Regulation of Gene Expression in Plants

Regulation of Gene Expression in Plants

The Role of Transcript Structure and Processing

Edited by

Carole L. Bassett
United States Department of Agriculture
Kearneysville, WV, USA

Springer

Carole L. Bassett
USDA - ARS
2217 Wiltshire Road
Kearneysville, WV 25430
USA

Library of Congress Control Number: 2006937882

ISBN-10: 0-387-35449-2 e-ISBN-10: 0-387-35640-1
ISBN-13: 978-0-387-35449-1 e-ISBN-13: 978-0-387-35640-2

Printed on acid-free paper.

9 8 7 6 5 4 3 2 1

springer.com

Dedication

The contributors would like to dedicate this book to their families and friends who have so patiently ignored the long hours and frustrations that are a natural part of scientific research and supported us intellectually and emotionally in our efforts. We hope they have also shared with us in the satisfaction of achievement and the joy of discovery that occur all too infrequently in the realm of science. The editor would also like to dedicate this book to her father, Dr. Boude Bowman Leavel (1919-2003), who inspired her interest in science and nurtured her love of learning. He was a dedicated physician and as close to a perfect husband and father as will ever be found.

Foreword

Regulation of gene expression in eukaryotes is complex and occurs on many levels, often overlapping each other. In a hierarchical sense, the chromatin structure itself is the first level influencing gene expression through methylation of select bases, posttranslational modification of histones, and alterations in scaffolding. The second level includes transcription and posttranscriptional processing, including export and turnover of mRNAs. It is with this level of gene regulation that the current book is primarily focused. The third level of regulation involves translation and posttranslational events. As we shall see in the following chapters, it is sometimes difficult to separate these processes from each other. Of necessity we have narrowed our emphasis on transcript expression to the structure and processing of mRNAs and have opted to minimize the mechanics of transcription initiation, elongation, and termination. This is not to say that such processes are not important, only that they have been amply covered in numerous recent reviews.

Preface

In almost all of the areas of gene expression control except one, plant research has lagged considerably behind studies in yeast, insects and vertebrates. Advances in animal gene expression control have also benefited plant research, as we continue to find that much of the machinery and mechanisms controlling gene expression have been preserved in all eukaryotes. However, there are some interesting differences in gene structure and regulation between plants and animals. First, vertebrate genes can be quite large, often spanning tens of thousands of base pairs and usually separated by numerous large introns, whereas plant genes tend to be much smaller (averaging between 1–2 kb (kilobases) with fewer and smaller introns. Second, as we shall see in the following chapters, plant transcripts retain introns more often than do animal transcripts (30% of all genes in the model plant, *Arabidopsis,* compared to 10% in humans). Unfortunately, at this time we have only a few plant models for gene regulation, and only *Arabidopsis*, rice and poplar have been fully sequenced. Since *Arabidopsis* was the first plant genome to be fully sequenced, most of our information has come from studies of its transcriptome, and it is not known to what extent it truly represents other members of the plant kingdom. In addition, as our knowledge of factors influencing gene expression increases, so too, does our recognition that the current annotations in the gene databases will need to be updated to reflect new information as it appears in the literature. Finally, compared to animals, plants have evolved different signaling mechanisms, partly because plant hormones do not exactly function as do those in animals, and partly because plants must cope with environmental changes differently than animals, since they cannot physically escape their environments except possibly through reproduction. Therefore, plants have evolved very complex, interacting signaling pathways in response to developmental signals and biotic/abiotic stresses. All of these observations ultimately reflect some of the differences plants display in regulating gene expression compared to yeasts, insects and vertebrates.

