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 2

RNA SECONDARY STRUCTURES

CHAPTER 2 RNA SECONDARY STRUCTURES

2.1 INTRODUCTION

Motifs that are conserved mostly in structure and less so in sequence can be found in messenger RNAs (mRNAs). The untranslated regions (UTRs) of the messenger RNA are more likely to contain structural motifs that have a biological role as they are usually not exposed to the high unravelling potential of the mature ribosome, which is formed when both the 40S and 60S ribosomal subunits bind at the site of translation (Niepel et al. 1999). Therefore, structural motifs in the UTRs can exhibit their effect on both transcriptional and post-transcriptional regulation of gene expression. Such effects include transcription termination, translation efficiency, mRNA localisation, stability, and alternative splicing (Ji et al. 2004).

The importance of secondary structures in UTRs of plant mRNAs is underscored in several reports. Firstly, Wang and Wessler (2001) showed that the maize transcriptional activator transcript, *Lc*, is translationally repressed by a 5′-UTR stem-loop via a proposed mechanism of reducing ribosomal loading. Also, Hulzink et al. (2002) 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. Finally, stem-loops in transcripts from plastid genes can also regulate chloroplast translation in spite of differences in their translational apparatus compared to cytosolic genes (Marin-Navarro et al. 2007). For example, 5′-UTR stem-loop of tobacco chloroplast *psbA* mRNA (Zou et al. 2003) and 3′-UTR stem-loop of tobacco chloroplast *petD* mRNA (Monde et al. 2000) tend to affect both mRNA stability and translation efficiency, respectively. More recently, it has been shown that the 5′-UTR stem-loop of *Chlamydomonas reinhardtii* chloroplast *psbD* mRNA (Klinkert et al. 2006) also affects translation efficiency by blocking the translation initiation site (TIS, AUG start codon).

A recent survey 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 from unaligned sequences, and most of these algorithms have constraints on the input data (e.g., input data size and sequence length limitations). One algorithm called RNAProfile has limited constraints on the input data (Pavesi et al. 2004), and has become an essential tool for structural biologist for finding conserved motifs as it takes into account both sequence and structural information (Pavesi et al. 2006). Despite the number of algorithms available, there has not been a large scale comparative analysis and discovery of conserved stem-loop motifs in plant transcriptomes.

The purpose of this work is to identify conserved stem-loop motifs in 5′-UTRs of mRNAs from cereal plants that could potentially regulate gene expression using a large collection of rice full-length cDNAs and other cereal ESTs (wheat, barley, and maize) for comparative analysis. To begin, RNAProfile was evaluated using two test sets: archaea 16S rRNAs (Gorodkin et al. 2001) and cereal *thioredoxin4* (*Trx4*) mRNAs (Baumann and Juttner 2002). Studies have suggested that *thioredoxin4* (*Trx4*) is post-transcriptionally regulated based on conserved long 5′-UTRs (Baumann and Juttner 2002; Juttner et al. 2000) and results obtained from transgenic plants (Fleur Dolman, personal communication). Next, RNAProfile was used to analyse long 5′-UTRs (200 to 1200 nt) of putative orthologue sequences from four cereals to find conserved stem-loops of length 20 to 40 nt. The secondary structure predictions will be useful for determining the prevalence of stem-loops in plant transcriptomes, and identifying classes of genes that are being regulated.

2.2 MATERIALS AND METHODS

2.2.1 Sequence data

KOME (Knowledge-based Oryza Molecular biological Encyclopedia) full-length rice cDNA sequences were obtained from ftp://cdna01.dna.affrc.go.jp/pub/data/CURRENT/INE_FULL_SEQUENCE_DB.zip. This file is dated Friday, 24 October 2003, and contains 32,127 full-length cDNA clones (originally 28,469). The Dana Farber Cancer Institute (DFCI) plant gene indices database (http://compbio.dfci.harvard.edu/tgi/) was used to obtain tentative contigs (TCs) from wheat (release 9.0), barley (release 8.0), and maize (release 16.0). Data cleaning was performed on the DFCI datasets to select for sequences that are designated as tentative contigs (identifiers prefixed with “TC”), thereby excluding all singletons. All data files were imported and managed using Microsoft Access 2003.

