Conserved control signals in the transcriptome of higher plants

Khanh Tran

Thesis submitted for the degree of Doctor of Philosophy

May 2010

Discipline of Plant and Pest Science
School of Agriculture, Food, and Wine
The University of Adelaide

CHAPTER 1

LITERATURE REVIEW

CHAPTER 1 LITERATURE REVIEW

1.1 INTRODUCTION

The control of gene expression in cells, also known as gene regulation, occurs at many different steps (Figure 1.1). These steps can be differentiated into two major points of gene regulation: transcriptional and post-transcriptional control. Understanding the mechanisms that control gene expression is an important goal in bioinformatics, a term referring to the application of information technology to the field of molecular biology. To date, there are less bioinformatics studies on post-transcriptional control, and most research has focused on the transcriptional control of gene expression. However, it is known that post-transcriptional control is important for more than 15% of animal and plant genes (Gygi et al. 1999; Munroe 2004).

Post-transcriptional control can be accomplished by short non-coding RNAs. One class, the microRNAs (reviewed in Cannel et al. (2008) and Zhang et al. (2007)), have received the most attention since their discovery 25 years ago (Lee et al. 1993; Wightman et al. 1993). MicroRNAs are an abundant class (~4,000) of short (21-25 nt) non-coding RNAs that regulate gene expression via perfect to near-perfect complementary binding to messenger RNAs, often within the 3′-untranslated region (UTR), which then trigger either mRNA cleavage or translational repression. Computational approaches have been used to dramatically increase the number of identified animal and plant miRNAs and their corresponding mRNA target sites (reviewed in Chaudhuri and Chatterjee (2007) and Yoon and De Micheli (2006)). This has led to intensive research into elucidating the biogenesis and mechanisms of miRNAs.

The untranslated regions (UTRs) of messenger RNAs have been implicated in playing important roles in post-transcriptional gene regulation in both animal and plant mRNAs (reviewed in Mignone et al. (2002)). Studies have shown that the untranslated regions, particularly the 5′-UTR, can harbour control signals, or regulatory motifs, that mediate mRNA translational efficiency, stability, and localization. Some identified control signals include upstream start codons (uAUGs) and upstream open reading frames (uORFs) (reviewed in Lovett and Rogers (1996), Meijer and Thomas (2002), and Vilela and McCarthy (2003)) and secondary structures and internal ribosome entry sites (reviewed in Pickering and Willis (2005). UTR-mediated post-transcriptional control is important in many animal biological processes including the homeostasis of iron (Rogers et al. 2002), hormones (Orso et al. 2004), and lipopolysaccarides (Cok et al. 2004). In plants, UTR-mediated post-transcriptional control is important for normal plant growth and development (Hanfrey et al. 2002; Wang and Wessler 1998).

This review will focus on the evidence for post-transcriptional control, the importance and mechanisms of post-transcriptional control through controlling mRNA translation, and highlight how bioinformatic identification of two important translational control signals, the uORFs and local secondary structures, has led to new insight into non-miRNA mediated translational control. Both these control signals will be discussed in detail with emphasis on those found in plant species.

1.2 EVIDENCE FOR POST-TRANSCRIPTIONAL CONTROL

1.2.1 Disparity between nuclear transcription rates and cytosolic mRNA levels

Post-transcriptional control is one of the major points of gene regulation. The evidence for this control was shown by the discrepancies found between transcription rates and steady state mRNA levels. Walling et al., (1986) were among the first to show that differences in transcription rates measured in isolated nuclei could not fully explain observed differences in cytosolic mRNA levels. For example, the soybean seed protein gene, E1.9, showed relatively steady levels of mRNA throughout embryogenesis even though there was more than a three fold increase in transcription rate in late embryogenesis. It was suggested that reduced mRNA stability could be responsible for the constant steady state mRNA levels. Subsequently, many groups have documented similar discrepancies between nuclear run-on transcription and steady state mRNA levels for a variety of plant genes. Examples include the E17 gene of tomato (Lincoln and Fischer 1988), the ribulose biphosphate carboxylase small subunit (rbcS) gene in soybean (Shirley and Meagher 1990), and the alcohol dehydrogenase-1 (ADH1) gene of maize (Rowland and Strommer 1986).

1.2.2 Cytosolic mRNA and protein levels do not always correlate

Further evidence of post-transcriptional control is seen when protein abundance does not correlate with steady state levels of mRNA. Gygi et al., (1999) showed that for many yeast genes the levels of mRNA are not highly correlated with protein levels. Briefly, they analysed the mRNA and protein levels of over 100 genes of the yeast *Saccharomyces cerevisiae*, and showed that for a representative sample of genes (73 of 106 genes or ~70%), where mRNA levels were below 10 copies/cell, the protein levels varied by as much as 30 fold (Figure 1.2). Indeed, a comparison of selected mRNA and protein abundances in human liver (Anderson and Seilhamer 1997) showed that a correlation between mRNA and protein abundance is seen only for secreted proteins (29 of 50 most abundant proteins). In plants, polyribosome-loading analyses have revealed alterations in mRNA translation following numerous environmental stimuli (Kawaguchi and Bailey-Serres 2002). Many studies have revealed that there are control signals all along the mRNA that determine differential rates of translation, and that the untranslated regions are a major contributor to these signals (Mignone et al. 2002).

1.2.3 Untranslated regions important for post-transcriptional control

One important step in post-transcriptional control is the regulation of mRNA translation (Figure 1, Step 4). The efficiency of translation of eukaryotic mRNA can be modulated by signals that influence the three steps of translation: initiation, elongation, and termination. Previously, much attention focused exclusively on 5′-UTR control signals in controlling translation initiation in eukaryotes (Vilela et al. 1999). It is now clear that 5′-UTR control signals also play a major role in modulating translational efficiency by applying its effect on the elongation and termination steps (Mignone et al. 2002). Examples include the carbamoyl phosphate synthetase (CPA1) mRNA of *Saccharomyces cerevisiae* (Delbecq et al. 1994), arg-2 mRNA of *Neurospora crassa* (Fang et al. 2004), and the mammalian S-adenosylmethionine decarboxylase (*SAMDC*) mRNA (Raney et al. 2000).

1.2.4 Control signals in untranslated regions mediate translational control

In mRNAs that encode important regulatory proteins (e.g., transcription factors), control signals are often found (~28%) in longer than average 5′-UTRs (>100 nucleotides), and suggest the need of these proteins to be finely and strongly regulated (Hayden and Jorgensen 2007; Kozak 1987a; Mignone et al. 2002; Tran et al. 2008). These long UTRs contain signals that mediate both negative and positive translational control (Kozak 1987a), and include upstream start codons and open reading frames (uORFs), stable secondary structures (e.g., stem loops), internal ribosome entry sites (IRESs) and various *cis*-acting elements that are bound by RNA-binding proteins (Mignone et al. 2002).

The upstream start codons are the first signals identified for post-transcriptional control and they can be found in the 5′-UTR of eukaryotic mRNA (Kozak 1987a). Upstream start codons that are in a good sequence context are often followed by a termination codon, thereby creating uORFs, the next signal discovered. Subsequenty, several other control signals were identified (e.g., secondary structures and IRESs). The two most studied control signals, uORFs and secondary structures, which can be found in the 5′-UTR, will be discussed in more depth below.

1.3 UPSTREAM OPEN READING FRAMES (uORFS)

1.3.1 Definition of uORFs

An upstream open reading frame (uORF), as found in some monocistronic eukaryotic mRNAs, is defined as an open reading frame demarcated by a 5′-UTR start codon (uAUG) followed by a downstream and inframe stop codon (uSTOP). More than one uORF can exist in the 5′-UTR, and depending on the position of the uSTOP codon uORFs can overlap other downstream uORFs and even the main coding region. In eukaryotic mRNAs, uORFs are important control signals that can regulate the translation efficiency and stability of the main coding region. They are over-represented in animal genes involved in the developmental processes (Mignone et al. 2002), mitochrondrial import and methlytransferase genes in Drosophila (Hayden and Bosco 2008), transcription factors in rice and Arabidopsis (Hayden and Jorgensen 2007), and genes involved in phosphorylation in higher plants (Tran et al. 2008).