Although we have touched upon some of the differences between animal and plant control of gene expression here, it is equally important to recognize and appreciate the common mechanisms they share. For example, the basic

transcriptional machinery *via* DNA-dependent RNA Polymerase II is virtually the same in plants and animals. Furthermore, some transcription factors, like *myb* and *myc* factors are similar in structure and function in both plants and animals. Plants and animals contain introns separating the coding regions of most genes and again, they utilize similar machinery to process the introns and form mature mRNAs. Since translation in all eukaryotes is basically the same, we see more similarities between plants and animals in this process, than differences. These mechanisms relate to mRNA structure, including sequences in the 5′ leader region and those in the 3′ untranslated region which influence the efficiency and selectivity of translation. It is hoped that the following chapters will expose the reader to some of the most recent, novel and fascinating examples of transcriptional and posttranscriptional control of gene expression in plants and, where appropriate, provide comparison to notable examples of animal gene regulation.

Contents

Contributors

Carole L. Bassett
USDA-ARS, Appalachian Fruit Research Station,
2217 Wiltshire Road, Kearneysville, WV 25430, USA

Robert E. Farrell, Jr.
Biology Department, Pennsylvania State University,
208 Mueller Lab, University Park, PA 16802-5301, USA

Jean-Michel Hily
Génomique dévelopment du Pouvoir pathogène, INRA-Université Bordeaux,
Segalen/Bordeaux, BP 81, 33883 Villenave d'Ornon, France

Arthur G. Hunt
Department of Plant and Soil Sciences, University of Kentucky,
Plant Sciences Building, 1405 Veterans Drive, Lexington, KY 40546-0312, USA

Zongrang Liu
USDA-ARS, Appalachian Fruit Research Station,
2217 Wiltshire Road, Kearneysville, WV 25430, USA

Eliezer S. Louzada
Department of Agronomy and Resource Sciences, Texas A & M University,
312 N. International Blvd., Weslaco, TX 78596, USA

1
The Regulation of Gene Expression in Plants and Animals

Robert E. Farrell, Jr.

1.1. Overview of Eukaryotic Transcription

The control of gene expression in all cells involves an elaborate and dynamic interplay among what might best be described as regulatory molecules. These molecules include RNA polymerases, myriad transcription factors, the DNA template, the RNA produced by transcription, and the protein produced by translation with its attendant processing. To interfere with or in some way modify any of these critical elements is to potentially cause a profound change in the phenotypic manifestation of one or more genes or gene relays. It is clear that transcription is a consequence of a series of well-orchestrated, ordered events. Contemporary methods that examine this aspect of gene expression have revealed the dynamic nature of this biochemical process.

The examination of gene expression often revolves around quantifying the abundance of a particular transcript. This assessment may be performed using state-of-the-art methods such as high throughput microarrays and real-time PCR, or by the use of the time-honored methods of Northern analysis, nuclease protection, or semi-quantitative PCR (described in Farrell, 2005). While all of these methods provide information about the transcriptional activity of genes, they measure steady-state levels only, i.e., the final accumulation of RNA in the cell at the moment of lysis; these methods fail to take into account the stability of the RNA, as well as the rate of transcription at the specific loci under investigation. In order to describe more completely the nature of gene expression modulation, it is incumbent upon the investigator to examine the level of the corresponding protein(s) so as to determine whether an alteration of transcript levels, protein levels, or both are in some way associated with a phenotypic change.

1.1.1. Regulation of Gene Expression

The modulation of transcript and protein levels is a nonstatic, on-going process that governs all cell and tissue function. Experiments performed using adenovirus and SV40 models were the first to elucidate some of the

biochemical events associated with the synthesis of mRNA. Current research has demonstrated that the secondary and tertiary structures formed by RNA and numerous RNA binding proteins influence the posttranscriptional fate of the transcript. While the number of potential gene regulatory control points is virtually infinite, in the *broadest* sense one might consider the following four major headings:

1. Transcriptional regulation
2. Posttranscriptional regulation
3. Translational regulation
4. Posttranslational regulation

