5.7 Authors' contributions

FF and MVS designed the study. FF performed programming and data analysis. MVS supervised the project. Both authors have participated in writing the manuscript and have read and approved the final version.

5.8 Acknowledgements

The authors acknowledge the Natural Sciences and Engineering Research Council of Canada (NSERC) for a Discovery grant to M.V.S. and an NSERC Postgraduate Scholarship (PGS D) to F.F. We also acknowledge le Fonds de recherche sur la nature et les technologies (FQRNT) and Centre SÈVE for financial support.

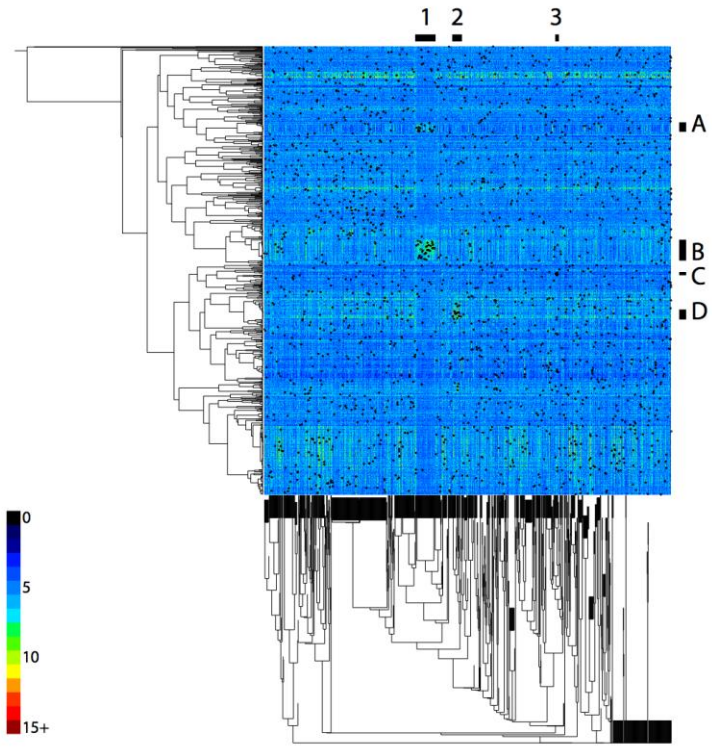


Figure 5.1: Co-clustering of transcription factor binding site similarity and function semantic similarity with a heat map of position-weight matrix scores of transcription factor binding sites on groups of orthologous promoters from five dicotyledonous plants.

Clusters of motifs (MCs, vertical axis) and clusters of promoters (PCs, horizontal axis) based on matrix similarity and semantic similarity, respectively. For each motif, the average of position-weight matrix scores in all sequences within one OP is reported on a heat map. Cold colors correspond to a low score, and warm colors correspond to a high score (see inset). Black circles correspond to significant promoter sequences/motif coordinates. (1-3), clusters of OPs (PCs); (A-D), clusters of motifs (MCs).

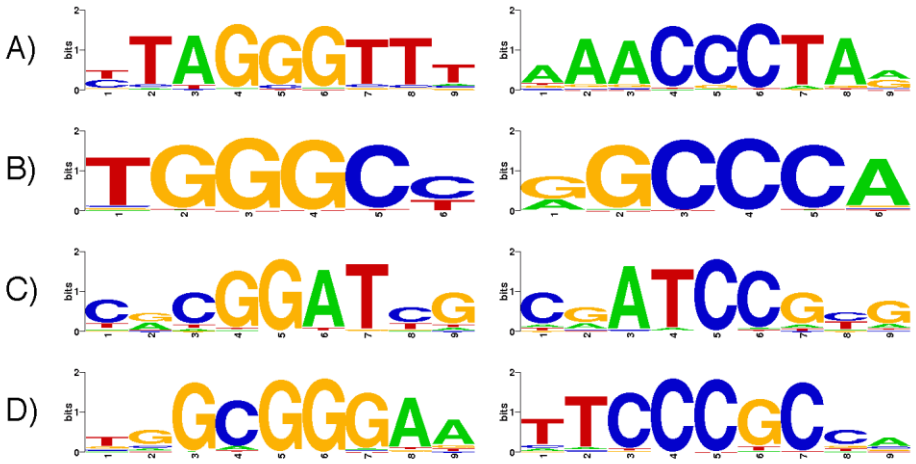


Figure 5.2: Familial binding profiles of motifs discovered in clusters of orthologous genes with similar gene function.

Sequence logos of four family binding profiles associated with DNA motifs discovered in groups of orthologous promoters. A) and B), (MC A and B) family binding profiles of the motifs discovered in a cluster of promoters of genes involved in translation (PC 1); C) (MC C) family binding profile of motifs discovered in a cluster of promoters of genes involved in nucleosome assembly (PC 3), D) (MC D) family binding profile of motifs discovered in a cluster of promoters of genes involved in DNA replication (PC 2). Forward and reverse complement motifs are on the left and right side, respectively.

Figure 5.3: The Database of Dicotyledonous Orthologous Promoter Motifs (DDOPM) web interface.

(A) From the main page users have two tool selections. (B) Selecting the "Database query" the user can search the database with a keyword, motif ID or GO ID. The annotation for that motif will be returned on the results page together with sequence logos generated on the fly. (C) Users' own motifs of interest can be matched with the annotated motifs in the database through the "Database matching" page.

Table 5.1: Number of groups of orthologous genes found in pairs of dicotyledonous plant species

	At	Mt	Gm	Pt
Mt	7,941			
Gm	12,582	10,797		
Pt	12,514	8,986	14,332	
Vv	11,369	8,126	12,815	13,074

At, Arabidopsis; Mt, Medicago; Gm, soybean; Pt, poplar; Vv, grapevine.

Table 5.2: Gene ontologies associated with clusters of groups of orthologous genes

Cluster	n. OGs	GO ID	GO term
1	47	GO:0006412	translation
	4	GO:0006414	translational elongation
2	9	GO:0006260	DNA replication
	3	GO:0015074	DNA integration
	3	GO:0006298	mismatch repair
	3	GO:0006270	DNA replication initiation
	1	GO:0032875	regulation of DNA endoreduplication
	1	GO:0006269	DNA replication, synthesis of RNA primer
	1	GO:0006268	DNA unwinding during replication
	1	GO:0006310	DNA recombination
	1	GO:0006265	DNA topological change
3	9	GO:0006334	nucleosome assembly

Cluster, cluster ID; n. OGs, number of OGs within cluster associated with the gene ontology; GO ID, gene ontology identification; GO term, gene ontology term.

6. Discussion

Plant biotechnology researchers and developers need greater and more accurate knowledge of plant promoters and plant CREs. The experimental identification of CREs is laborious, and computational methods may facilitate this task by identifying putative candidates for further experimental validation. In this thesis, my main objective was to develop and apply an algorithm for fast and reliable DNA motif discovery in plant promoters.

My first hypothesis was that computational methods for DNA motif discovery could be improved by enumerative discriminative seeding. To test this hypothesis, the objective was to design and benchmark an exact discriminative seeding DNA motif discovery algorithm. In Chapter 3, the design, software implementation and benchmarks of the Seeder algorithm is reported. The algorithm outperformed other tools on biological benchmark data, which confirmed the first hypothesis. The performance of the algorithm over other tools was attributed to the judicious combination of word enumeration for seed selection and PWM scoring for site selection, and the use of an empirical background model and reliable statistics for motif seed selection.

My second hypothesis was that a data structure, based on the geometry of the similarity matrix between combinations of nucleotide symbols, could

accelerate enumerative DNA motif discovery. To test this hypothesis, the second objective was to create the data structure and to evaluate its performance in accelerating DNA motif discovery. In Chapter 3, the design and software implementation the SMD index is reported. The use of this data structure accelerated background and motifs computations by several orders of magnitude, which confirmed the second hypothesis. The primer to this design was the observation that HDs between sorted lists of words were characterized by a square recursive fractal geometry, which was successfully deciphered for the mathematical formulation and efficient implementation of the SMD index.