2.2.2 Orthologue searches

The reciprocal best hit method (rbh) was adapted to account for alternative splice forms that are present in the KOME dataset that would otherwise give many false negatives. The problem with alternative splice forms is that they will never have the highest score in the reverse BLAST because the presence of a longer alternative splice form will always be listed higher on the hit list due to the way BLAST (Altschul et al. 1990) ranks hits (according to score and e-value). To account for alternative splice forms, the top hit and also similar hits in the reverse BLAST (percent identity to top hit: Δ -5%, similar length to top hit: +/- 20%) were examined for symmetry with the top hit in the forward BLAST. If there is symmetry between the forward and reverse BLASTs then we considered the reciprocal pair to be orthologous. General parameters for similarity searches were: tblastx program, expect threshold value at 1.0e-50, scoring parameters, BLOSUM62 matrix; gap costs (existence, 11; extension 1), and filter and masking, off. Only sequence alignments with at least 70% sequence coverage were considered further (applied in Section 2.3.7). Similarity searches were performed at The South Australian Partnership for Advanced Computing (SAPAC) (http://www.sapac.edu.au/).

2.2.3 Algorithm for finding conserved methionine positions

A Java program, CMPSCAN, was written by Michael Tran to find conserved start methionine positions in closely related species as indicators of potential translation initiation sites (Appendix 1.3). Step 1 of CMPSCAN is to annotate the white spaces in the annotation line of CLUSTALW protein alignment outputs to include information on INDELs and mismatches (referred to as anomalies). The specific INDEL or mismatch that was most frequently observed in the alignment was referred to as the highest scoring anomaly. Awareness of the highest scoring anomaly can help identify those sequences that are problematic in aligning translation start methionines in all orthologous. In step 2, the conserved start methionines are detected by scanning the annotation line (N-terminal to C-terminal), from a starting position that is 70 aa (to allow for signal peptide) upstream of the estimated start of the coding region (see next paragraph), until the first occurrence of an alignment of at least 2 (out of 4) methionines. In the final step (step 3), the program reports the conserved start methionine positions, as well as the highest scoring anomaly (e.g., a rice deletion), and the presence of other upstream start and stop codons that can be used to access the reliability of the translation initiation site prediction.

Estimating the start of the conserved coding region (in step 2) was achieved by finding a sub-alignment (60 aa long) that has the highest count of the conservation symbols (*, :, and .) in the annotation line. The program then scans backwards (C-terminal to N-terminal) until there is a breakdown in the conservation symbols, indicating the boundary between the 5′-UTR (usually not conserved) and the start of the coding region has been reached.

2.2.4 RNAProfile algorithm

RNA secondary structure prediction was performed using the RNAProfile algorithm (Pavesi et al. 2004), and is available online as supplementary material. A script was written to enable RNAProfile to run in “batch mode”, so that more than one set of related sequences can be analysed in a single session. RNAProfile was run with the following parameter settings unless otherwise specified: region length (-l and –L) set to 20 and 40 nt, respectively, number of iterations (-I) set to 12, and random picking order.