1.3.2 Types of uORFs

Two types of functional uORFs have been described with a demonstrated activity either *in vitro* or *in vivo*: a) uORFs encoding bioactive peptides (Crowe et al. 2006; Hayden and Bosco 2008; Hayden and Jorgensen 2007; Iacono et al. 2005) that either affect the translational machinery or have biological roles other than reducing the translation of the main ORF, and therefore can be described as sequence-dependent, and b) sequence-independent uORFs. A sequence-dependent uORF encodes a small peptide, and some of these uORF-encoded peptides have been shown to directly affect translation via either ribosomal stalling during translation of the uORF or termination of translation by inhibiting the peptidyl transferase activity of the ribosome and thus peptide bond formation (Gaba et al. 2001; Luo and Sachs 1996). The exact mechanism of interaction between sequence-dependent uORFs and the translating ribosome is still unknown, possibly attributed to the ribosome heterogeneity (Byrne 2009). For sequence-independent uORFs, the uORF-encoded peptide is not important for translational control, but other factors like uORF recognition, length, stop codon environment, and the downstream intercistronic sequence (length and structure) can affect reinitiation efficiency at the downstream ORF (Meijer and Thomas 2002; Vilela and McCarthy 2003). Sequence-independent uORFs can also indirectly affect translation by allowing ribosomes to bypass inhibitory stem structures (Hemmings-Mieszczak et al. 2000) or activate dormant IRESs (Yaman et al. 2003) via conformational changes induced by the translation of the uORF. These distinct mechanisms of translational control were proven to be important through *in vitro* genetic (mutational analyses) and biochemical (toe-printing) assays (Gaba et al. 2001).

1.3.3 uORFs can influence mRNA stability

Studies in yeast have indicated that both sequence-dependent and sequence-independent uORFs can destabilise mRNAs. Currently, two known pathways have been described for uORF-mediated destabilisation. The first is the nonsense-mediated decay (NMD) pathway (Ruiz-Echevarria and Peltz 2000). Ruiz-Echevarria and Peltz (2000) hypothesised that mutations in the mRNA 5′-UTR that create a uORF trigger the NMD pathway and lead to decapping of the mRNA. Indeed, two recent studies in Arabidopsis have provided experimental evidence for the NMD pathway, showing that the length of a uORF (>153 nt) was important to efficiently trigger the NMD pathway (Nyiko et al. 2009), and that the presence of uORFs were associated with higher levels of uncapped mRNAs (Jiao et al. 2008).

Alternatively, mRNA destabilisation can occur via the termination dependent decay pathway (Vilela et al. 1999). In this pathway, the 40S ribosomal subunits are released from the mRNA due to features such as the uORF stop codon environment (i.e., a GC rich region surrounding the uORF stop codon) or short intercistronic sequence containing a secondary structure. Release of 40S ribosomal subunits prevents them from reinitiating at a downstream ORF, usually the main ORF, and the mRNA becomes susceptible to decay. The mechanisms underlying both uORF nonsense-mediated decay and post-termination mediated decay remain unclear.

1.3.4 Start codon context of uORFs

The sequence context around uORFs differs strikingly from functional initiator codons (Kozak 1987a). In mammals, the consensus sequence around functional initiator codons is GCC(A/G)CCAUGG, and the most conserved nucleotides are the purines, usually adenine, in position –3, and guanine in position +4. The nucleotides in a sequence context are designated at positions relative to the start codon (A of AUG is at position +1). In higher plants, the consensus sequence is caA(A/C)aAUGGCg with small variations between monocots and dicots, where letters in uppercase are more highly conserved (Joshi et al. 1997). The importance of the two positions, a purine in the -3 and a guanine in +4, has been demonstrated experimentally as mutations in these positions result in the greatest reduction in translation efficiency (Kozak 2005). Therefore, start codons with a sequence context containing a purine in the -3 position and a guanine in +4 position are referred to as optimal. However, it has been shown that even uORFs with a sub-optimal sequence context can be recognised efficiently by scanning ribosomes both *in vivo* and *in vitro* (David-Assael et al. 2005), indicating that uORF recognition is not solely based on the sequence context surrounding the uORF start codon.

Generally, uORFs do not have an optimal sequence context (Hayden and Jorgensen 2007; Tran et al. 2008). It is believed that leaky scanning of ribosomes bypasses non-functional upstream start codons due to their different sequence context. It is not known how ribosomes discriminate between different consensus sequence contexts. However, leaky scanning can be a deliberate process that results in the production of multiple different proteins from one mRNA transcript (Mignone et al. 2002).

1.3.5 Translation of messenger RNA containing functional uORFs

Upstream ORFs are translated in the same way as main ORFs. In the initiation step, the messenger RNA is loaded with the small 40S ribosomal subunits and circularized by the interaction between the poly-A tail and the 5′-cap structure (Kozak 2005). For mRNAs containing a functional uORF, a bound 40S ribosomal subunit will scan in the 5′ to 3′ direction until it reaches the first uORF start codon. The sequence context around the uORF start codon determines the frequency at which the uORF start codon is recognised by the 40S ribosomal subunit as a functional initiator codon (Section 1.3.4). For example, a uORF with a weak sequence context will be initiated at a much lower frequency compared to a uORF with an optimal sequence context. At the uORF start codon, the large 60S ribosomal subunit will bind to the momentarily paused 40S ribosomal subunit to form an active 80S ribosome. At this stage, the translation process enters the elongation step and begins the synthesis of a uORF peptide. The newly formed peptide, if sequence-dependent (Section 1.3.2), may then act on the 80S ribosome and cause it to stall during the elongation or termination step of translation. The translation process is then halted, and the mRNA may be a target for decay. If the uORF peptide does not act on the 80S ribosome then the small peptide will terminate at the uORF stop codon, the 60S ribosomal subunit detaches, and the 40S subunit either resumes scanning and re-initiates translation at a downstream start codon or leaves the mRNA, thereby preventing translation of the main ORF (Mignone et al. 2002).

The resumption of scanning may be dependent on the interaction between eukaryotic initiation factor 4F (eIF4F) and the scanning ribosome being maintained while the ribosome translates the uORF (Poyry et al. 2004). If the 40S ribosomal subunit resumes scanning after translating the uORF, the level of translation is dependant on the frequency of ribosomal re-initiation at a downstream main start codon. Re-initiation by post-terminated ribosomes (ribosomes that have completed translating the uORF) is not an efficient mechanism, and is modulated by many properties (e.g., uORF recognition by the 40S ribosomal subunit, length, and intercistronic distance, and stop codon environment) of the sequence-independent uORF (Section 1.3.2).

1.3.6 Approaches for identifying functional uORFs

Identifying uORFs involved in regulation of gene expression remains a challenge (Gaba et al. 2001; Spevak et al. 2006; Wu et al. 2007). Recently it has been estimated that it would take 20 man-months to find a single functional uORF by random selection and testing of yeast mRNAs (Selpi et al. 2006). To overcome this problem Selpi *et al.* (2006) used an artificial intelligence approach called inductive logic programming to identify likely functional uORFs. The approach used rules based on background knowledge of uORFs in yeast mRNAs and as such may not be applicable to other organisms. This limitation was also noted by Cvijovic *et al.* (2007) in their rule based approach for finding putative regulatory uORFs in the yeast genome.

A comparative approach for finding functional uORFs that is not limited to any particular species was developed that uses evolutionary conservation of uORF peptide sequence (Crowe et al. 2006; Hayden and Bosco 2008; Hayden and Jorgensen 2007; Tran et al. 2008). These homology-based approaches tend to favour the identification of bioactive uORF peptides (sequence-dependent uORFs) that either act on translation or have other biological roles other than reducing the translation of the main ORF.

A recent approach for finding sequence-independent uORFs was recently described (Kochetov et al. 2008). Kochetov *et al.* (2008) selected human mRNAs with specific sequence organisation (i.e., uORF overlapping the main ORF) that could facilitate reinitiation at downstream start codons. If the downstream start codons were nested in-frame with the main ORF then potentially N-terminally truncated variants of the main protein could be produced via reinitiation. Kochetov *et al.* (2008) reported that 297 out of 754 mRNAs (39% of the sub-sample) contained this specific sequence organisation with an average intercistronic spacer of 66±77 nt, which provides sufficient space for reinitiation. This novel approach highlights another way in which uORFs can be functional via the generation of novel protein isoforms.

1.3.7 Plant uORFs

The frequency of reported uORFs in plants is rare in comparison to mammalian systems. Early estimates on the number of characterised uORFs in plants were less than 100 (0.3%) (Tran et al. 2008), and most are described in four cereal transcriptomes. They include the uORFs of the S-adenosylmethionine decarboxylase gene (*SAMDC*) in both monocots and dicots (Franceschetti et al. 2001; Hu et al. 2005; Tassoni et al. 2007), rice *myb7* gene (Locatelli et al. 2002); transcription factors such as maize *Opaque-2* (Lohmer et al. 1993), maize *R* (Wang and Wessler 1998), and maize *Lc* (Wang and Wessler 2001). Also, uORFs have been identified in dicot plant genes that include transcription factors *MtHAP2-1* (Combier et al. 2008), *AtbHLH* (Imai et al. 2006), *AtB2/AtbZIP11* (Rahmani et al. 2009; Wiese et al. 2004; Wiese et al. 2005), *ABI3* (Ng et al. 2004), and *CpbZIP2* (Ditzer and Bartels 2006); tonoplast transporter *AtMHX* (David-Assael et al. 2005), phosphoethanolamine N-methyltransferase *AtPEAMT* (Tabuchi et al. 2006), ornithine decarboxylase gene *ODC* (Kwak and Lee 2001); and auxin responsive factor genes *ETT* and *MP* (Nishimura et al. 2005). These characterised uORFs (<0.3%) in plants are much lower than the estimated number of genes that contain uORFs, which can vary from 11% (Pesole et al. 2000) to 60% (Hayden and Jorgensen 2007). Of the aforementioned plant uORFs, the sequence-dependent small uORF of *SAMDC* and the sequence-independent uORF of maize *Lc* are both well-described. Therefore, both these uORFs and other novel uORFs will be discussed in more detail.