My third research hypothesis was that plant tissue-specific gene promoters contain conserved CRMs in related plant species, and different combinations of motifs in different plant families. To test this hypothesis, the objective was to perform motif discovery in plant seed-specific promoters, and to compare conserved motifs in different plant families. In Chapter 4, the Seeder algorithm was used for an in-depth analysis of 54 SSP gene promoters from three plant families, namely *Brassicaceae*, *Fabaceae* and *Poaceae* using backgrounds based on complete sets of promoters from a representative species in each family. Three conserved motifs (two RY-like and one ACGT-like) were discovered in *Brassicaceae* and *Fabaceae* SSP gene promoters. A novel, seed-specific E2Fb-like motif was also identified in *Fabaceae*. Conserved motifs

identified in *Poaceae* SSP gene promoters included a GCN4-like motif, two prolamin-box-like motifs and an Skn-1-like motif. Evidence of the presence of a variant of the TATA-box was found in the SSP gene promoters from the three plant families. Altogether, these results confirmed the third hypothesis, since conserved motifs were found in the promoters of seed-specific genes in related species (within each plant family), but different sets of motifs were found in different plant families.

My fourth hypothesis was that the promoters of dicotyledonous orthologous genes contain conserved CRMs. To test this hypothesis, the objective was to perform motif discovery in the promoters of dicotyledonous orthologous genes. In Chapter 5, the large-scale analysis of thousands of OPs among five dicotyledonous species is reported. A total of 1,773 significant motifs were discovered; this result strongly confirms the fourth hypothesis. The co-clustering analysis of DNA motif similarity and GO semantic similarity further allowed the identification of major clusters of conserved motifs, found in the promoters of genes involved in fundamental cellular processes. In Chapter 5, an online database resource is also presented, the DDOPM. This database is available for query and for download on the Internet, and constitutes a great reference for the annotation of plant CRMs.

7. Future research directions

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

1. The Seeder algorithm would benefit from elimination of the motif width parameter, which could be automatically adjusted step-wise by recursively extending the seed matrix and evaluating the statistical significance of the extended matrix.
2. The Seeder algorithm could be improved for more versatility in the motif models, *e.g.* to include more than one match per sequence. This will however require a thorough restructuration of the algorithm and related data structures.
3. Motifs found in dicotyledonous promoters could be tested experimentally in monocotyledonous promoters and vice-versa, and carefully examined for relative strength in the embryo *vs.* the endosperm. It would also be interesting to test such promoters and motifs in endospermic dicotyledonous species such as guar (*Cyamopsis tetragonolobus*).
4. The new E2Fb-like motif discovered in *Fabaceae* promoters needs to be validated experimentally.

5. It would be very interesting to test different designs of synthetic promoters based on motifs identified in SSPs, and to evaluate their relative strength and tissue specificity.
6. As more species are sequenced, the DDOPM will have to be extended, for example to include a monocotyledonous plant species division.
7. The analysis of motifs conserved in dicotyledonous orthologous promoters could be improved by a more thorough analysis of paralogs, and by going deeper in the analysis of gene function in relation with promoter motif composition and experimentally characterized CREs.

8. Contribution to knowledge

8.1 Contributions from Chapter 3

- The discriminative seeding approach to DNA motif discovery implemented in the Seeder algorithm significantly improved the detection and statistical evaluation of conserved motifs in eukaryotic promoter sequences.
- The SMD index, a data structure based on the mathematical properties of ordered lists of combinations of nucleotide symbols, accelerated enumerative DNA motif discovery computations by several orders of magnitude.
- Two seed-specific motifs were identified in Arabidopsis seed-specific marker genes.

8.2 Contributions from Chapter 4

- Quantitative models were generated for known seed-specific motifs conserved in SSP promoters from three plant families, and were shown to be an accurate signature of SSP gene promoters.
- A new, E2Fb-like motif was discovered in *Fabaceae*.
- Important differences were highlighted in SSP promoter organization between monocotyledonous and dicotyledonous plant species.

8.3 Contributions from Chapter 5

- Groups of orthologous genes were identified in five dicotyledonous plant species.
- Over 1,700 significant conserved motifs were identified in groups of orthologous promoters from five dicotyledonous plant species.
- The clustering analysis of gene functions and motif similarities highlighted the presence of three large clusters of motifs found in promoters of genes involved in fundamental cellular processes.
- A database resource, constituting a large reference for the exploration and the annotation of plant CRMs, was developed and released on the internet.

List of references

Adams, K. L., and Wendel, J. F. (2005). Polyploidy and genome evolution in plants. *Curr Opin Plant Biol, 8*(2), 135-141.

AGI. (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature, 408*(6814), 796-815.

Alexeyenko, A., Tamas, I., Liu, G., and Sonnhammer, E. L. L. (2006). Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics, 22*(14), E9-E15.

Allen, R. D., Bernier, F., Lessard, P. A., and Beachy, R. N. (1989). Nuclear factors interact with a soybean beta-conglycinin enhancer. *Plant Cell, 1*(6), 623-631.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol, 215*(3), 403-410.

Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., et al. (2004). UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res, 32*(Database issue), D115-119.

Axelos, M., Bardet, C., Liboz, T., Le Van Thai, A., Curie, C., and Lescure, B. (1989). The gene family encoding the Arabidopsis thaliana translation elongation factor EF-1 alpha: molecular cloning, characterization and expression. *Mol Gen Genet, 219*(1-2), 106-112.

Bailey, T. L., and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol, 2*, 28-36.

Barta, E., Sebestyen, E., Palfy, T. B., Toth, G., Ortutay, C. P., and Patthy, L. (2005). DoOP: Databases of Orthologous Promoters, collections of clusters of orthologous upstream sequences from chordates and plants. *Nucleic Acids Res, 33*(Database issue), D86-90.

Baud, S., bastien, Dubreucq, B., Miquel, M., Rochat, C., and Lepiniec, L. (2008). Storage Reserve Accumulation in Arabidopsis: Metabolic and Developmental Control of Seed Filling. *The Arabidopsis Book*.

Baumlein, H., Nagy, I., Villarroel, R., Inze, D., and Wobus, U. (1992). Cis-analysis of a seed protein gene promoter: the conservative RY repeat CATGCATG within the legumin box is essential for tissue-specific expression of a legumin gene. *Plant J, 2*(2), 233-239.

Baumlein, H., Wobus, U., Pustell, J., and Kafatos, F. C. (1986). The legumin gene family: structure of a B type gene of Vicia faba and a possible legumin gene specific regulatory element. *Nucleic Acids Res, 14*(6), 2707-2720.

Beccari, J. B. (1745). De frumento. De bononiensi scientarium et artium. *Instituto atque Academia Commentarii, Bologna, 2*, 122-127.

Benoist, C., and Chambon, P. (1981). In vivo sequence requirements of the SV40 early promotor region. *Nature, 290*(5804), 304-310.

Blackwood, E. M., and Kadonaga, J. T. (1998). Going the distance: a current view of enhancer action. *Science, 281*(5373), 60-63.

Bode, J., Goetze, S., Heng, H., Krawetz, S. A., and Benham, C. (2003). From DNA structure to gene expression: mediators of nuclear compartmentalization and dynamics. *Chromosome Res, 11*(5), 435-445.

Bown, D., Ellis, T. H., and Gatehouse, J. A. (1988). The sequence of a gene encoding convicilin from pea (Pisum sativum L.) shows that convicilin differs from vicilin by an insertion near the N-terminus. *Biochem J, 251*(3), 717-726.

Brzeski, J., and Jerzmanowski, A. (2004). Plant chromatin -- epigenetics linked to ATP-dependent remodeling and architectural proteins. *FEBS Lett, 567*(1), 15-19.

Bulyk, M. L., Johnson, P. L., and Church, G. M. (2002). Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors. *Nucleic Acids Res, 30*(5), 1255-1261.