Eukaryotic RNA molecules are synthesized by transcription in the nucleus, in mitochondria, or in chloroplasts. This is transcriptional-level regulation: a gene either is or is not transcribed. The same subcellular compartment where transcripts are synthesized is also the location where they are processed to maturity. Maturation, stability, and export of mRNAs represent posttranscriptional levels of gene regulation. With respect to nuclear transcripts, mature mRNA molecules may be exported into the cytoplasm, usually lacking the intron sequences that were part of heterogeneous nuclear RNA (hnRNA). Nucleocytoplasmic transport is an extremely discriminating mechanism, involving movement to and then through nuclear pore complexes (for reviews see Piñol-Roma and Dreyfuss, 1993; Maquat, 1997; Chapter 6). It is very likely that mature mRNA will become translated, but this is not always the case. The chemical stability of mRNA is itself an important posttranscriptional regulatory control point, as is the formation of nontranslatable mRNA:protein complexes previously known as "informosomes" (Preobrazhensky and Spirin, 1978; Bag, 1991). Translation initiation is another major control point affecting gene expression. mRNAs that bind eukaryotic initiation factors (eIFs) with low affinity typically do not synthesize as much polypeptide product as mRNAs with a higher binding affinity. This level of control occurs at the translational level and includes the elongation and termination of polypeptide synthesis. If translation of a particular mRNA does occur, then the translated polypeptide itself is subject to extensive posttranslational modifications. The magnitude and diverse nature of these posttranslational modifications is only now coming to light with the advent of the emerging field of proteomics. Some of the common posttranslational modifications include, but are not limited to, proteolytic cleavage (with or without ubiquitination), acylation, methylation, prenylation, carboxylation, glycosylation, phosphorylation, neddylation, acetylation, sumoylation, and hydroxylation.

The biochemical status of a cell or tissue sample can be qualitatively and quantitatively defined as a function of mRNA complexity. For example, one may investigate whether modulation of gene expression is regulated transcriptionally or by a posttranscriptional event(s). Thus, insightful consideration

must be given to the method of cellular disruption and RNA purification because of the numerous sub-populations of transcripts in the cell (for protocols, see Farrell, 2005). In order to paint a more complete picture of the methods by which one or more genes of experimental interest are being regulated in a model system, the abundance of the corresponding protein(s), and modulation thereof, should also be scrutinized.

1.1.2. Nature of Transcription

Transcription is that process by which the information residing in a double-stranded DNA molecule (dsDNA) is transferred to a nascent RNA molecule. "Transcription" is so-named because the code contained within the sequence of the DNA is rewritten in the same language, namely, the language of nucleic acids. RNA polynucleotides are assembled at a rate of about 40 nucleotides (nt)/sec in eukaryotic and in prokaryotic cells. Translation, in contrast, is that process by which a protein or polypeptide is synthesized from the information encoded in the RNA; it is so-named because the information contained within the RNA is rewritten in an entirely different language, namely, the language of proteins. Proteins are assembled from amino acid monomers at a rate of about 5–8 amino acids (aa)/sec in eukaryotes and about twice as fast in bacteria. The mechanics of the processes of transcription and translation ensure the requisite colinearity and accuracy of information transfer from the DNA to the resulting protein which, in turn, ensures proper biochemical function. To do otherwise would constitute a mutation, with potentially serious consequences for both the cell and the organism.

In eukaryotic cells, transcription is compartmentalized within membrane-bound organelles and, in the case of the nucleus, transcription is both spatially and temporally separated from translation. In both plant and animal cells, the half-life of a typical mRNA transcript is significantly greater than the half-life of mRNA molecules in bacteria; prokaryotic transcripts are often being both translated and degraded from their 5′ ends even while transcription is still on-going at the 3′ ends of the "nascent" transcripts. The enhanced stability of eukaryotic transcripts clearly facilitates transport across the nuclear membrane into the cytoplasm where mRNAs may have an opportunity to engage the protein translation apparatus.