1.3.8 Mechanisms of regulation by the plant *SAMDC* and Lc uORFs

The *SAMDC* gene encodes the *S*-adenosylmethionine decarboxylase protein, a key enzyme in the biosynthesis of polyamines (i.e., spermidine and spermine). Polyamines are multivalent cations implicated in a wide range of cellular physiological processes including chromatin organisation, mRNA translation, cell proliferation, and apoptosis (Hanfrey et al. 2002). The most remarkable characteristic of all expressed plant *SAMDC* genes is the presence of a long 5′-UTR (~500 bp) containing a pair of highly conserved uORFs that overlap by one nucleotide (Franceschetti et al. 2001). The 5′ tiny uORF consist of two or three codons and the 3′ small uORF encodes 50-54 codons. The small uORF is highly conserved between monocot, dicot, and gymnosperm species, and therefore strongly suggestive of a conserved regulatory mechanism in translation (Hanfrey et al. 2002).

To evaluate the function of the tiny and small uORFs, Hanfrey et al. (2002) placed the Arabidopsis *SAMDC1* 5′-UTR between the plant viral cauliflower mosaic virus 35 S promoter and the *Escherichia coli* glucuronidase (GUS) reporter gene. Transgenic tobacco plants were generated expressing this reporter construct and mutant constructs generated by site-directed mutagenesis of the small uORFs. By relating GUS activity to the level of corresponding *GUS* mRNA, it was determined that the small uORF is responsible for the translational repression of *SAMDC* (Figure 1.3). The stunted mutant phenotype of derepressed tobacco plants indicates that translational regulation of *SAMDC* is essential for normal plant development. Hu *et al.* (2005) also showed evidence for the *SAMDC* 5′-UTR in transcriptional control, but did not test if the *SAMDC* uORFs were solely responsible.

Unlike the SAMDC small uORF, the other tiny uORF has controversial functions. One report concludes that in response to high polyamine levels, the tiny uORF acts to maintain normal levels of polyamines by favouring the preferential recognition of the repressive small uORF either by leaky scanning or -1 frameshifting (Hanfrey et al. 2005). Another report suggests that the tiny and the small uORFs have the same affect in response to high polyamine levels; that is downstream translational repression (Hu et al. 2005).

A good example of a sequence-independent uORF in plants is the 38-codon uORF found in the long 5′-UTR (235 nt) of the maize *R* gene, *Lc*, encoding a transcriptional activator of the anthocyanin biosynthetic pathway (Damiani and Wessler 1993). Damiani and Wessler (1993) showed that the uORF decreased translation 25 to 30 fold in an *in vivo* particle bombardment assay. Furthermore, co-bombardment experiments showed that the uORF decreased translation in *cis* and not in *trans*, possibly indicating that the uORF peptide was not directly involved in translational control. Rather, the codon usage within the uORF was important for the stalling of ribosomes as nonsynonymous codon changes showed higher translation efficiency than synonymous codon changes. Damiani and Wessler (1993) concluded that the uORF translational control mechanism prevented the overexpression of the Lc protein, which could otherwise result in developmental defects or lethality in plants.

A more recent study on the maize *Lc* gene also showed that the uORF was involved in translational control (Wang and Wessler 1998) but via a different mechanism to that suggested by Damiani and Wessler (1993). They demonstrated that minor and major changes (e.g., point and frame-shift mutations) in the uORF sequence did not affect the repression activity of the uORF, suggesting that the uORF codons and the encoded peptide were not directly involved in the translational control. Instead, they found that ribosomes that translated the uORF did not reinitiate efficiently downstream (~30%) possibly due to multiple stop codons in the intercistronic sequence, as random generated intercistronic sequences improved reinitiation frequency almost three fold. However, it is not known if and how stop codons prevent post-terminated ribosomes from resuming scanning.

1.3.9 Novel regulatory mechanisms by plant uORFs

The *Medicago truncatula MtHAP2-1* transcript, which encodes a HAP2-type transcription factor of the CCAAT-box-binding family (CBF), is post-transcriptionally regulated by microRNA169 (miR169) (Combier et al. 2006) and uORF1p (Combier et al. 2008) during nodule growth and development. This is the first example of post-transcriptional regulation by a uORF and a microRNA, which targets the 5′ and 3′-UTRs respectively. Combier at al. (2006) showed that the small uORF1 is translated more efficiently during nodule growth when alternative splicing of *MtHAP2* retains a large intron (*MtHAP2-1*) positioned 5′ of uORF1. Also, the over-expression of *MtHAP2-1* without uORF1 results in nodule developmental defects, indicating that uORF1 plays a key role in regulating nodule development. Finally, western blot detection of HA-tagged uORF1p confirmed that uORF1 is translated and specifically binds *in trans* to *MtHAP2-1* exon E1 and/or E2 in the 5′-UTR, according to RNA pull-down assays of exon deleted *MtHAP2-1*. These major findings by Combier at al. (2006) show for the first time that a uORF-encoded peptide can bind specifically *in trans* to the 5′-UTR, instead of the ribosome translational machinery as seen in some *cis*-acting uORFs (Lovett and Rogers 1996). It is believed that uORFp1 promotes transcript degradation in an unknown manner, thus down-regulating gene expression. It remains to be determined if uORF1p can also bind *in cis*.

Another novel regulatory mechanism by uORFs is the positive translational control mechanism via 5′-UTR remodelling (Yaman et al. 2003). Yaman et al. (2003) showed that a uORF in the cat-1 transcript, which encodes an arginine/lysine transporter, uses the innate ability of the 80S ribosomal subunit to unwind secondary structures during uORF translation, resulting in the activation of a dormant 3′-IRES induced by new long range RNA interaction with the 5′-UTR. The IRES is a *cis*-acting mRNA element which facilitates cap-independent translation, a more efficient form of translation that allows the 40S ribosomal subunit to engage the mRNA from within the 5′-UTR, thus avoiding the 5′-cap structure and the need for other recruited proteins (e.g., eIF4F) that are required for ribosome binding.

The above reports of novel regulatory mechanisms by uORFs indicate that they have diverse regulatory mechanisms in reducing translation of the main ORF. More mechanisms of regulation by uORFs may yet be found as new uORFs have been identified in different species via computational approaches (Hayden and Bosco 2008; Hayden and Jorgensen 2007; Kochetov et al. 2008; Tran et al. 2008).

1.4 MESSENGER RNA (mRNA) STRUCTURES

1.4.1 Introduction to RNA, structure, and stability

RNA is a similar molecule to DNA and differs in that it uses uracil instead of thymine as a base structure, it does not have a complementary partner, and its backbone is more flexible allowing it to bind to itself at complementary regions (Alberts et al. 2002). The complementary binding between paired nucleotides, known as base-pairing, can occur over short or long range distances along the RNA molecule. Such base-pairing interactions allow for the formation of RNA secondary and tertiary (interactions between two or more secondary structures) structures (Alberts et al. 2002). The base-pairing interactions tend to occur in "stacks" (multiple consecutive base pairing) as this contributes to a more stable structure (Figure 1.4A). The base-pair stacking can be interspersed by unpaired nucleotides to form symmetrical and asymmetrical structures called loops and bulges (Alberts et al. 2002) (Figure 1.4C and D).

One class of RNA called the messenger RNA can form local secondary structures that span across small domains of the RNA molecule rather than from end to end of the molecule as seen with non-coding RNAs. In the cell, eukaryotic mRNA is synthesised from the transcription of genes in the nucleus (Figure 1.1). The mRNAs, bounded by ribosomes, are then exported to the cytoplasm where they undergo maturation (splicing, 5′-capping, and 3′-polyadenylation) and circularisation to form a stable complex for the translational machinery to bind and translate the mRNAs into proteins.

The stability of a RNA molecule is dependent on several interactions that include the intramolecular base-pairing by covalent bonding of the sugar-phosphate backbone, intermolecular base-pairing (Watson-Crick) by hydrogen bonding, tertiary interactions between RNA secondary structures, and ionic interactions (e.g., Mg^{2+}) that stabilise the RNA tertiary structure (Chen 2008). The stability of a RNA molecule can be determined by melting experiments, which determine the amount of free energy (measured in calories per molecule) that is required to unfold the molecule (SantaLucia and Turner 1997; Xia et al. 1998). Also, RNA stability can be estimated from known thermodynamically characterised structures (Davis and Znosko 2007).