Butler, J. E., and Kadonaga, J. T. (2002). The RNA polymerase II core promoter: a key component in the regulation of gene expression. *Genes Dev, 16*(20), 2583-2592.

Chaboute, M. E., Clement, B., Sekine, M., Philipps, G., and Chaubet-Gigot, N. (2000). Cell cycle regulation of the tobacco ribonucleotide reductase small subunit gene is mediated by E2F-like elements. *Plant Cell, 12*(10), 1987-2000.

Chamberland, S., Daigle, N., and Bernier, F. (1992). The legumin boxes and the 3' part of a soybean beta-conglycinin promoter are involved in seed gene expression in transgenic tobacco plants. *Plant Mol Biol, 19*(6), 937-949.

Chandrasekharan, M. B., Bishop, K. J., and Hall, T. C. (2003). Module-specific regulation of the beta-phaseolin promoter during embryogenesis. *Plant J, 33*(5), 853-866.

Chen, Z. L., Pan, N. S., and Beachy, R. N. (1988). A DNA sequence element that confers seed-specific enhancement to a constitutive promoter. *EMBO J, 7*(2), 297-302.

Chen, Z. L., Schuler, M. A., and Beachy, R. N. (1986). Functional analysis of regulatory elements in a plant embryo-specific gene. *Proc Natl Acad Sci U S A, 83*(22), 8560-8564.

Claverie, J. M., and Audic, S. (1996). The statistical significance of nucleotide position-weight matrix matches. *Comput Appl Biosci, 12*(5), 431-439.

Cornish-Bowden, A. (1985). Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Res, 13*(9), 3021-3030.

D'Haeseleer, P. (2006). How does DNA sequence motif discovery work? *Nat Biotechnol, 24*(8), 959-961.

Depigny-This, D., Raynal, M., Aspart, L., Delseny, M., and Grellet, F. (1992). The cruciferin gene family in radish. *Plant Mol Biol, 20*(3), 467-479.

Derose, R. T., Ma, D. P., Kwon, I. S., Hasnain, S. E., Klassy, R. C., and Hall, T. C. (1989). Characterization of the Kafirin Gene Family from Sorghum Reveals Extensive Homology with Zein from Maize. *Plant Molecular Biology, 12*(3), 245-256.

Dickinson, C. D., Evans, R. P., and Nielsen, N. C. (1988). RY repeats are conserved in the 5'-flanking regions of legume seed-protein genes. *Nucleic Acids Res, 16*(1), 371.

Doyle, J. J., Schuler, M. A., Godette, W. D., Zenger, V., Beachy, R. N., and Slightom, J. L. (1986). The glycosylated seed storage proteins of Glycine max and Phaseolus vulgaris. Structural homologies of genes and proteins. *J Biol Chem, 261*(20), 9228-9238.

Dynan, W. S., and Tjian, R. (1985). Control of eukaryotic messenger RNA synthesis by sequence-specific DNA-binding proteins. *Nature, 316*(6031), 774-778.

Ellerstrom, M., Stalberg, K., Ezcurra, I., and Rask, L. (1996). Functional dissection of a napin gene promoter: identification of promoter elements required for embryo and endosperm-specific transcription. *Plant Mol Biol, 32*(6), 1019-1027.

Elliott, K. A., and Shirsat, A. H. (1998). Promoter regions of the extA extensin gene from Brassica napus control activation in response to wounding and tensile stress. *Plant Mol Biol, 37*(4), 675-687.

Elnitski, L., Jin, V. X., Farnham, P. J., and Jones, S. J. (2006). Locating mammalian transcription factor binding sites: a survey of computational and experimental techniques. *Genome Res, 16*(12), 1455-1464.

Ericson, M. L., Muren, E., Gustavsson, H. O., Josefsson, L. G., and Rask, L. (1991). 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*(3), 741-746.

Eyal, Y., Curie, C., and McCormick, S. (1995). Pollen specificity elements reside in 30 bp of the proximal promoters of two pollen-expressed genes. *Plant Cell, 7*(3), 373-384.

Ezcurra, I., Ellerstrom, M., Wycliffe, P., Stalberg, K., and Rask, L. (1999). 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*(4), 699-709.

Fauteux, F., Blanchette, M., and Stromvik, M. V. (2008). Seeder: Discriminative Seeding DNA Motif Discovery. *Bioinformatics, 24*(20), 2303-2307.

Forde, B. G., Heyworth, A., Pywell, J., and Kreis, M. (1985). Nucleotide sequence of a B1 hordein gene and the identification of possible upstream regulatory elements in endosperm storage protein genes from barley, wheat and maize. *Nucleic Acids Res, 13*(20), 7327-7339.

Fujiwara, T., and Beachy, R. N. (1994). Tissue-specific and temporal regulation of a beta-conglycinin gene: roles of the RY repeat and other cis-acting elements. *Plant Mol Biol, 24*(2), 261-272.

Garner, M. M., and Revzin, A. (1981). A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system. *Nucleic Acids Res, 9*(13), 3047-3060.

Gatehouse, J. A., Bown, D., Gilroy, J., Levasseur, M., Castleton, J., and Ellis, T. H. (1988). Two genes encoding 'minor' legumin polypeptides in pea (Pisum sativum L.). Characterization and complete sequence of the LegJ gene. *Biochem J, 250*(1), 15-24.

Gerstein, M. B., Bruce, C., Rozowsky, J. S., Zheng, D., Du, J., Korbel, J. O., et al. (2007). What is a gene, post-ENCODE? History and updated definition. *Genome Res, 17*(6), 669-681.

Goff, S. A., Ricke, D., Lan, T. H., Presting, G., Wang, R., Dunn, M., et al. (2002). A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). *Science, 296*(5565), 92-100.

Gordon, D. M. (1998). A survey of fast exponentiation methods. *Journal of Algorithms, 27*(1), 129-146.

Grinstead, C. M., and Snell, J. L. (1997). Sums of Random Variables *Introduction to Probability* (pp. 285-304). Providence, RI: American Mathematical Society.

Gruss, P., Dhar, R., and Khoury, G. (1981). Simian virus 40 tandem repeated sequences as an element of the early promoter. *Proc Natl Acad Sci U S A, 78*(2), 943-947.

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

Guilfoyle, T. (1997). The structure of plant gene promoters. In J. K. Setlow (Ed.), *Genetic Engineering: Principles and Methods* (Vol. 19, pp. 15-47). New-York: Plenum Press.

Guo, A. Y., Chen, X., Gao, G., Zhang, H., Zhu, Q. H., Liu, X. C., et al. (2008). PlantTFDB: a comprehensive plant transcription factor database. *Nucleic Acids Res, 36*(Database issue), D966-969.

Gutierrez, L., Van Wuytswinkel, O., Castelain, M., and Bellini, C. (2007). Combined networks regulating seed maturation. *Trends Plant Sci, 12*(7), 294-300.

Hahn, S. (2004). Structure and mechanism of the RNA polymerase II transcription machinery. *Nat Struct Mol Biol, 11*(5), 394-403.

Hamming, R. W. (1950). Error Detecting and Error Correcting Codes. *Bell System Technical Journal, 29*(2), 147-160.

Harada, J. J., Barker, S. J., and Goldberg, R. B. (1989). Soybean beta-conglycinin genes are clustered in several DNA regions and are regulated by transcriptional and posttranscriptional processes. *Plant Cell, 1*(4), 415-425.

Heintzman, N. D., and Ren, B. (2007). The gateway to transcription: identifying, characterizing and understanding promoters in the eukaryotic genome. *Cell Mol Life Sci, 64*(4), 386-400.

Heintzman, N. D., Stuart, R. K., Hon, G., Fu, Y., Ching, C. W., Hawkins, R. D., et al. (2007). Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet, 39*(3), 311-318.

Hertz, G. Z., and Stormo, G. D. (1999). Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics, 15*(7-8), 563-577.

Higo, K., Ugawa, Y., Iwamoto, M., and Higo, H. (1998). PLACE: a database of plant cis-acting regulatory DNA elements. *Nucleic Acids Res, 26*(1), 358-359.