RNA is produced from different loci at different rates. The quantity of a particular RNA species in a cell is referred to as its abundance. It is common to classify genes based on the abundance category into which their transcripts fall and to describe measured changes in the expression of a gene at the RNA level as a change in the relative abundance of that transcript. This terminology is used to describe (1) a change in the quantity of a particular transcript relative to another transcript, such as a housekeeping gene, or (2) a change in the quantity of a transcript after experimental manipulation relative to an untreated control.

1.1.2.1. Important Plant and Animal Models for Studying Transcription

Plants: The genomes of *Arabidopsis* (*Arabidopsis thaliana*), rice (*Oryza sativa*), and poplar (*Populus trichocarpa*) have been completely sequenced. Projects to sequence tomato (*Lycopersicon esculentum*), *Medicago truncatula* (a forage legume) and lotus (*Lotus japonica*) and many other plants are currently underway. Because *Arabidopsis* was the first plant whose genome was completely sequenced, it has served as a model plant for gene organization and structure, as well as for transcript processing and regulation.

Models for inheritance and gene mapping studies have historically been maize (*Zea mays*) and tomato. More recently wheat and barley, as well as soybean and alfalfa have been the focus of trait mapping, especially QTL (quantitative trait loci) identification of disease and stress resistance. Genetic studies in plants are often complicated by the fact that some of the more important crops are polyploids or aneuploids. Common plant models for the study of non-Mendelian inheritance include shoot variegation in the four o'clock *Mirabilis jalapa*, cytoplasmic male sterility in corn, and uniparental inheritance of mating type in the alga, *Chlamydomonas reinhardtii*.

Animals: Much of the pioneering research in the area of transcription and transcriptional control of gene expression involved the use of such unassuming animals as *Drosophila*, sea urchins, *Caenorhabditis elegans* (nematode), rat, mouse, and just as importantly, animal cell culture models, such as the CHO-K1 (Chinese Hamster Ovary) and NIH 3T3 (swiss mouse fibroblast) cell lines.

1.1.2.2. Genome Organization Affects Nuclear Gene Transcription

The eukaryotic genome is both large and quite sophisticated. If the DNA from a typical eukaryotic cell were stretched end to end it would reach several feet in length. Organization of the DNA is critically important for packaging into a microscopic nucleus, which must be done in such a way that specific DNA sequences (genes) and their regulatory elements (promoters) can be periodically accessed in order to be transcribed. Supercoiling is the key.

The first level of eukaryotic nuclear DNA organization involves the formation of nucleosome particles. Nucleosomes consist of an association of 146 bp of core DNA wrapped around a group of eight histone proteins, i.e., two molecules each of the histones H2A, H2B, H3, and H4. The orderly assemblage of nucleosomes is often described as beads on a string when viewed by electron microscopy. A fifth histone protein, H1, is located on the outside of the nucleosome particle and can be readily removed from the complex under conditions of low ionic strength.

Prior to the initial interaction of the chromatin with the early initiators of transcription, this three-dimensional nucleosome architecture prevents RNA polymerase and transcription factor association with the promoter. The local unwinding of tightly packaged DNA is a first step needed to provide transcription factors the opportunity to interact with the regulatory elements governing expression of a particular gene. Beyond the formation of

nucleosomes, subsequent levels of coiling ultimately lead to the characteristic formation of chromatin which, at the onset of mitosis, further condenses into recognizable chromosomes.

1.1.2.3. Gene Organization Affects Nuclear Gene Transcription

The typical eukaryotic genome is well-known for its excessive amount of non-coding, intergenic spacers, as well as the fact that the structural portions of genes are divided into coding regions, known as exons, and intervening, noncoding regions known as introns. While the precise origin of interrupted genes is anyone's guess, one of the more rational attempts to explain this phenomenon is that introns are remnants of genome invasion attempts by primordial prokaryotes. The intron sequences that are observed distributed throughout plant and animal genes are variable in number, length, and nucleotide content, and infrequently are found to contain open reading frames (ORFs). The only predictable feature of introns is the presence of highly conserved dinucleotide pairs that define the intron boundaries: introns begin with GU and end with AG. By recognizing these dinucleotides, and the adjacent bases that are conserved to a lesser extent, introns are excised from hnRNA transcripts during and shortly after transcription. A considerable amount of the typical eukaryotic genome does not appear to encode anything, and these intergenic regions are often impressive in size.