The structure of a RNA molecule is typically determined experimentally by X-ray crystallography (Ferre-D'Amare and Doudna 2001; Mooers 2009; Pikovskaya et al. 2009) and by nuclear magnetic resonance (NMR) (Furtig et al. 2007; Furtig et al. 2003), and computationally by phylogenetic analysis (James et al. 1989; Shapiro et al. 2007). Phylogenetic analysis, which searches for compensatory mutations in many evolutionary related sequences, is the gold standard for RNA structure prediction as it is technically less difficult and is considered very reliable if enough sequences are available for comparison (Mathews and Turner 2002; Zuker et al. 1991).

1.4.2 Types of secondary structures

The local base pair interactions that can occur in mRNAs are important for secondary structure formation that is required for correct translational control. There are two types of RNA base pairing: canonical and non-canonical (Lemieux and Major 2002). Canonical base pairing refers to the Watson-Crick base pairing, where adenine (A) pairs with uracil (U) and guanine (G) pairs with cytosine (C) by hydrogen bonding. Conversely, non-canonical base pairing refers to any other base pair combination other than the Watson-Crick base pairs (i.e., AG/GA and AC/CA). GU base-pairs should be considered as canonical as they occur frequently in RNA structure (Gautheret et al. 1995), and have comparable thermodynamic stability and structure/shape to Watson-Crick base pairs (Varani and McClain 2000), and furthermore their energetic contributions to annealing energy have been measured in many different contexts so that they are routinely used in RNA structure prediction programs (Mathews and Turner 2006; Pavesi et al. 2004).

Common RNA secondary structure motifs (re-occurring patterns) that form as a result of base pairing include stems, loops, and bulges (Figure 1.4). These motifs represent the building blocks of RNA secondary structures, and allow the formation of complex structures, such as the stem-loop. Several different types of stem-loops exist: a plain stem-loop and a stem-loop with internal loop(s) and/or bulges(s). A plain stem-loop (Figure 1.4B) is simply a secondary structure consisting of a stem and a loop region. More complicated stem-loops build on the plain stem-loop to include internal loops and/or bulges (Figure 1.4C and D). The internal loops and bulges of stem-loops are often critical for binding to RNA-binding proteins (Ke et al. 1998).

Secondary structures can link together to form composite structures and can also interact with one another to form tertiary structures (Batey et al. 1999). The interactions between two or more secondary structures can be short or long-ranged. For long ranged interactions, secondary structures that are separated by a long distance in a RNA sequence are brought closer together by convoluted folding of the RNA such that base-pairing can occur. An excellent example of a tertiary structure is the “pseudo-knot”, a knot-shaped three-dimensional structure containing at least two intercalated stem-loop structures (reviewed in Giedroc and Cornish (2009), Brierley et al. (2008), Giedroc et al. (2000), and Hilbers et al. (1998)).

1.4.3 Position and stability of secondary structures

It has been shown that a very stable stem-loop (ΔG >25 kcal/mol) in mRNA, not just moderately stable (ΔG ~18 kcal/mol), can cause ribosomes to completely stall proximal to the stem-loop, and in turn stop translation (Niepel et al. 1999). A stem loop of moderate stability can also impede the ribosomal scanning process. In this case, the scanning process is slowed down by the inhibitory structure of the stem-loop, resulting in decreased ribosomal loading and downstream translation (Niepel et al. 1999). Furthermore, the position of secondary structures, regardless of their stability, can also repress downstream translation by influencing the binding of the 40S ribosomal subunit near the 5′ end; however, once the 40S subunit is engaged it can disrupt base-paired structures downstream (Baim and Sherman 1988). This has been shown in stem loops of mRNA from mammalian cells (Kozak 1989) and in *CYC1* mRNA of *S. cerevisiae* (Baim and Sherman 1988). There have been relatively few reports of stem-loop mediated translational control in plants, but the stem-loops found in mRNA from maize transcriptional activator *Lc* (Wang and Wessler 2001), tobacco *psbA* (Wang and Wessler 1998), pollen *ntp303* (Hulzink et al. 2002), and two chloroplast mRNAs (Monde et al. 2000; Zou et al. 2003) are well described.

1.4.4 Secondary structures in mRNAs from non-plant organisms

Translational repression by a stem-loop located in the 5′-UTR has been seen in α -globin mRNA (Kozak et al. 1994), amyloid precursor mRNA (Rogers et al. 2002), and ferritin mRNA (Muckenthaler et al. 1998). Of more interest, the ferritin mRNA contains a conserved stem-loop structure known as the iron responsive element (IRE). The iron resposive element can be found in the 5′ or 3′-UTRs of various mRNAs coding for proteins involved in celluar iron metabolism (Ke et al. 2000). The interaction of the IRE with the iron-responsive protein (IRP) blocks entry of the 40S ribosomal subunit and thereby reduces translation of the ferritin mRNA in times of low iron. Therefore, the IRP-IRE mechanism modulates the translation of the mRNA according to the amount of iron present in the cell. Stem-loops that deviate from the classical IRE consensus sequence are also able to bind to IRPs (dos Santos et al. 2008). Dos Santos et al. (2008) showed that the 3′-UTR IRE-like stem-loop of human alpha-hemoglobin-stabilizing protein (AHSP) mRNA was shown to co-immunoprecipitate with IRPs in a manner that is dependent on the stem-loop structure. Conversely, small molecules other than the IRP, such as anthracyclines, a class of chemotherapeutic drugs, are also able to bind specifically around the IRE internal C-bulge residue (Canzoneri and Oyelere 2008).

Stem-loops are able to form stable composite structures known as kissing complexes through loop to loop interactions (Brunel et al. 2002; Duconge and Toulme 1999). The stem-loop identified in an internal ribosome entry site of the hepatitis C virus mRNA is capable of forming loop to loop interactions with selected stem-loops that are GC rich and contain contiguous complementary sites in the loop region (Dausse et al. 2008). Such loop to loop interactions were shown to reduce *in vitro* translation (by more than 60%) of a luciferase reporter gene. Although the authors did not provide an explanation for the reduction in translation, it can be reasonably expected that the stable loop to loop interactions prevent the IRES from functioning as an entry site for ribosomes. However, it was concluded that stem-loops can act as ligands that recognise and bind to specific regions of other stem-loops via loop to loop interactions, and that targeted stem-loop ligands have potential therapeutic value by regulating gene expression.

Another example of a RNA composite structure is the conserved selenocysteine insertion sequence (SECIS) element that consists of two stem-loops interspersed by an internal loop. The SECIS element, found in the 3′-UTR of eukaryotic mRNAs encoding selenoproteins, is part of a novel mechanism that allows ribosomes to read the UGA stop codon as a selenocysteine amino acid (21st amino acid) (Martin et al. 1996). This translational read through mechanism of the UGA stop codon requires the correct positioning of the SECIS element (51 to 111 nt downstream of UGA stop codon) and GA non-canonical base-pairing in the upper stem region. Recently, an evolutionarily conserved GTPase-activating protein, termed GAPsec, was shown to be necessary to support SECIS-dependent UGA read-through activity (Hirosawa-Takamori et al. 2008). Also recently is the development of a web-base tool, SECISaln, for extensive structure alignments of SECIS elements (Chapple et al. 2009).

1.4.5 Secondary structures in mRNAs from plant organisms

Wang and Wessler (2001) showed that translation of the maize *Lc* mRNA is also repressed by a stem-loop of moderate stability (ΔG of –15.6 kcal/mol) in the 5′-UTR. More importantly, this stem-loop repression activity is independent of the uORF, and therefore shows that the 5′-UTR is capable of mediating two independent levels of repression. The confirmation of the activity of the stem-loop was achieved by comparing the reporter gene expression of the wild-type and mutant *Lc* 5′-UTR constructs. It was found that the partial deletion of the hairpin (five nucleotide deletion of the stem) increases downstream reporter gene expression. This evidence supports the model that the stem-loop is responsible for decreasing ribosomal loading and repressing downstream translation (Figure 1.5). The translational control of *Lc* by both a 5′ stem-loop and a uORF has not been reported before in other species. It remains to be determined if the dual mechanism of translational control by both a 5′ stem-loop and a uORF in *Lc* is conserved in orthologous genes, and whether more such examples exists in the animal and plant transcriptome.

Hulzink et al. (2002) also reported that two predicted stem-loops (H-I and H-II) in the 5′-UTR of pollen *ntp303* transcripts are responsible for modulating translational efficiency and mRNA stability during pollen tube growth, respectively. Deletion studies revealed that a specific feature [$(GAA)_8$ repeat] in the loop part of the H-I stem-loop caused almost complete inhibition of translation (~94%), indicating that the primary sequence was important for the translation inhibition. In contrast, deletion of the H-II stem-loop predominately affects mRNA stability due to a decrease (~2 fold) in transcript accumulation.