Hogues, H., Lavoie, H., Sellam, A., Mangos, M., Roemer, T., Purisima, E., et al. (2008). Transcription factor substitution during the evolution of fungal ribosome regulation. *Mol Cell, 29*(5), 552-562.

Huang, H., Tudor, M., Weiss, C. A., Hu, Y., and Ma, H. (1995). The Arabidopsis MADS-box gene AGL3 is widely expressed and encodes a sequence-specific DNA-binding protein. *Plant Mol Biol, 28*(3), 549-567.

Hyndman, R., and Fan, Y. (1996). Sample quantiles in statistical packages. *The American Statistician, v50*(n4), p361(365).

Inze, D., and Veylder, L. D. (2006). Cell Cycle Regulation in Plant Development. *Annu Rev Genet*.

Istrail, S., and Davidson, E. H. (2005). Logic functions of the genomic cis-regulatory code. *Proc Natl Acad Sci U S A, 102*(14), 4954-4959.

Itoh, Y., Kitamura, Y., Arahira, M., and Fukazawa, C. (1993). cis-acting regulatory regions of the soybean seed storage 11S globulin gene and their interactions with seed embryo factors. *Plant Mol Biol, 21*(6), 973-984.

Itoh, Y., Kitamura, Y., and Fukazawa, C. (1994). The glycinin box: a soybean embryo factor binding motif within the quantitative regulatory region of the 11S seed storage globulin promoter. *Mol Gen Genet, 243*(3), 353-357.

Jacob, F., and Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol, 3*, 318-356.

Jaillon, O., Aury, J. M., Noel, B., Policriti, A., Clepet, C., Casagrande, A., et al. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature, 449*(7161), 463-467.

Jensen, E. O., Marcker, K. A., Schell, J., and Bruijn, F. J. (1988). Interaction of a nodule specific, trans-acting factor with distinct DNA elements in the soybean leghaemoglobin lbc(3) 5' upstream region. *Embo J, 7*(5), 1265-1271.

Josefsson, L. G., Lenman, M., Ericson, M. L., and Rask, L. (1987). Structure of a gene encoding the 1.7 S storage protein, napin, from Brassica napus. *J Biol Chem, 262*(25), 12196-12201.

Kaufman, L., and Rousseeuw, P. J. (1990). *Finding groups in data : an introduction to cluster analysis*. New York: Wiley.

Keich, U., and Pevzner, P. A. (2002). Finding motifs in the twilight zone. *Bioinformatics, 18*(10), 1374-1381.

Keller, B., and Heierli, D. (1994). Vascular expression of the grp1.8 promoter is controlled by three specific regulatory elements and one unspecific activating sequence. *Plant Mol Biol, 26*(2), 747-756.

Khoury, G., and Gruss, P. (1983). Enhancer elements. *Cell, 33*(2), 313-314.

Kitamura, Y., Arahira, M., Itoh, Y., and Fukazawa, C. (1990). The complete nucleotide sequence of soybean glycinin A2B1a gene spanning to another glycinin gene A1aB1b. *Nucleic Acids Res, 18*(14), 4245.

Klepper, K., Sandve, G. K., Abul, O., Johansen, J., and Drablos, F. (2008). Assessment of composite motif discovery methods. *BMC Bioinformatics, 9*, 123.

Klinedinst, S., Pascuzzi, P., Redman, J., Desai, M., and Arias, J. (2000). A xenobiotic-stress-activated transcription factor and its cognate target genes are preferentially expressed in root tip meristems. *Plant Mol Biol, 42*(5), 679-688.

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., et al. (2007). Clustal W and Clustal X version 2.0. *Bioinformatics, 23*(21), 2947-2948.

Lawrence, C. E., Altschul, S. F., Boguski, M. S., Liu, J. S., Neuwald, A. F., and Wootton, J. C. (1993). Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science, 262*(5131), 208-214.

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

Lelievre, J. M., Oliveira, L. O., and Nielsen, N. C. (1992). 5'CATGCAT-3' Elements Modulate the Expression of Glycinin Genes. *Plant Physiol, 98*(1), 387-391.

Lemon, B., and Tjian, R. (2000). Orchestrated response: a symphony of transcription factors for gene control. *Genes Dev, 14*(20), 2551-2569.

Lessard, P. A., Allen, R. D., Bernier, F., Crispino, J. D., Fujiwara, T., and Beachy, R. N. (1991). Multiple nuclear factors interact with upstream sequences of differentially regulated beta-conglycinin genes. *Plant Mol Biol, 16*(3), 397-413.

Lessard, P. A., Allen, R. D., Fujiwara, T., and Beachy, R. N. (1993). Upstream regulatory sequences from two beta-conglycinin genes. *Plant Mol Biol, 22*(5), 873-885.

Lindstrom, J. T., Vodkin, L. O., Harding, R. W., and Goeken, R. M. (1990). Expression of soybean lectin gene deletions in tobacco. *Dev Genet, 11*(2), 160-167.

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

Luo, H., Song, F., Goodman, R. M., and Zheng, Z. (2005). Up-regulation of OsBIHD1, a rice gene encoding BELL homeodomain transcriptional factor, in disease resistance responses. *Plant Biol (Stuttg), 7*(5), 459-468.

Mahony, S., and Benos, P. V. (2007). STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res, 35*(Web Server issue), W253-258.

Meinke, D. W., Chen, J., and Beachy, R. N. (1981). Expression of Storage-Protein Genes during Soybean Seed Development. *Planta, 153*(2), 130-139.

Mena, M., Vicente-Carbajosa, J., Schmidt, R. J., and Carbonero, P. (1998). An endosperm-specific DOF protein from barley, highly conserved in wheat, binds to and activates transcription from the prolamin-box of a native B-hordein promoter in barley endosperm. *Plant J, 16*(1), 53-62.

Mizukami, Y., Huang, H., Tudor, M., Hu, Y., and Ma, H. (1996). Functional domains of the floral regulator AGAMOUS: characterization of the DNA binding domain and analysis of dominant negative mutations. *Plant Cell, 8*(5), 831-845.

Molina, C., and Grotewold, E. (2005). Genome wide analysis of Arabidopsis core promoters. *BMC Genomics, 6*(1), 25.

Moreno-Risueno, M. A., Gonzalez, N., Diaz, I., Parcy, F., Carbonero, P., and Vicente-Carbajosa, J. (2008). FUSCA3 from barley unveils a common transcriptional regulation of seed-specific genes between cereals and Arabidopsis. *Plant J, 53*(6), 882-894.

Morton, R. L., Quiggin, D., and Higgins, T. J. V. (1995). Regulation of seed storage protein gene expression. In J. Kigel & G. Galili (Eds.), *Seed Development and Germination* (pp. 103–138). New-York, NY: Marcel Dekker Inc.

Muller, M., and Knudsen, S. (1993). The nitrogen response of a barley C-hordein promoter is controlled by positive and negative regulation of the GCN4 and endosperm box. *Plant J, 4*(2), 343-355.

Nakayama, T., Sakamoto, A., Yang, P., Minami, M., Fujimoto, Y., Ito, T., et al. (1992). Highly conserved hexamer, octamer and nonamer motifs are positive cis-regulatory elements of the wheat histone H3 gene. *FEBS Lett, 300*(2), 167-170.

Newbigin, E. J., Delumen, B. O., Chandler, P. M., Gould, A., Blagrove, R. J., March, J. F., et al. (1990). Pea Convicilin - Structure and Primary Sequence of the Protein and Expression of a Gene in the Seeds of Transgenic Tobacco. *Planta, 180*(4), 461-470.

Nielsen, N. C., Dickinson, C. D., Cho, T. J., Thanh, V. H., Scallon, B. J., Fischer, R. L., et al. (1989). Characterization of the glycinin gene family in soybean. *Plant Cell, 1*(3), 313-328.

O'Connor, T. R., Dyreson, C., and Wyrick, J. J. (2005). Athena: a resource for rapid visualization and systematic analysis of Arabidopsis promoter sequences. *Bioinformatics, 21*(24), 4411-4413.