Most animal cells are diploid with the exception of gametes. As such, each cell has at least two copies of each gene, and there are many genes that exist in multiple copies. The genome is also subject to change, not only by random mutation, but also through the orderly duplication of specific genes in response to external stimuli. For example, the drug methotrexate, which is commonly used to treat patients with either cancer or arthritis, is known to induce amplification of the dihydrofolate reductase (DHFR) gene (Schimke, 1981). Another example of the nonstatic nature of the animal genome occurs in lymphocytes, in which the immunoglobulin genes are spliced in order to facilitate the production of a customized antibody in response to exposure to a previously unknown antigen (reviewed by Lewin, 2004).

In plants, the parameters governing genome stability are similar to those in animals, though it is not unusual to find triploid, tetraploid, and octaploid tissues and organisms in the plant kingdom. In addition, plants seem better able to tolerate aneuploidy than mammals, and some hybrid progeny from interspecies crosses in the plant kingdom are fertile, whereas this rarely happens in higher animals.

1.1.2.4. RNA Polymerases I, II, III, and IV

Genes in plant cells, as in animal cells, are transcribed by enzymes known as DNA-dependent RNA polymerases. In eukaryotic organisms there are three nuclear RNA polymerases, designated RNA polymerase I, RNA polymerase II, and RNA polymerase III. These enzymes, which are composed of

multiple subunits and are among the largest and most complex enzymes in the cell, each transcribe a different class of genes. This approach to transcription is markedly different compared to prokaryotic cells where one type of RNA polymerase is responsible for the transcription of all genes. Each eukaryotic RNA polymerase is functional only in the presence of DNA, which acts as a template. These enzymes also exhibit an absolute cofactor requirement for Mg^{++} and are dependent upon the presence of myriad transcription factors to initiate and support transcription.

RNA polymerase I is responsible for the transcription of the ribosomal genes, and is therefore ultimately responsible for the production of 28S ribosomal RNA (rRNA), 18S rRNA, and 5.8S rRNA. RNA polymerase II transcribes genes that encode proteins, first producing hnRNA which can mature into mRNA molecules that can support translation. RNA polymerase III transcribes the transfer RNA (tRNA) genes and a small ribosomal rRNA gene that produces the 5S rRNA. RNA polymerases are also responsible for the synthesis of small nuclear RNA (snRNA), small nucleolar RNA (snoRNA), and small cytoplasmic RNA (scRNA). In mammals, a fourth RNA polymerase (single-polypeptide nuclear RNA polymerase IV [spRNAP-IV]) has recently been described (Kravchenko *et al*., 2005). This polypeptide is derived from a nuclear-encoded mitochondrial RNA polymerase by alternative splicing which results in a truncated polypeptide lacking 262 amino acids near the amino terminus, including the mitochondrial transit sequence. spRNAP-IV is located in the nucleus and regulates a subset of nuclear-encoded genes.

Similarly, a novel polymerase, RNA polymerase IV, or simply Pol IV, has been discovered in plants (Onodera *et al*., 2005). Pol IV does not appear to be necessary for viability, nor does its activity overlap with the known functions of RNA polymerases I, II, and III; in contrast, it is used by the cell to promote methylation-associated higher-order heterochromatin formation (Kanno *et al*., 2005; Chapter 5).