Secondary structures in the 3′-UTR can also reduce translation efficiency (Niepel et al. 1999). Niepel et al. (1999) found that the repressive effect requires a stem-loop of much greater stability than that required in the 5′-UTR, and is only observed when the stem-loop is placed adjacent to the stop codon and not further downstream. These results suggest that the 5′-UTR is more sensitive to the repressive effect of secondary structures. Interestingly, secondary structures in the 3′-UTR do not appear to affect mRNA stability (Niepel et al. 1999). However, this could be attributed to the limited studies on secondary structures found in the 3′-UTR as at least one other example, the 3′-UTR stem-loop of *petD* (Monde et al. 2000) was found to affect both mRNA stability and translation efficiency.

The limited examples of secondary structures involved in post-transcriptional control of gene expression is likely to be attributed to the limitation of early RNA motif prediction programs, but this is changing as improved programs are beginning to emerge (Hu 2002; Pavesi et al. 2004). The advent of these improved programs will provide us with a better insight into the mechanism of action of secondary structures. It is clear, however, that secondary structures are important in determining the translational efficiencies of mRNAs in higher plants (Hulzink et al. 2002; Wang and Wessler 2001). Both the position and structural stability of secondary structures can repress downstream ORF translation. Unlike uORF, where leaky scanning can be used to bypass the activity of the uORF, secondary structures of very high stability remain a physical barrier for ribosomal scanning.

1.4.6 Secondary structures in mRNAs from plastid genes

Stem-loops in plastid transcripts can regulate chloroplast translation in spite of the differences in their translational apparatus compared to cytosolic genes (Marin-Navarro et al. 2007). For example, the plant chloroplast 5′-UTR stem-loop of *psbA* (Zou et al. 2003) and 3′-UTR stem-loop of *petD* (Monde et al. 2000), tend to affect both mRNA stability and translation efficiency. More recently, the 5′-UTR stem-loop of *psbD* mRNA (Klinkert et al. 2006) also affects translation efficiency by blocking the translation initiation site (AUG start codon).

A detailed study is required to determine how many plant chloroplast transcripts contain stable RNA secondary structures in their untranslated regions. Also of interest is the number of structures in chloroplast precursor transcripts that are post-translationally imported into the chloroplasts, as some of these structures could potentially regulate translation in a similar manner as structures found in transcripts encoded in the chloroplast genome. Currently, it is understood that secondary structures in the untranslated regions of chloroplast genes generally affect translation initiation (e.g., inhibit ribosomal scanning) (Manuell et al. 2007; Marin-Navarro et al. 2007). The identification of new RNA secondary structures in untranslated regions of chloroplast transcripts may lead to additional regulatory mechanisms of translational control by chloroplast structures.

1.5 *IN SILICO* DISCOVERY OF RNA MOTIFS

The recent advances in genomic sequencing technology have generated large amounts of biological data for functional genomics (Lewis et al. 2000). The non-coding regions of RNA have attracted interest because of the role they may play in gene regulation. In fact, there are now several public databases that specialise in non-coding RNA (Pavesi et al. 2004), and programs developed for the analysis and comparison of non-coding RNA sequences (Pavesi et al. 2004). These programs predict *cis*-acting motifs; that is, motifs that perform their biological functions at the locations where they are encoded (e.g., stem loops). This section focuses on the problems associated with predicting RNA motifs, the current state-of-the-art in RNA prediction programs, and the efficient and effective RNAProfile program for finding conserved secondary structures.

1.5.1 RNA vs DNA motif discovery

The nature of RNA motifs has challenged biologists for decades. Unlike DNA motifs, where the sequence variation is the main factor that distinguishes one motif from another, RNA motifs are much more complex. This complexity is attributed to the combination of both sequence and structural constraints, such that the primary sequence may vary, but the overall structure is much conserved (Gorodkin et al. 2001; Hu 2002; Pavesi et al. 2004). In addition, the individual bases that make up a binding site in a RNA motif can interact with each other giving rise to correlations (covariation in positions); that is, the contribution of each base to the binding affinity is no longer independent, as seen with DNA binding sites (Gorodkin et al. 2001; Stormo 2000). It is these complexities associated with RNA motifs that have made it difficult to develop reliable RNA motif prediction algorithms.

1.5.2 RNA can exist in different states

It is generally understood that a RNA molecule is kinetically controlled, meaning it can exist in different states as it traverses from an unfolded state to its native state (biologically active, Figure 1.6) (Chen and Dill 2000). Also, the *in vivo* folding process is remarkably complex and can involve electrostatic interactions, base-pairing and stacking, conformational entropy, and other molecules (e.g., RNA binding proteins) (Chen 2008). For these reasons it is difficult to accurately predict *in silico* RNA structures based on their sequence information alone (i.e., *ab-initio*).

One common solution to improving the accuracy of RNA folding algorithms that model the *in vivo* folding process based on stable base-pairing and stacking (known as thermodynamics) is to also report sub-optimal structures in the prediction of RNA structures. The rationale for this is that the optimal structure with the lowest free energy is not always the native form of a RNA molecule due to other complex interactions that change the kinetics of the folding pathway and the associated free energy of the structure. It is estimated that the native state of an RNA molecule should lie within the top 10% of the optimal structures (Zuker 1989; Zuker et al. 1991).

All RNA folding algorithms make assumptions about the state in which a RNA molecule exists. For example, the commonly used mFOLD algorithm is used to find a global RNA structure (Zuker 2003) whereas a relatively new algorithm called RNAProfile finds local RNA structures (Pavesi et al. 2004). Algorithms based on global structures are suitable for the prediction of highly structured RNAs (e.g., tRNA and rRNA). Less structured RNAs like mRNAs are more accurately predicted by algorithms that are optimised for finding local structures.

1.5.3 Early RNA motif prediction algorithms

In an important dynamic programming algorithm that was never implemented, it was shown that it is computationally infeasible [O(n^{3k}), where k and n equal the number and length of sequences, respectively] to provide an optimal multiple sequence and structure alignment, even for a small dataset of RNA nucleotide sequences (Sankoff 1985). As a result, many of the early RNA motif prediction algorithms either focused exclusively on conserving the RNA primary sequence or resorted to finding RNA motifs in a single sequence. Not surprisingly, algorithms that considered only the primary sequence in RNA motif prediction did not perform well. Examples include, among others, the consensus pattern recognition (Hertz et al. 1990) and optimised multiple patterns and pattern repeats (Lawrence et al. 1993). Algorithms which incorporated secondary structure information, such as the free-energy minimisation (Zuker and Stiegler 1981) and comparative sequence analysis (Gutell et al. 1994) performed better.

As a means to reduce algorithm complexity in multiple sequence and structure alignments, a method known as heuristics started to emerge, and was applied to the RNA alignment algorithms. Heuristic based methods define a set of commonsense rules aimed to reduce algorithm complexity at a cost of finding solutions that cannot be guaranteed as optimal. Examples of these include the stochastic context-free grammars (SCFGs) (i.e., COVE) (Eddy and Durbin 1994) and genetic algorithms (Chen et al. 2000b). Both these algorithms aim to find global RNA alignments, and as such require sequences to be similar in base-pairing and in length, and therefore are not adept at finding local consensus motifs. To deal with this problem, a simplified version of the Sankoff algorithm (Sankoff 1985), known as FOLDALIGN, was first developed to predict RNA motifs conserved in both sequence and structure in a set of unaligned sequences (Gorodkin et al. 1997). FOLDALIGN reduces the computational complexity by scoring the structure based on the number of base pairs, instead of the stacking energies, and by disallowing branch structures.

Despite these improvements, the time complexity ($O(\mathrm{L}^4\mathrm{N}^\mathrm{N})$, where O, L, and N refer to the big Oh function notation, sequence length, and the number of sequences, respectively) of FOLDALIGN is still too high for practical use (Hu 2002). To overcome this limitation, a new system called the Stem-Loop Align SearcH (SLASH) (Gorodkin et al. 2001) was developed. SLASH combines both FOLDALIGN and COVE, and the resulting time complexity has been shown to be acceptable for real-time applications for sequences of length up to ~300 nt.

1.5.4 State-of-the-art RNA motif prediction

A recent survey of the literature indicates a surge in the number of RNA alignment and motif searching algorithms over the last few years (Ferre et al. 2007). A moderate number (<10) of these algorithms include those that identify conserved structures in unaligned sequences. Examples include Dynalign, CARNAC, ComRNA, GPRM, RNApromo, and RNAProfile. Most of these algorithms have constraints on the input data (Table 1.1).