Ogawa, M., Hanada, A., Yamauchi, Y., Kuwahara, A., Kamiya, Y., and Yamaguchi, S. (2003). Gibberellin biosynthesis and response during Arabidopsis seed germination. *Plant Cell, 15*(7), 1591-1604.

Orphanides, G., and Reinberg, D. (2002). A unified theory of gene expression. *Cell, 108*(4), 439-451.

Osborne, T. B. (1924). *The Vegetable Proteins*. London: Longmans, Green.

Palaniswamy, S. K., James, S., Sun, H., Lamb, R. S., Davuluri, R. V., and Grotewold, E. (2006). AGRIS and AtRegNet. a platform to link cis-regulatory elements and transcription factors into regulatory networks. *Plant Physiol, 140*(3), 818-829.

Pandey, S. P., and Krishnamachari, A. (2006). Computational analysis of plant RNA Pol-II promoters. *Biosystems, 83*(1), 38-50.

Pape, U. J., Rahmann, S., and Vingron, M. (2008). Natural similarity measures between position frequency matrices with an application to clustering. *Bioinformatics, 24*(3), 350-357.

Paterson, A. H., Bowers, J. E., Bruggmann, R., Dubchak, I., Grimwood, J., Gundlach, H., et al. (2009). The Sorghum bicolor genome and the diversification of grasses. *Nature, 457*(7229), 551-556.

Pavesi, G., Mereghetti, P., Mauri, G., and Pesole, G. (2004). Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Res, 32*(Web Server issue), W199-203.

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

Pedersen, K., Devereux, J., Wilson, D. R., Sheldon, E., and Larkins, B. A. (1982). Cloning and Sequence-Analysis Reveal Structural Variation among Related Zein Genes in Maize. *Cell, 29*(3), 1015-1026.

Pizzi, C., Bortoluzzi, S., Bisognin, A., Coppe, A., and Danieli, G. A. (2005). Detecting seeded motifs in DNA sequences. *Nucleic Acids Res, 33*(15), e135.

Potenza, C., Aleman, L., and Sengupta-Gopalan, C. (2004). Targeting transgene expression research, agricultural, and environmental applications: promoters used in plant transformation. *In vitro cellular & developmental biology - Plant, 40*(1), 1-22.

Rafalski, J. A. (1986). Structure of wheat gamma-gliadin genes. *Gene, 43*(3), 221-229.

Remm, M., Storm, C. E., and Sonnhammer, E. L. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J Mol Biol, 314*(5), 1041-1052.

Ren, B., Robert, F., Wyrick, J. J., Aparicio, O., Jennings, E. G., Simon, I., et al. (2000). Genome-wide location and function of DNA binding proteins. *Science, 290*(5500), 2306-2309.

Rerie, W. G., Whitecross, M. I., and Higgins, T. J. V. (1990). Nucleotide-Sequence of an a-Type Legumin Gene from Pea. *Nucleic Acids Research, 18*(3), 655-655.

Richmond, T. J., and Davey, C. A. (2003). The structure of DNA in the nucleosome core. *Nature, 423*(6936), 145-150.

Riechmann, J. L., Heard, J., Martin, G., Reuber, L., Jiang, C., Keddie, J., et al. (2000). Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. *Science, 290*(5499), 2105-2110.

Rodin, J., Sjodahl, S., Josefsson, L. G., and Rask, L. (1992). Characterization of a Brassica napus gene encoding a cruciferin subunit: estimation of sizes of cruciferin gene families. *Plant Mol Biol, 20*(3), 559-563.

Rombauts, S., Florquin, K., Lescot, M., Marchal, K., Rouze, P., and van de Peer, Y. (2003). Computational approaches to identify promoters and cis-regulatory elements in plant genomes. *Plant Physiol, 132*(3), 1162-1176.

Ryan, A. J., Royal, C. L., Hutchinson, J., and Shaw, C. H. (1989). Genomic sequence of a 12S seed storage protein from oilseed rape (Brassica napus c.v. jet neuf). *Nucleic Acids Res, 17*(9), 3584.

Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W. W., and Lenhard, B. (2004). JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res, 32*(Database issue), D91-94.

Sandve, G. K., Abul, O., Walseng, V., and Drablos, F. (2007). Improved benchmarks for computational motif discovery. *BMC Bioinformatics, 8*, 193.

Santos-Mendoza, M., Dubreucq, B., Baud, S., Parcy, F., Caboche, M., and Lepiniec, L. (2008). Deciphering gene regulatory networks that control seed development and maturation in Arabidopsis. *Plant J, 54*(4), 608-620.

Scheets, K., and Hedgcoth, C. (1988). Nucleotide-Sequence of a Gamma-Gliadin Gene - Comparisons with Other Gamma-Gliadin Sequences Show the Structure of Gamma-Gliadin Genes and the General Primary Structure of Gamma-Gliadins. *Plant Science, 57*(2), 141-150.

Schlicker, A., and Albrecht, M. (2008). FunSimMat: a comprehensive functional similarity database. *Nucleic Acids Res, 36*(Database issue), D434-439.

Schlicker, A., Domingues, F. S., Rahnenfuhrer, J., and Lengauer, T. (2006). A new measure for functional similarity of gene products based on Gene Ontology. *BMC Bioinformatics, 7*, 302.

Schmid, M., Davison, T. S., Henz, S. R., Pape, U. J., Demar, M., Vingron, M., et al. (2005). A gene expression map of Arabidopsis thaliana development. *Nature Genetics, 37*(5), 501-506.

Schneider, T. D. (2002). Consensus sequence Zen. *Appl Bioinformatics, 1*(3), 111-119.

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

Shannon, C. (1948). A mathematical theory of communication. *The Bell System Technical Journal, 27*, 379-423.

Shewry, P. R., Jones, H. D., and Halford, N. G. (2008). Plant biotechnology: transgenic crops. *Adv Biochem Eng Biotechnol, 111*, 149-186.

Shewry, P. R., Napier, J. A., and Tatham, A. S. (1995). Seed storage proteins: structures and biosynthesis. *Plant Cell, 7*(7), 945-956.

Shirsat, A., Wilford, N., Croy, R., and Boulter, D. (1989). Sequences responsible for the tissue specific promoter activity of a pea legumin gene in tobacco. *Mol Gen Genet, 215*(2), 326-331.

Shiu, S. H., Shih, M. C., and Li, W. H. (2005). Transcription factor families have much higher expansion rates in plants than in animals. *Plant Physiol, 139*(1), 18-26.

Sims, T. L., and Goldberg, R. B. (1989). The glycinin Gy1 gene from soybean. *Nucleic Acids Res, 17*(11), 4386.

Sinha, S. (2003). Discriminative motifs. *J Comput Biol, 10*(3-4), 599-615.

Sinha, S. (2006). On counting position weight matrix matches in a sequence, with application to discriminative motif finding. *Bioinformatics, 22*(14), e454-463.

Smale, S. T., and Kadonaga, J. T. (2003). The RNA polymerase II core promoter. *Annu Rev Biochem, 72*, 449-479.

Smith, A. D., Sumazin, P., and Zhang, M. Q. (2005). Identifying tissue-selective transcription factor binding sites in vertebrate promoters. *Proc Natl Acad Sci U S A, 102*(5), 1560-1565.

Smith, S. W. (1997). Convolution *The Scientist and Engineer's Guide to Digital Signal Processing* (pp. 107-122). San Diego, CA: California Technical Publishing.

Smith, T. F., and Waterman, M. S. (1981). Identification of common molecular subsequences. *J Mol Biol, 147*(1), 195-197.

So, J. S., and Larkins, B. A. (1991). Binding of an endosperm-specific nuclear protein to a maize beta-zein gene correlates with zein transcriptional activity. *Plant Mol Biol, 17*(3), 309-319.

Staden, R. (1984). Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Res, 12*(1 Pt 2), 505-519.