As with all nucleic acid polymerases, RNA polynucleotides are assembled $5' \rightarrow 3'$, and in this case, in a DNA-dependent manner. Mutations notwithstanding, the products of transcription are identical to the base sequence found on the DNA coding strand, with the exception of the substitution of uracil in RNA in place of the thymine found in DNA. Thus, the newly transcribed RNA is likewise complementary to the DNA template strand upon which it was synthesized directly. Curiously, more than 70% of all transcribed RNA is subsequently degraded in the nucleus, never maturing into corresponding cytoplasmic species (Soeiro *et al*., 1968; Jackson *et al*., 2000). While this may seem like an inordinate waste of cellular resources, it is nonetheless a well-documented mode of posttranscriptional regulation of gene expression.

1.1.2.5. Phage-Type RNA Polymerases: The RpoT Genes

Angiosperms possess a small, unique group of nuclear-encoded RNA polymerases that resemble the well-characterized RNA polymerases associated with the T7 and T3 bacteriophages. These genes, known as the RpoT

polymerase family, are believed to have arisen by duplication from an ancestral gene encoding the mitochondrial RNA polymerase. At least some of the genes in this family are found in both monocots and dicots. The RpoT genes were initially identified and characterized in *Arabidopsis thaliana* (Hedtke *et al*., 1997), *Cheopodium album* (Weihe *et al*., 1997), *Physcomitella patens* (Richter *et al*., 2002), maize (Young *et al*., 1998; Chang *et al*., 1999), and wheat (Ikeda and Gray, 1999). In *Arabidopsis*, the three members of this family that have been found so far are known as RpoT1, RpoT2, and RpoT3. RpoT1 is responsible for the transcription of mtDNA (mitochondrial DNA) while RpoT3 transcribes cpDNA (chloroplast DNA). Particularly interesting is RpoT2, the N-terminal amino acid composition of which directs RpoT2 into chloroplasts and mitochondria, suggesting a dual transcriptional role for this enzyme involving two different genomes (Hedtke *et al*., 2000). In moss, the RpoT2 mRNA contains two inframe AUG codons near the 5′ terminus (Richter *et al*., 2002). When translation initiation occurs at the second AUG codon, the resulting enzyme is directed into the mitochondria. Translation initiation from the first AUG codon, in contrast, produces a chloroplast-targeted translation product.

1.1.3. Transcription Factors and Promoter Elements

Initiation of the process of transcription is a complex event, involving far more components than were appreciated even as recently as five years ago. The specific DNA sequences required for the binding of transcription-associated proteins and enzymes are known generically as regulatory elements, and these include motifs commonly known as the TATA box, the CAAT box, enhancers, and the GC-rich elements. The highly variegated proteins that bind to the DNA and to each other to facilitate the initiation of transcription are collectively referred to as transcription factors. Those DNA sequences that are spatially associated with a particular genetic locus are known as *cis*-acting elements, while the various RNA polymerases and all of the transcription factors that are encoded at other genetic loci, are referred to as *trans*-acting factors. For example, in plants, the *cis*-acting motif, (T)(T)TGAC(C/T), is known as a W box, to which members of the WRKY superfamily of transcription factors bind in response to wounding and certain other stresses (Cheong *et al*., 2002). Before an RNA polymerase can engage a transcription template, a pre-initiation complex consisting of several different transcription factors binds to the DNA promoter itself, and this precedes the binding of RNA polymerase (reviewed by Lewin, 2004). An overview of transcription and associated processes is illustrated in Figure 1.1.

The transcription factors associated with genes that encode proteins, as opposed to genes which encode tRNA and rRNA, appear to be highly conserved among eukaryotes, as is RNA polymerase II, which transcribes these genes. Basal transcription factors associated with genes that encode mRNA are generically referred to as $TF_{II}X$, with X representing a specific