2.3 RESULTS

2.3.1 Selection of a suitable RNA motif prediction program

A suitable RNA motif prediction program was required for finding conserved local secondary structures in a moderate set of related plant 5′-UTR sequences. In Chapter 1, programs that predict common structures from unaligned sequences were reviewed and the results were summarised in Table 1.1. Based on this comparison, it was clear that RNAProfile was the algorithm of choice for further evaluation for the following reasons. Firstly, the linear time complexity of RNAProfile, which is unseen in other RNA motif prediction algorithms of its type, means it can handle many long sequences. RNA motif prediction based on comparative approaches performs better with more sequences as positions that covary are more pronounced. Also, allowing long sequences improves the detection of secondary structures as longer sequences have a greater potential for secondary structure formation. Algorithms like CARNAC and Dynalign can handle long sequences but are limited to a comparative analysis of only a pair of sequences, since the algorithms have a higher time complexity. GPRM on the other hand requires a minimum of 10 sequences for a reliable prediction making it unsuitable for a comparative analysis of smaller sets of sequences. Secondly, RNAProfile is suitable for the discovery of unknown motifs as it does not require any prior information about the structure of the motifs to be predicted (i.e., constraints on the size, number and position are not required). Some RNA motif prediction algorithms, like GPRM, require the user to specify the maximum number and length of stems and loops comprising the motif. Finally, RNAProfile has been successfully tested on mRNA sequences containing local secondary structures (human and mouse IREs, mRNAs encoding selenoproteins, and *Drosophila* nanos mRNA 3′-UTRs), as it uses local rather global alignments (Dynalign, CARNAC, and comRNA make use of global alignments).

2.3.2 Creating a 16S rRNA dataset to test RNAProfile

To assess the performance of RNAProfile, a new test dataset was created that differed from the datasets used in the original paper (Pavesi et al. 2004). The initial dataset consisted of 34 “archaea” 16S rRNA from Gorodkin et al. (2001). Each 16S rRNA sequence contains a secondary structure as determined by comparative sequence analysis (Wuyts et al. 2004). According to the European ribosomal RNA database (Wuyts et al. 2004), seven sequences (X92172, D85507, D85519, D26489, D85520, L07510, X92171) from the initial dataset were not archaea but bacterial in origin, and another three sequences (U51469, U17593, L77117 U67459-U67608) did not have reliable structural alignments. Therefore, these ten sequences were removed from the dataset. The remaining 24 sequences were used in a pilot test to determine whether RNAProfile can: 1) predict the same rRNA structural motifs as annotated in the European rRNA database, and 2) identify a conserved structural motif in a set of rRNA sequences.

The criterion for selecting conserved rRNA motifs reported by RNAProfile is based on the motif fitness score, which gives an indication of how well the motif fits the consensus structure (known as the profile). It has been demonstrated that a positive fitness score means that the reported motif fits well in the profile, whereas a negative fitness score indicates that the reported motif does not fit the consensus structure (Pavesi et al. 2004). Using this criterion, 23 (out of 24) of the RNAProfile predicted structures agreed with the predicted motif in the European rRNA database (Table 2.1), of which 14 were conserved in the set of rRNA sequences. The conserved structures are predicted to be plain stem-loops, and therefore did not contain internal loops or bulges as seen in the other structures that did not fit the consensus, indicating that RNAProfile can discriminate between motifs based on small structural variations. Therefore, the 14 rRNA sequences (D85038, AJ224936, AF028690, AF028693, AE000940, M36507, X16932, M60880, M38637, Z75218, X05567, M36474, M32222, AJ002946) that shared a highly conserved plain stem-loop motif that agreed with the European rRNA structural annotations formed the test dataset used to further evaluate RNAProfile. Note that it is possible to expand on the test dataset to include less structurally similar rRNA motifs, but this may compromise the robustness of the test dataset.

2.3.3 RNAProfile test case 1: Archaea 16S rRNAs

To evaluate the performance of RNAProfile under different conditions and parameters, RNAProfile was run with various contaminated test datasets to test its ability to discriminate conserved rRNA motifs from spurious motifs. The majority of the test datasets were deliberately made to contain a fixed number (four) of rRNAs and a variable number (one to four) of “contaminating sequences” (sequences that do not contain the expected conserved motif or the absence of a motif), so as to reflect the set of four putative orthologous cereal sequences that are available (Section 2.3.7). In this experiment, sequences containing the iron responsive element (IRE) were chosen as contaminating sequences. The IRE is an ideal motif to use for the feasibility of secondary structure prediction by RNAProfile because it is known to have a conserved RNA motif in animals (Ke et al. 1998; Muckenthaler et al. 1998; Proudhon et al. 1996; Rogers et al. 2002) but appears to be absent in plants (Kozak 2005). Using the IRE motif as a contaminant will provide an indication of specificity in the secondary structure predictions by RNAProfile. Specificity refers to the percentage of contaminating sequences that have a negative fitness score from the total number of contaminating sequences. Similarly, the percentage of rRNA sequences that have positive fitness score (i.e., contain a conserved rRNA motif) from the total number of rRNAs refers to the sensitivity of RNAProfile.