Staden, R. (1989). Methods for calculating the probabilities of finding patterns in sequences. *Comput Appl Biosci, 5*(2), 89-96.

Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., et al. (2002). The Bioperl toolkit: Perl modules for the life sciences. *Genome Res, 12*(10), 1611-1618.

Stalberg, K., Ellerstom, M., Ezcurra, I., Ablov, S., and Rask, L. (1996). Disruption of an overlapping E-box/ABRE motif abolished high transcription of the napA storage-protein promoter in transgenic Brassica napus seeds. *Planta, 199*(4), 515-519.

Steffens, N. O., Galuschka, C., Schindler, M., Bulow, L., and Hehl, R. (2004). AthaMap: an online resource for in silico transcription factor binding sites in the Arabidopsis thaliana genome. *Nucleic Acids Res, 32*(Database issue), D368-372.

Sticklen, M. B. (2008). Plant genetic engineering for biofuel production: towards affordable cellulosic ethanol. *Nat Rev Genet, 9*(6), 433-443.

Storey, J. D., and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A, 100*(16), 9440-9445.

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

Stormo, G. D., and Fields, D. S. (1998). Specificity, free energy and information content in protein-DNA interactions. *Trends Biochem Sci, 23*(3), 109-113.

Stormo, G. D., Schneider, T. D., Gold, L., and Ehrenfeucht, A. (1982). Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. *Nucleic Acids Res, 10*(9), 2997-3011.

Streatfield, S. J. (2007). Approaches to achieve high-level heterologous protein production in plants. *Plant Biotechnol J, 5*(1), 2-15.

Sumner-Smith, M., Rafalski, J. A., Sugiyama, T., Stoll, M., and Soll, D. (1985). Conservation and variability of wheat alpha/beta-gliadin genes. *Nucleic Acids Res, 13*(11), 3905-3916.

Sundt, B., and Dickson, D. C. M. (2000). Comparison of methods for evaluation of the n-fold convolution of an arithmetic distribution. *Bulletin of the Association of Swiss Actuaries*, 129–140.

Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R., and Wu, C. H. (2007). UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics, 23*(10), 1282-1288.

Takaiwa, F., Yamanouchi, U., Yoshihara, T., Washida, H., Tanabe, F., Kato, A., et al. (1996). Characterization of common cis-regulatory elements responsible for the endosperm-specific expression of members of the rice glutelin multigene family. *Plant Mol Biol, 30*(6), 1207-1221.

Takei, Y., Yamauchi, D., and Minamikawa, T. (1989). Nucleotide sequence of the canavalin gene from Canavalia gladiata seeds. *Nucleic Acids Res, 17*(11), 4381.

Tao, Y., Kassatly, R. F., Cress, W. D., and Horowitz, J. M. (1997). Subunit composition determines E2F DNA-binding site specificity. *Mol Cell Biol, 17*(12), 6994-7007.

Thibaud-Nissen, F., Wu, H., Richmond, T., Redman, J. C., Johnson, C., Green, R., et al. (2006). Development of Arabidopsis whole-genome microarrays and their application to the discovery of binding sites for the TGA2 transcription factor in salicylic acid-treated plants. *Plant J, 47*(1), 152-162.

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

Thijs, G., Moreau, Y., De Smet, F., Mathys, J., Lescot, M., Rombauts, S., et al. (2002). INCLUSive: integrated clustering, upstream sequence retrieval and motif sampling. *Bioinformatics, 18*(2), 331-332.

Thomas, M. S., and Flavell, R. B. (1990). Identification of an enhancer element for the endosperm-specific expression of high molecular weight glutenin. *Plant Cell, 2*(12), 1171-1180.

Tompa, M., Li, N., Bailey, T. L., Church, G. M., De Moor, B., Eskin, E., et al. (2005). Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol, 23*(1), 137-144.

Tuerk, C., and Gold, L. (1990). Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science, 249*(4968), 505-510.

Tuskan, G. A., Difazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., et al. (2006). The genome of black cottonwood, Populus trichocarpa (Torr. & Gray). *Science, 313*(5793), 1596-1604.

Ullmann, A., Perrin, D., Jacob, F., and Monod, J. (1965). [Identification, by in vitro complementation and purification, of a peptide fraction of Escherichia coli beta-galactosidase]. *J Mol Biol, 12*(3), 918-923.

Vandepoele, K., Vlieghe, K., Florquin, K., Hennig, L., Beemster, G. T. S., Gruissem, W., et al. (2005). Genome-wide identification of potential plant E2F target genes. *Plant Physiology, 139*(1), 316-328.

Venter, M. (2007). Synthetic promoters: genetic control through cis engineering. *Trends Plant Sci, 12*(3), 118-124.

Vicente-Carbajosa, J., and Carbonero, P. (2005). Seed maturation: developing an intrusive phase to accomplish a quiescent state. *Int J Dev Biol, 49*(5-6), 645-651.

VicenteCarbajosa, J., Moose, S. P., Parsons, R. L., and Schmidt, R. J. (1997). A maize zinc-finger protein binds the prolamin box in zein gene promoters and interacts with the basic leucine zipper transcriptional activator Opaque2. *Proceedings of the National Academy of Sciences of the United States of America, 94*(14), 7685-7690.

Vincentz, M., Leite, A., Neshich, G., Vriend, G., Mattar, C., Barros, L., et al. (1997). ACGT and vicilin core sequences in a promoter domain required for seed-specific expression of a 2S storage protein gene are recognized by the opaque-2 regulatory protein. *Plant Mol Biol, 34*(6), 879-889.

Wang, D. Y., Kumar, S., and Hedges, S. B. (1999). Divergence time estimates for the early history of animal phyla and the origin of plants, animals and fungi. *Proc Biol Sci, 266*(1415), 163-171.

Wang, H. W., Zhang, B., Hao, Y. J., Huang, J., Tian, A. G., Liao, Y., et al. (2007). The soybean Dof-type transcription factor genes, GmDof4 and GmDof11, enhance lipid content in the seeds of transgenic Arabidopsis plants. *Plant J, 52*(4), 716-729.

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

Wasserman, W. W., and Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. *Nat Rev Genet, 5*(4), 276-287.

Watanabe, Y., and Hirano, H. (1994). Nucleotide sequence of the basic 7S globulin gene from soybean. *Plant Physiol, 105*(3), 1019-1020.

Welchen, E., and Gonzalez, D. H. (2005). Differential expression of the Arabidopsis cytochrome c genes Cytc-1 and Cytc-2. Evidence for the involvement of TCP-domain protein-binding elements in anther- and meristem-specific expression of the Cytc-1 gene. *Plant Physiol, 139*(1), 88-100.

Weschke, W., Baumlein, H., and Wobus, U. (1987). Nucleotide sequence of a field bean (Vicia faba L.var.minor) vicilin gene. *Nucleic Acids Res, 15*(23), 10065.

White, R. J., and Jackson, S. P. (1992). The TATA-binding protein: a central role in transcription by RNA polymerases I, II and III. *Trends Genet, 8*(8), 284-288.

Wingender, E., Dietze, P., Karas, H., and Knuppel, R. (1996). TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res, 24*(1), 238-241.

Wu, C., Washida, H., Onodera, Y., Harada, K., and Takaiwa, F. (2000). 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*(3), 415-421.

Yamamoto, Y. Y., Ichida, H., Matsui, M., Obokata, J., Sakurai, T., Satou, M., et al. (2007). Identification of plant promoter constituents by analysis of local distribution of short sequences. *BMC Genomics, 8*, 67.

Yanagisawa, S., and Schmidt, R. J. (1999). Diversity and similarity among recognition sequences of Dof transcription factors. *Plant J, 17*(2), 209-214.