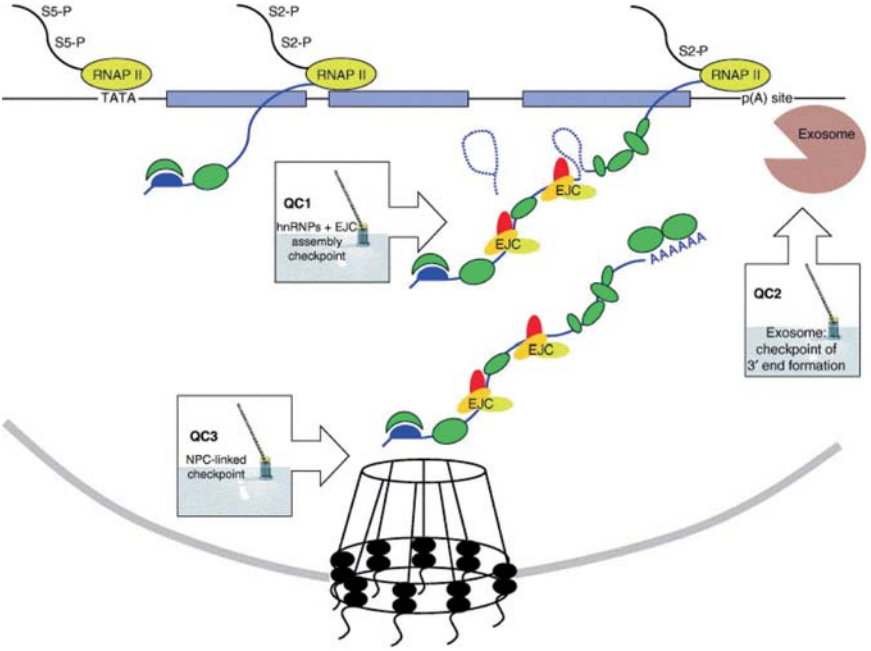

FIGURE 1.1. Key features and quality control checkpoints of the 'mRNA factory'. In the CTD (C-terminal domain) of RNAP II, the changing phosphorylation of repeat residues Serine 5 (S5) and Serine 2 (S2) during transcription regulates the dynamic interactions of the CTD with enzymes of mRNA maturation (capping, splicing and polyadenylation factors, not shown). The 5′ cap is added early in transcription, whereas splicing of introns usually occurs cotranscriptionally but can happen later as well. Splicing leads to a deposition of the exon junction complex (EJC) upstream of the splice junctions. Numerous hnRNPs and other RNA binding proteins (represented as green symbols of different shapes and sizes), including cap binding protein eIF4E and poly(A) binding protein, associate with maturing transcripts; the complement of these factors is likely to be distinct for different mRNAs. The appropriate assembly of hnRNP proteins and completion of 3′ end formation is monitored at quality control checkpoints QC1 and QC2, respectively; messages improperly matured at these stages are subject to retention at the transcription site and degradation by the exosome complex. The nuclear pore complex-linked checkpoint, QC3, ensures that only spliced transcripts are exported from the nucleus and causes mRNA to be kept at the site of transcription when the nuclear-pore complex-linked export step is blocked. Reproduced with permission, from Belostotsky and Rose, 2005, Trends in Plant Science, 10(7), 347–353, © Elsevier Ltd. (*See Color Plates*)

transcription factor, *e.g.*, $TF_{II}D$. Table 1.1 lists several representative classes of transcription factors that have been identified in plants and in animals.

Many promoters exhibit a DNA sequence known as a TATA box, so named because of the highly conserved sequence, TATAA, or a close variant

TABLE 1.1. Classes of eukaryotic transcription factors[1].