Dynalign is another simplified version of the Sankoff algorithm (Sankoff 1985) for finding the secondary structure common to two unaligned RNA sequences (Mathews and Turner 2002). As Dynalign is based on the original Sankoff algorithm, it uses dynamic programming to simultaneous align sequence and structure of RNA sequences. Thermodynamic rules are used to derive favourable structures, and heuristics is then applied to the alignment of those structures that restricts the maximum distance allowed between aligned nucleotides, thereby reducing the computational complexity. The algorithm has recently been extended to report a set of low energy structures, instead of the previous lowest energy structure, that are common in a pair of sequences (Mathews 2005). The limitations of Dyalign are: 1) the comparative sequence analysis is limited to two sequences, 2) it is still computationally expensive, and 3) it performs a global sequence and structure alignment, and therefore the pair of sequences should be of similar length and overall structure (e.g., tRNAs/rRNAs). Despite these limitations, Dynalign is an improvement over structure prediction in a single sequence.

CARNAC is an algorithm for finding a common structure shared by two homologous RNAs (Perriquet et al. 2003; Touzet 2007; Touzet and Perriquet 2004). The main points of difference of CARNAC to other RNA motif prediction algorithm are the use of both intrinsic information in each individual sequence and mutual information from the comparative analysis of stems rather than whole sequences to find a common structure. For example, the first step in the CARNAC algorithm is to find local stems (a stack of base pairs) in each sequence that meets a minimum threshold. A threshold is determined for each individual stem and is based on short-range rather than long-range base-pairing interactions so as to limit stems that may have occurred by chance. In the following steps, the algorithm searches for matchable stems between two sequences based on two criteria: a) stems must be located in similar positions as determined by consistency with ‘anchor points’, which are highly conserved regions as determined using a probabilistic based measure selection, and b) stems must have at least one position that covary (variable in sequence but maintain base-pairing). Finally, a common folding between matchable stems is performed using dynamic programming reminiscent of the Sankoff algorithm to find a pair with the lowest free energy. As CARNAC requires candidate stems to be anchored by highly conserved regions, RNA motifs that are located in the untranslated regions of mRNAs or in highly diverged homologous sequences will not be detected.

comRNA, like CARNAC, finds a common RNA structure based on stem comparisons (Ji et al. 2004). The heuristics applied in comRNA is similar to that of CARNAC, and so takes advantage of local sequence similarity information to reduce the search space and the run-time of the algorithm. The main difference is that comRNA uses graph theory (set of stems represented as nodes in a graph for stem comparison) for finding and assembling conserved stems, it can handle multiple sequences, pseudoknots are not a problem, and can report several best candidate common structures. The computational complexity of comRNA is O(M^m) (m equals number of stems), which means it is impractical on large numbers (>20) of long (>300) sequences.

GPRM (genetic programming for RNA motifs) is an algorithm for finding a consensus structural motif in a large set of coregulated RNA sequences (Hu 2003). The algorithm, based on the concept of biological evolution, is programmed into three steps that are applied iteratively in each generation. In the first step, GPRM generates a population of putative structural motifs based on user input constraints on a “training data”. The constraints specify a maximum number and length of segments (stems or loops). The training data is a collection of two datasets: a “positive” dataset of coregulated RNA sequences and a “negative” dataset of random sequences that is automatically generated using the alphabet of the “positive” dataset. In the second step, each putative structural motif is evaluated by a fitness function that measures the quality of each motif. In the last step, the motifs undergo either “mutation” to reconstruct alternative base paired structures or “crossover” to generate a new offspring by randomly selecting a segment from each parent motif. In each passing generation the population is halved to reflect the survival of the motifs with higher fitness, and eventually converging upon a single motif with the highest fitness. The GPRM algorithm has been shown to be an improvement over previous RNA prediction systems. For example, GPRM showed similar results to SLASH for finding stem-loops, but in constrast, it is more flexible in finding other complex structures such as pseudoknots, which most algorithms cannot detect. On the downside, GPRM requires the “positive” dataset to contain at least 10 sequences, and the running time is n^3 (n equals number of sequences), which means that it can compare thousands of sequences (limited to ~60,000) in the order of days. The program is expected to be improved further by incorporating background knowledge such as thermodynamic or phylogenetic information.

RNApromo (RNA prediction of motifs), is a stochastic context-free grammars-based method for finding local structural RNA motifs in sets of unaligned RNAs (Rabani et al. 2008). Stochastic context-free grammars are a class of probabilistic models used for modeling RNA secondary structure motifs, and has shown promising results in discovering RNA motifs in non-coding RNAs (Yao et al. 2006). There are two major parts to the motif discovery scheme of RNApromo. In the first part, thermodynamically stable structures are identified in each RNA input sequence using ViennaRNA (Hofacker et al. 1994) set in a sliding window of 200 bp (with 100 bp overlap) to avoid the low accuracy of folding long sequences and to reduce running time. Those identified structures that appear in as many input RNA sequences as possible are selected and then iteratively refined by SCFG-modelling, which includes a parameter estimation step that helps the algorithm converge to a final set of RNA structures. RNApromo claims to overcome the limitations of existing SCFG programs by integrating thermodynamic considerations and restricting the search space to a predefined and limited number of structures for each RNA input, making it feasible to scan large RNA sets or long RNA sequences. The computational complexity of RNApromo is estimated to be $O(L^2)$ (L is the length of the RNA sequence), and the algorithm has been tested on 1) 3′-UTRs of fast- and slow-decaying yeast mRNAs, 2) 5′- and 3′-UTRs of co-localised mRNAs in mouse neurons and fly embryos, and 3) plant and animal pre-microRNAs.

One of the most recent RNA motif prediction algorithm is RNAProfile: an algorithm for finding conserved secondary structures in unaligned RNA sequences (Pavesi et al. 2004; Pavesi et al. 2006). The program is split into two parts: selecting initial candidates and finding similar regions. In the first part, the program takes as input a set of related RNA sequences expected to share a conserved stem-loop motif, and begins by finding regions in each sequence that can fold into a stem-loop. To maximise the number of stem-loop candidates, the stem-loops only need to fold with a lower free energy value than the unfolded state (ΔG of 0). In the second step, an iterative process is used to evaluate all the candidate stem-loop regions. In each iteration, a set of stem-loop regions (one from each sequence) are compared in a progressive pair-wise alignment, to build a group of alignments, each described by a profile (represents a consensus structure) and scored with a suitable function *S(M)* that reflects the degree of sequence and structure similarity between the stem-loop regions comprising it. Only the best profile scores are kept and built on in further iterations in order to find the group of regions that builds the best multiple alignment (Figure 1.7). The RNAProfile program uses heuristics to keep the computational complexity to a minimum, which has been estimated to be *O*(n), making it feasible for long sequences (up to 2 kb). The other advantages of RNAProfile is that it uses energy-based rules (thermodynamics) to fold candidate regions, which is considered reliable for stem-loop structures that are small in size (15 to 50 nt); it has limited constraints on the input data; and good results were obtained for the test cases of the animal ferritin IRE, selenocysteine insertion sequence (SECIS), and ribonuclease P (RNase P). The main drawback of RNAProfile is the inherent lack of support for the prediction of complicated structures that includes pseudoknots.

1.6 SUMMARY AND AIMS OF THE PRESENT STUDY

Post-transcriptional control of gene expression in plants is not well understood. In the past, much attention has been directed to transcriptional control of gene regulation, and only recently have UTRs of mRNA been given the attention they deserve as major players in post-transcriptional control.

There are many well-characterised examples of UTR-mediated post-transcriptional control in mammalian systems. Most of these examples describe uORFs in the 5′-UTR that control the efficiency of mRNA translation. Other examples that are less described, in both mammalian and plant systems, involve secondary structures found within the 5′- and 3′-UTRs that decrease ribosomal loading, and hence translation. It is clear from these examples that uORFs and secondary structures are important features found within the UTRs that control gene expression. In particular, there is strong evidence of uORFs being highly conserved across diverse taxonomic groups (Franceschetti et al. 2001; Hu et al. 2005; Tassoni et al. 2007). It is possible that high conservation exists for certain secondary structures, and their discovery may now be possible due to recent advances in RNA motif prediction algorithms and the availability of large mRNA datasets. These provide a strong foundation for the study of UTRs in regulating post-transcriptional gene control.

Recently, a database of over 30,000 full-length cDNA clones of japonica rice has been made publicly available at the Knowledge-based Oryza Molecular Biological Encyclopedia (KOME, http://cdna01.dna.affrc.go.jp/cDNA) (Kikuchi et al. 2003). This provides an excellent resource for data-mining plant-specific UTRs. In addition, the TIGR gene indices database (http://compbio.dfci.harvard.edu/tgi/) contains large datasets of tentative contigs (TC) for wheat (44,954), barley (23,176), and maize (56,687). This additional resource provides an excellent opportunity to take advantage of the power of comparative RNAomics to identify conserved 5′-UTR signals in other cereals. For signals that cannot be identified by sequence similarity alone, RNA-motif prediction algorithms will be used to identify potential structural elements. These predictions will be evaluated by using a combined approach of site-directed mutagenesis and luciferase assay measurements.