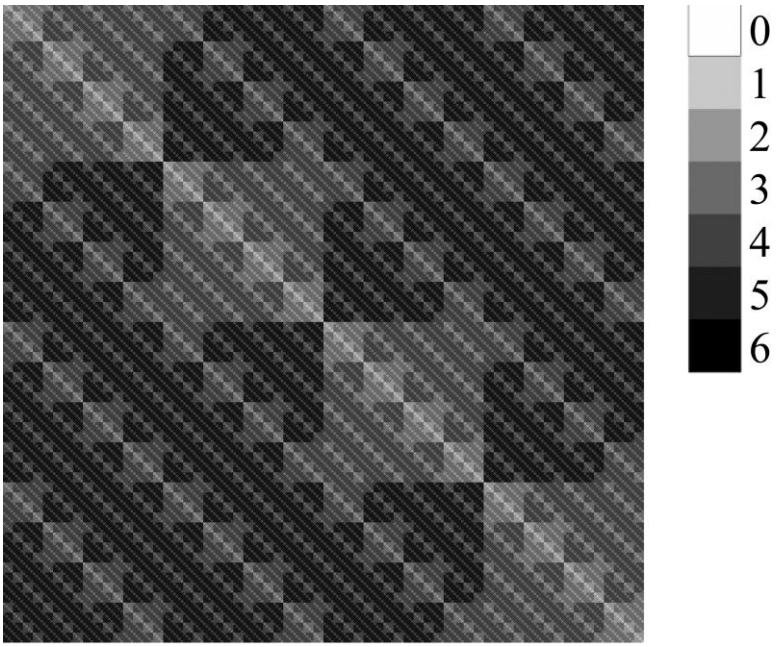
Yoshino, M., Kanazawa, A., Tsutsumi, K. I., Nakamura, I., and Shimamoto, Y. (2001). Structure and characterization of the gene encoding alpha subunit of soybean beta-conglycinin. *Genes Genet Syst, 76*(2), 99-105.

Yoshino, M., Nagamatsu, A., Tsutsumi, K., and Kanazawa, A. (2006). The regulatory function of the upstream sequence of the beta-conglycinin alpha subunit gene in seed-specific transcription is associated with the presence of the RY sequence. *Genes Genet Syst, 81*(2), 135-141.

Yu, J., Hu, S., Wang, J., Wong, G. K., Li, S., Liu, B., et al. (2002). A draft sequence of the rice genome (Oryza sativa L. ssp. indica). *Science, 296*(5565), 79-92.

Zilberman, D., Gehring, M., Tran, R. K., Ballinger, T., and Henikoff, S. (2007). Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. *Nat Genet, 39*(1), 61-69.

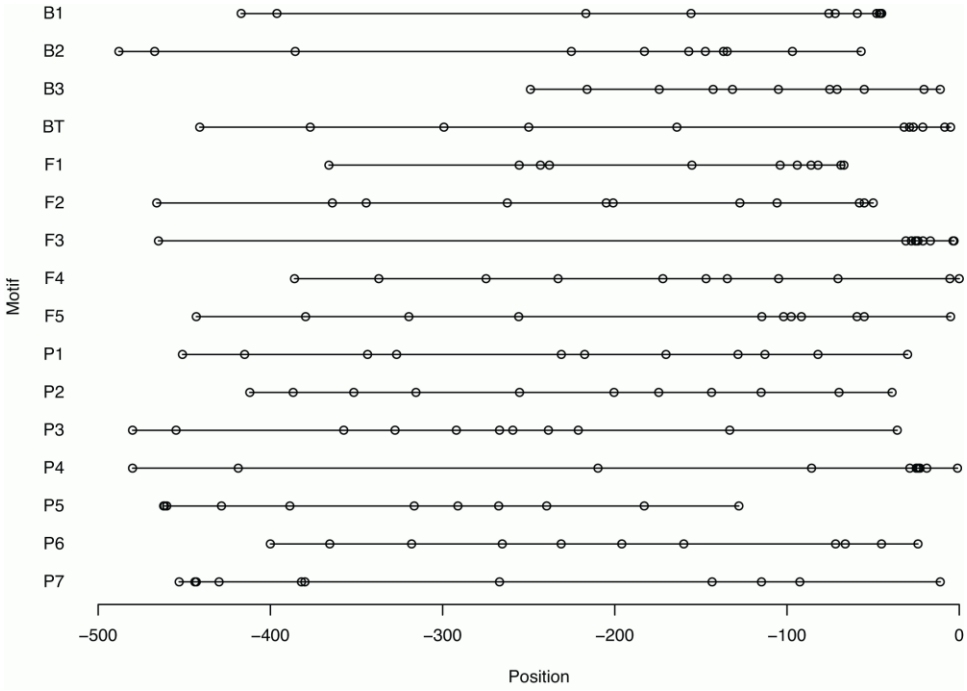
Appendix 1.



Supplementary Figure 1: Heat map of Hamming distances between words of length six.

Hamming distances are plotted in an alphabetically sorted list of nucleotide symbols (words) against itself. Identities are found on the diagonal. Hamming distances are organized in square recursive fractal geometry. At each level (in 0 to given word length), the square is divided into 16 squares (4 x 4), and a value of 1 is added to every square but to the four diagonal; this operation is iterated on each of the smaller squares defined in the previous operation. White corresponds to a HD of zero and black corresponds to a HD of six (see inset).

Appendix 2.



Supplementary Figure 2: Minimum, maximum and sample deciles for the position of SSP gene promoter motifs

The minimum, maximum and deciles of positions of best matching subsequences to motifs discovered in *Brassicaceae* (B1-3, BT), *Fabaceae* (F1-5, FT), and *Poaceae* (P1-5) are mapped on a 500 bp sequence (horixontal axis).

Appendix 3.

Supplementary Table 1: List of top-scoring Arabidopsis and rice promoters for the presence of seed storage protein gene promoter motifs

Species	GeneID	Description
Arabidopsis thaliana	AT4G27150.1	2S seed storage protein precursor
Arabidopsis thaliana	AT3G29075.1	glycine-rich protein
Arabidopsis thaliana	AT4G27170.1	2S seed storage protein precursor
Arabidopsis thaliana	AT4G27160.1	2S seed storage protein precursor
Arabidopsis thaliana	AT4G28520.4	12S seed storage protein precursor
Arabidopsis thaliana	AT4G27140.1	2S seed storage protein precursor
Arabidopsis thaliana	AT5G44120.3	12S seed storage protein precursor
Arabidopsis thaliana	AT5G07530.2	glycine-rich protein
Arabidopsis thaliana	AT2G28490.1	cupin family protein
Arabidopsis thaliana	AT5G54740.1	2S seed storage protein precursor
Oryza sativa	12007.t00931	prolamin precursor
Oryza sativa	12007.t00932	prolamin precursor
Oryza sativa	12012.t01542	prolamin PPROL 17 precursor
Oryza sativa	12002.t01476	glutelin type-B 4 precursor
Oryza sativa	12002.t01475	glutelin type-B 4 precursor
Oryza sativa	12002.t01308	glutelin type-B 7 precursor
Oryza sativa	12012.t01555	prolamin PPROL 17 precursor
Oryza sativa	12005.t02317	prolamin PPROL 14 precursor
Oryza sativa	12005.t02289	prolamin PPROL 14 precursor
Oryza sativa	12005.t02300	prolamin PPROL 14 precursor

Species, binomial name of species; GeneID, accession number; Description, functional gene annotation.

Appendix 4.