	Transcription factor categories			
Super family	Examples of major classes	Observed in plants	Observed in animals	Transcriptional factor examples
Basic domains	Leucine zippers	Yes	Yes	c-jun (human) c-fos (mouse) CPRF-3 (parsley)
	Helix-loop-helix	Yes	Yes	Id2 (mouse) Lc (maize) NUC-1 (*Neurospora*)
	NF-1	No	Yes	CTF-3 (human) NF-1C1 (pig) NF-1A1 (chick)
Zinc-coordinating DNA binding Domains	Cys4 zinc finger (nuclear receptor type)	No	Yes	T3R-α (rat) AR (human) CF1 (fruit fly)
	Cys2-His2 zinc finger domain	Yes	Yes	$TF_{III}A$ (ubiquitous TF WT1 (human) GAL4 (yeast)
	Diverse Cys4 zinc fingers	Yes	Yes	GATA-1 (frog) AREA/NIT-2 (*Neurospora*)
Helix-turn-helix	Homeo domain	Yes	Yes	HOXC6 (zebra fish) Knox3 (barley) KN1 (maize)
	Heat shock factors	Yes	Yes	HSF1 (human) HSF3 (chick) HSF24 (tomato)
	Tryptophan clusters	Yes	Yes	c-myb (human) GL1 (*Arabidopsis*) MYB.pH3 (petunia)
β-Scaffold factors	STAT	No	Yes	STAT2 (human) STAT4 (mouse) STAT5A (sheep)
	p53	No	Yes	p53 (human) p53as (mouse)
	MADS Box	Yes	Yes	GLO (tobacco) AG (*Arabidopsis*) ZEM1 (maize)
	TATA-binding proteins	Yes	Yes	TBP (human) TBP (*Arabidopsis*) TBP (yeast)
Other types of transcription factors	Copper fist proteins (fungal)	No	No	ACE1 (yeast) AMT1 (yeast)
	HMGI(Y)	No	Yes	HMG Y (human) HMGI-C (mouse)
	Pocket domain	No	Yes	Rb (mink) CBP (mouse) p107 (human)

(Continued)

TABLE 1.1. Classes of eukaryotic transcription factors[1]. — Cont'd

Super family	Transcription factor categories: Examples of major classes	Observed in plants	Observed in animals	Transcriptional factor examples
	E1A-like factors (adenovirus)	No	No	E1A 12S (adenovirus) E1A 13S (adenovirus)
	AP2/EREBP-related factors	Yes	No	AP2 (*Arabidopsis*) EREBP-1 (tobacco) ARF1 (*Arabidopsis*)

[1] Some of this information was derived from www.gene-regulation.com/pub/databases/transfac/cl.html.
Visit this site for a more comprehensive listing of known TFs.

thereof. Most often a TATAA box is situated at –25, meaning 25 nts upstream from the transcription start site (TSS). The TATA box is considered a fundamental part of transcript initiation, although transcription can also be initiated from TATA-less promoters. In animals this occurs when the transcription factor IIB recognition element ($TF_{II}B$; Lagrange et al, 1998) and downstream promoter element (DPE; Burke and Kadonaga, 1996) are present in place of the TATA box element. An additional motif, known as an initiator sequence (Inr) is able to promote transcription autonomously or in conjunction with a TATA or other element (Smale and Baltimore, 1989). The precise geometry of these elements will determine the TSS. In *Arabidopsis*, the TATA box is the primary regulatory element required to initiate transcription in less than one-third of promoters analyzed (assuming that full length cDNAs represented the primary transcript; Molina and Grotewold 2005). This means fully two-thirds of the putative *Arabidopsis* promoters are without TATA boxes. It is noteworthy that the DPE and Inr sequences mentioned above in TATA-less animal promoters were not over-represented in the TATA-less *Arabidopsis* promoters, suggesting that there are fundamental architectural differences between plant and animal promoters.

Tissue-specific expression of particular genes is fundamental to proper function at the tissue level and beyond. As such, identification of the functional TATA box and its associated regulatory element in a particular tissue provides a greater understanding of gene regulation in that tissue. Currently available software is particularly useful for the *in silico* identification of potentially functional TATA boxes, especially when the region upstream of the TSS is AT-rich. While no single software product can predict TATA box functionality with absolute confidence, merging the results generated through the use of several different gene analysis programs allows the investigator to make a more informed judgment. For example, the presence of three operational TATA boxes was recently reported in the promoter region of the