This research aims to improve the understanding of plant 5′-UTR-mediated post-transcriptional control at the level of mRNA translation. Some unanswered questions include how prevalent secondary structures and conserved uORFs are in the cereal transcriptomes, and what classes of genes are being regulated by them. This will be addressed in a general way through bioinformatics and then more specifically using molecular techniques to test a small sample of genes, which will be chosen based on their potential importance in controlling translation. The knowledge gained will have important practical applications.

Table 1.1. Comparison of programs that predict common structures in unaligned sequences

Features	Dynalign	CARNAC	comRNA	GPRM	RNApromo	RNAProfile
Complexity	$O(M^3N^3)^a$	$O(N^2)$	$O(M^n)$	$O(N^3)$	$O(L^2)$	$O(N)$
Input data constraints	A pair of sequences only.	A pair of sequences only.	Maximum of 20 sequences of less than 300 nt.	Minimum of 10 sequences required. Must specify the maximum number and length of “segments” (stems/loops).	A predefined and limited number of structures in each input RNA sequence.	None.
Detectable structures	Secondary structures without pseudoknots.	Secondary structures without pseudoknots.	Secondary structures including pseudoknots.	Secondary structures including pseudoknots.	Secondary structures without pseudoknots.	Secondary structures without pseudoknots.
Alignment type	Global	Local and Global	Local and Global	Local	Local	Local
Tested datasets	tRNAs, 5S rRNA, and Drosophila R2 3'-UTRs.	Rnase P RNA and 16S rRNA.	Bacterial *α* operon 5′-UTRs, bacterial S15 5′-UTRs, mosaic viral 3′-UTRs, and bacterial G-box sequences.	16S rRNA, IREs, mosaic viral 3′-UTRs.	3′-UTRs of fast- and slow-decaying yeast mRNAs, 5′- and 3′-UTRs of co-localised mRNAs in mouse neurons and fly embryos, and plant and animal pre-microRNAs.	IREs, SECIS, Drosophila Nanos mRNA 3′-UTR, and RNase P RNA.
Source	(Mathews and Turner 2002; Mathews 2005)	(Perriquet et al. 2003)	(Ji et al. 2004)	(Hu 2003)	(Rabani et al. 2008)	(Pavesi et al. 2004)

O – Big Oh function notation
[a] N – Length of the shorter sequence
M – Maximum number of stems
N/n – Number of sequences
L – Length of sequence

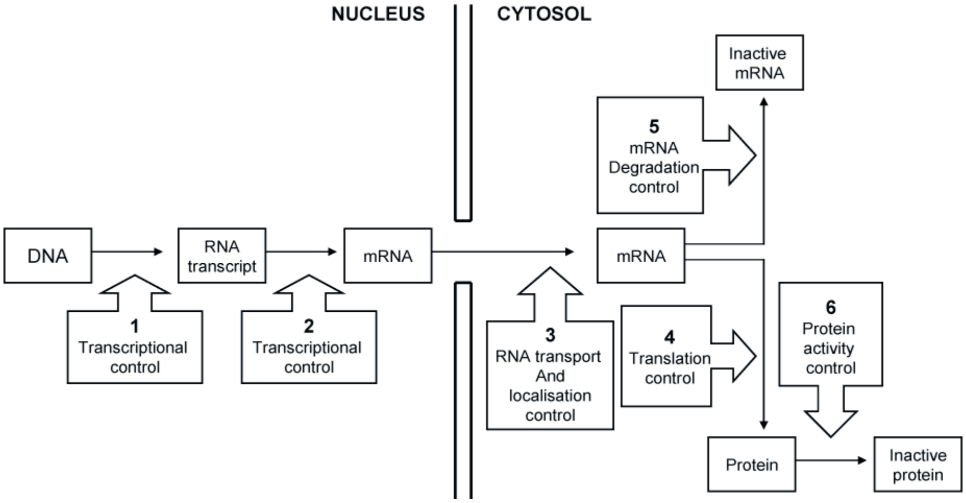

NUCLEUS
CYTOSOL
DNA
RNA transcript
mRNA
1 Transcriptional control
2 Transcriptional control
3 RNA transport And localisation control
mRNA
4 Translation control
5 mRNA Degradation control
Inactive mRNA
6 Protein activity control
Protein
Inactive protein

Figure 1.1 Six steps at which eukaryotic gene expression can be controlled. Image adapted from Alberts et al. (2002).

NOTE:
This figure is included on page 34
of the print copy of the thesis held in
the University of Adelaide Library.

Figure 1.2 Correlation between protein and mRNA levels for 106 genes (Δ) in yeast growing at log phase with glucose as a carbon source. The inset box shows the variation for low abundance genes (o, 0-10 mRNA/cell) (Gygi et al. 1999).

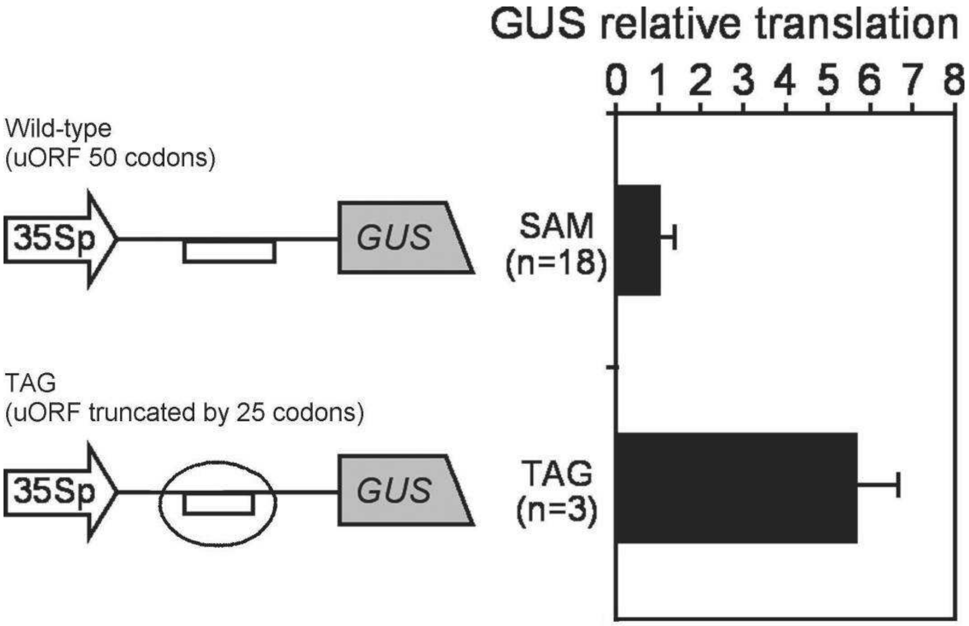


Figure 1.3 Translational efficiency of GUS in leaves of transgenic tobacco plants. 35Sp – 35S promoter, SAM – wild-type construct, and TAG – mutant construct. GUS translational efficiency was calculated as the GUS activity divided by the GUS mRNA level for each transformant (Wang and Wessler 2001).