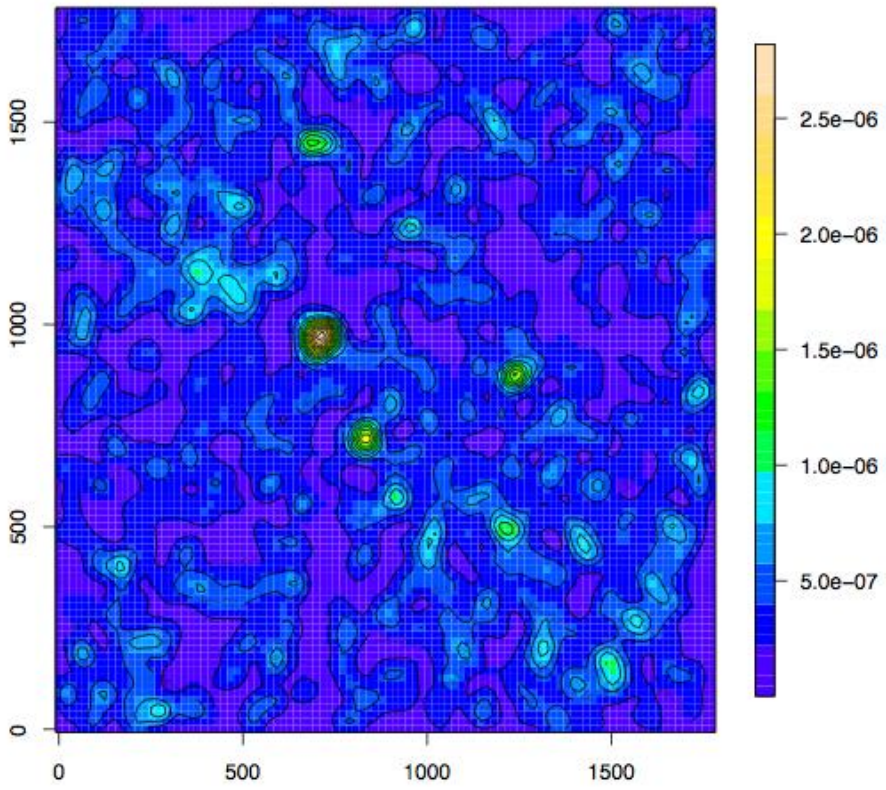
Supplementary Table 2: List of seed-storage protein gene promoters included in the analysis

Uniprot ID	GenPept ID	GenBank ID	Start	Stop	Strand	Species
P15455	BAB10979.1	AB005239.1	16870	17369	minus	Arabidopsis thaliana
P15456	AAD10680.1	AC003027.1	71129	71628	minus	Arabidopsis thaliana
P15457	AAA32743.1	M22032.1	418	917	plus	Arabidopsis thaliana
P15458	CAA80871.1	Z24745.1	2429	2928	plus	Arabidopsis thaliana
P15459	CAB38846.1	AL035680.1	31641	32140	plus	Arabidopsis thaliana
P15460	CAA80869.1	Z24744.1	1625	2124	plus	Arabidopsis thaliana
Q96318	AAB17379.1	U66916.1	460	959	plus	Arabidopsis thaliana
Q9ZWA9	AAD10679.1	AC003027.1	67682	68181	minus	Arabidopsis thaliana
Q9SK09	AAD21484.1	AC006587.5	15866	16365	minus	Arabidopsis thaliana
P33524	CAA41985.1	X59295.1	110	609	plus	Brassica napus
P33525	CAA44042.1	X62120.1	720	1219	plus	Brassica napus
P11090	CAA32692.1	X14555.1	181	680	plus	Brassica napus
P01090	AAA87348.1	J02798.1	603	1102	plus	Brassica napus
P17333	CAA35580.1	X17542.1	446	945	plus	Brassica napus
Q02498	CAA42478.1	X59808.1	218	717	plus	Raphanus sativus
P10562	CAA33172.1	X15076.1	809	1308	plus	Canavalia gladiata
P13917	BAA03681.1	D16107.1	864	1363	plus	Glycine max
P11827	AAB01374.1	M13759.1	407	906	plus	Glycine max
P25974	AAB23463.1	S44893.1	411	910	plus	Glycine max
P04776	CAA33215.1	X15121.1	140	639	plus	Glycine max
P04405	CAA33216.1	X15122.1	302	801	plus	Glycine max
Q2HUY7	ABD32862.1	AC149038.2	125028	125527	minus	Medicago truncatula
P19329	AAA33753.1	M68913.1	2695	3194	plus	Phaseolus vulgaris
Q42460	CAA90585.1	Z50202.1	1153	1652	plus	Phaseolus vulgaris
P62930	AAC61879.1	M81864.1	34	533	plus	Pisum sativum
P13915	CAA29695.1	X06398.1	63	562	plus	Pisum sativum
P13919	AAA33660.1	M73805.1	520	1019	plus	Pisum sativum
P15838	CAA35056.1	X17193.1	991	1490	plus	Pisum sativum

P05692	CAA30067.1	X07014.1	64	563	plus	Pisum sativum
P13918	CAA32239.1	X14076.1	1035	1534	plus	Pisum sativum
P05190	CAA27313.1	X03677.1	2199	2698	plus	Vicia faba
P08438	CAA68559.1	Y00506.1	2029	2528	plus	Vicia faba
P06470	CAA26889.1	X03103.1	14	513	plus	Hordeum vulgare
P06293	CAA36015.1	X51726.1	527	1026	plus	Hordeum vulgare
Q0DN94	CAA34926.1	X17074.1	246	745	plus	Oryza sativa
P17048	CAA46197.1	X65064.1	113	612	plus	Oryza sativa
P29835	BAA09308.1	D50643.1	427	926	plus	Oryza sativa
P07728	BAB61225.1	AP003256.3	166219	166718	plus	Oryza sativa
P07730	BAA00462.1	D00584.1	1843	2342	plus	Oryza sativa
Q09151	AAA50314.2	M28158.1	311	810	plus	Oryza sativa
P14323	CAA38212.1	X54314.1	781	1280	plus	Oryza sativa
Q02897	BAD19800.1	AP005511.3	98052	98551	plus	Oryza sativa
P14614	AAM97692.1	AF537221.2	604	1103	plus	Oryza sativa
P0C5E5	AAW80678.1	AY896773.1	132	631	plus	Oryza sativa
Q42465	BAC20110.1	AP005160.3	156007	156506	minus	Oryza sativa
P20698	ABA97054.1	DP000011.2	9685896	9686395	plus	Oryza sativa
P14690	CAA34230.1	X16104.1	288	787	plus	Sorghum bicolor
P04726	CAA26383.1	X02538.1	28	527	plus	Triticum aestivum
P21292	AAA34272.1	M36999.1	191	690	plus	Triticum aestivum
P10388	CAA31395.3	X12928.4	3340	3839	plus	Triticum aestivum
P10387	CAA31396.1	X12929.2	2378	2877	plus	Triticum aestivum
P04706	AAA33468.1	M16066.1	23	522	plus	Zea mays
P04704	CAA24717.1	V01470.1	390	889	plus	Zea mays
P08031	CAA37595.1	X53515.1	673	1172	plus	Zea mays

Uniprot ID, UniProt/SwissProt sequence identifier; GenPept ID, GenPept accession number; GenBank ID, GenBank accession number; Start, start coordinate for the coding sequence; Stop, stop coordinate for the coding sequence; Strand, strand (+/-) of the coding sequence; Species, binomial name of species.

Appendix 5.



Supplementary Figure 3: Hybrid image-contour plot of two-dimensional kernel density estimation of DNA motifs clustered by similarity and plotted on promoters clustered by GO semantic similarity.

Motifs clusters (vertical axis) are plotted against semantic similarity clusters (horizontal axis) and co-clusters are visualized using two-dimensional density. Yellow spots highlight regions of higher density, which indicates the presence of similar motifs in the promoters of genes of similar function.

Appendix 6.

Supplementary Table 3: DDOPM MySQL table description.

Field	Type	Null	Key
cluster_id	smallint(6)	YES	NULL
n_motif	tinyint(4)	YES	NULL
go_id	char(10)	YES	NULL
go_term	varchar(101)	YES	NULL
go_type	varchar(18)	YES	NULL
short_name	varchar(103)	YES	NULL
full_name	varchar(829)	YES	NULL

Appendix 7.

Chapter 3 and Chapter 4 are Open Access articles:

Fauteux, F., Blanchette, M., & Stromvik, M. V. (2008). Seeder: Discriminative Seeding DNA Motif Discovery. Bioinformatics, 24(20), 2303-2307.

Fauteux, F., and Stromvik, M. V. (2009). Seed storage protein gene promoters contain conserved DNA motifs in *Brassicaceae*, *Fabaceae* and *Poaceae*. BMC Plant Biol. 9:126.