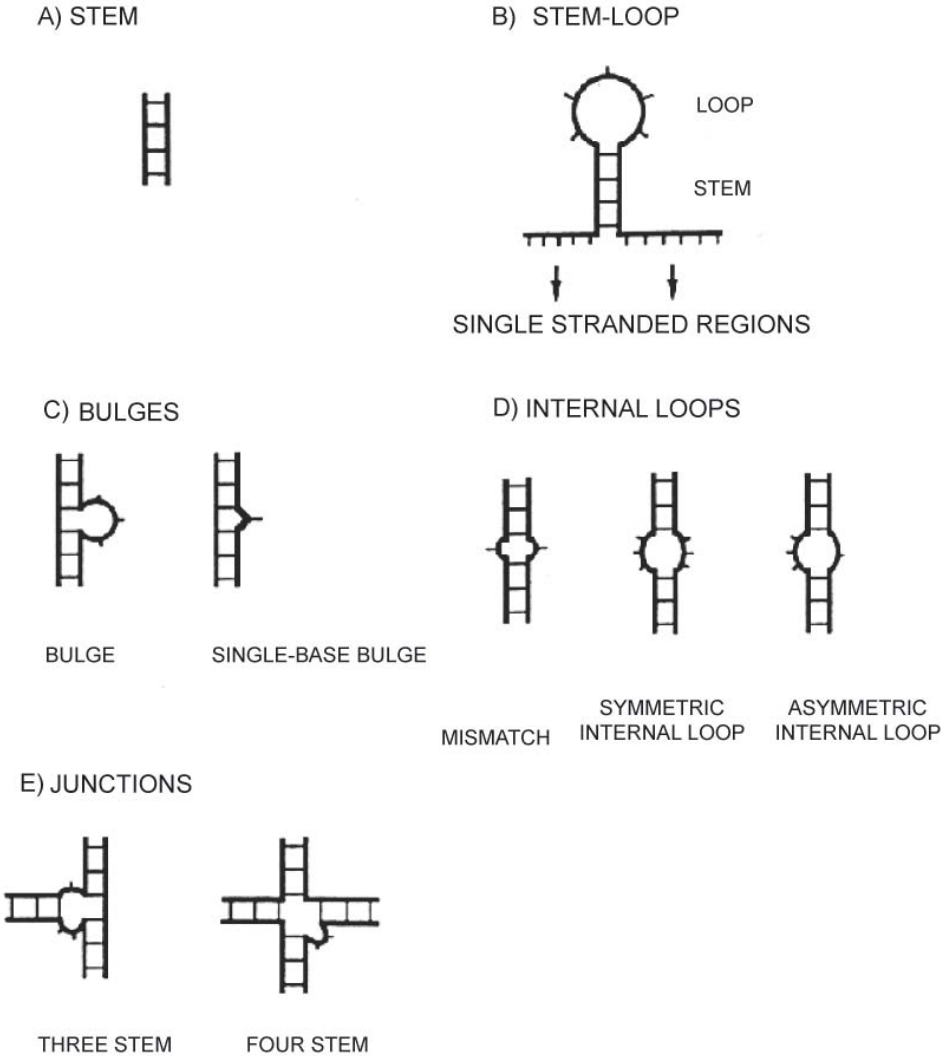


Figure 1.4 Common RNA secondary structure motifs. Image adapted from Chastain and Tinoco (1991).

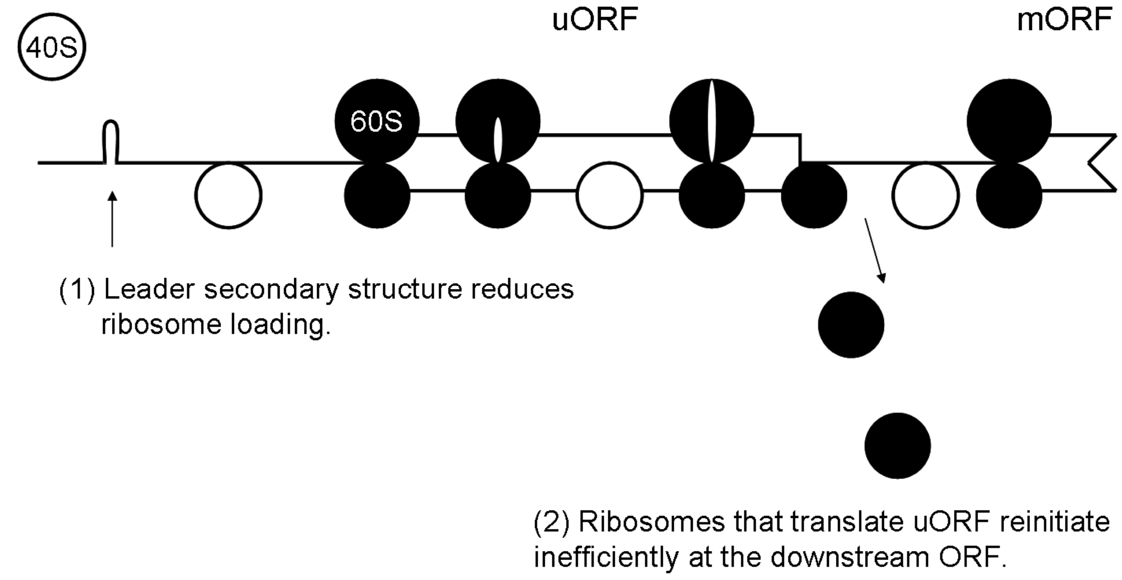
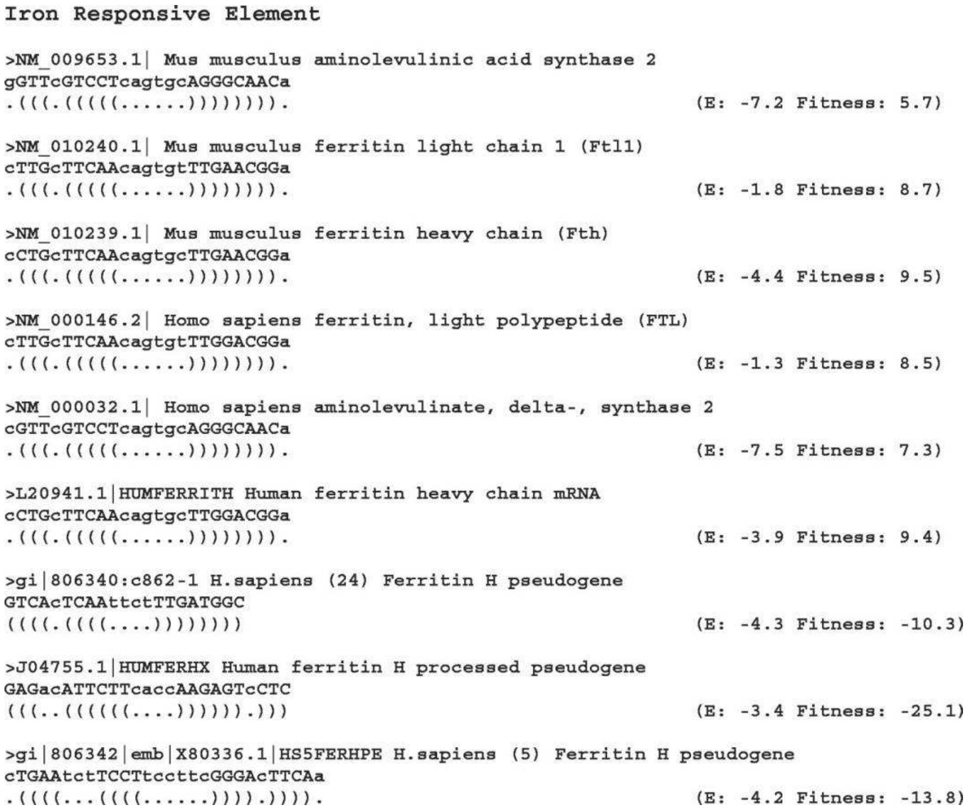

40S
uORF
mORF
60S
(1) Leader secondary structure reduces ribosome loading.
(2) Ribosomes that translate uORF reinitiate inefficiently at the downstream ORF.

Figure 1.5 Two levels of translation repression mediated by the maize *Lc* (1) leader secondary structure and (2) uORF. White and black circles represent 40S ribosomal subunits in scanner and initiation modes, respectively. Image adapted from Wang and Wessler (2001)

Figure 1.6 RNA states and folding energy profile (Chen and Dill 2000).



```
Iron Responsive Element

>NM_009653.1| Mus musculus aminolevulinic acid synthase 2
gGTTcGTCCTcagtgcAGGGCAACa
.(((.(((((......)))))))).                         (E: -7.2 Fitness: 5.7)

>NM_010240.1| Mus musculus ferritin light chain 1 (Ftl1)
cTTGcTTCAAcagtgtTTGAACGGa
.(((.(((((......)))))))).                         (E: -1.8 Fitness: 8.7)

>NM_010239.1| Mus musculus ferritin heavy chain (Fth)
cCTGcTTCAAcagtgcTTGAACGGa
.(((.(((((......)))))))).                         (E: -4.4 Fitness: 9.5)

>NM_000146.2| Homo sapiens ferritin, light polypeptide (FTL)
cTTGcTTCAAcagtgtTTGGACGGa
.(((.(((((......)))))))).                         (E: -1.3 Fitness: 8.5)

>NM_000032.1| Homo sapiens aminolevulinate, delta-, synthase 2
cGTTcGTCCTcagtgcAGGGCAACa
.(((.(((((......)))))))).                         (E: -7.5 Fitness: 7.3)

>L20941.1|HUMFERRITH Human ferritin heavy chain mRNA
cCTGcTTCAAcagtgcTTGGACGGa
.(((.(((((......)))))))).                         (E: -3.9 Fitness: 9.4)

>gi|806340:c862-1 H.sapiens (24) Ferritin H pseudogene
GTCAcTCAAttctTTGATGGC
((((.((((....))))))))                             (E: -4.3 Fitness: -10.3)

>J04755.1|HUMFERHX Human ferritin H processed pseudogene
GAGacATTCTTcaccAAGAGTcCTC
(((..((((((....)))))).)))                         (E: -3.4 Fitness: -25.1)

>gi|806342|emb|X80336.1|HS5FERHPE H.sapiens (5) Ferritin H pseudogene
cTGAAtctTCCTtccttcGGGAcTTCAa
.((((...((((......)))).)))).                      (E: -4.2 Fitness: -13.8)
```

Figure 1.7 Highest scoring motif occurrences output by RNAProfile on the IRE dataset with their respective energy and fitness value. Default parameters were used by RNAProfile. The parentheses and the dots indicate paring and nonpairing respectively (Pavesi et al. 2004).