The results in Table 2.2 show that RNAProfile had a low sensitivity (56.1%) in finding the conserved rRNA stem-loop motif with a corresponding high level of false negatives (42.9%) if no contaminating sequences were added. However, including at least one type (a to f) of IRE-contaminating sequences (combinations of 6 IRE, and 1 to 4 IRE) dramatically improved both the sensitivity (100%) and specificity (100%) of RNAProfile. Increasing the number of contaminating sequences did not change the performance of RNAProfile, provided the proportion of contaminating sequences did not exceed 50%. For example, when the number of contaminating sequences was greater than the number of rRNAs, both the sensitivity and specificity of RNAProfile was reduced to 33.3% and 25%, respectively.

To examine the effect of varying the “region length” for folding (-l and -L options) and the number of iterations (-I option) on RNAProfile secondary structure prediction on four 16s rRNA, the region length was varied from 10 to 40 bp and number of iterations from 1 to 24. The *number of iterations* parameter defines the number of times the program is run consecutively, and each run is initiated with a different random seed. The results in Table 2.3 show that a short region length (10 to 19 bp) was not long enough to identify the conserved rRNA motif (24 bp) found in 14 rRNA. Increasing the region length from 20 to 29 bp allowed RNAProfile to correctly identify the conserved rRNA motif (100% sensitivity) and the IRE contaminant (100% specificity). Changing the region length from 30 to 39 bp showed a decrease in specificity for the conserved rRNA motif (0%) and an increase in sensitivity (100%) for a different and larger conserved motif, which contained the rRNA motif nested within. Broadening the region length from 10 to 40 bp showed that RNAProfile was unable to identify the conserved rRNA motif (0% sensitivity), with an increased sensitivity (75%) for detecting a different motif. Using the default region length (20 to 40 bp), RNAProfile was able to identify the IRE contaminant (100% specificity), and the conserved rRNA motif that was nested within a larger motif (100% sensitivity for a different motif). This result is consistent with the fact that many simple plant and animal secondary structures usually range in size between 20 and 40 bp (Chaudhuri and Chatterjee 2007; Klinkert et al. 2006; Monde et al. 2000; Theil 1993; Wang and Wessler 2001; Zou et al. 2003), and is consistent with why this region length was chosen as the default region length in RNAProfile (Pavesi et al. 2004). Table 2.3 also shows that 12 iterations was the optimal number for identifying the conserved rRNA motif (100% sensitivity) and the IRE contaminant (100% specificity). Using the number of iterations of one (default), three, and six resulted in an increased sensitivity (100%) for a larger motif (conserved rRNA motif nested within). Increasing the number of iterations to 24 resulted in decreased specificity (0%). Based on these results, all following experiments using RNAProfile were run with a region length between 20 to 40 bp (default), 12 iterations, and with one IRE contaminant.

2.3.4 RNAProfile test case 2: Cereal *Trx4* mRNAs

To test further the ability of RNAProfile to detect a conserved secondary structure when given a set of related sequences expected to contain an unknown structural element, the cereal *Trx4* mRNA sequences were used as a test case as their 5′-UTRs has been shown to contain conserved regions (Figure 2.1). It was of interest to determine if RNAProfile could find conserved secondary structures in the 5′-UTR of orthologous cereal genes, specifically those with long 5′-UTRs, as later studies on conserved secondary structures in plant mRNAs may use RNAProfile to predict conserved structures in orthologous cereal genes containing long 5′-UTRs, which might be post-transcriptionally regulated. The results in Figure 2.2A show that 5 (out of 6) motif instances reported by RNAProfile had positive fitness scores. These five motif scores, although positive, were extremely low (less than 1.0) in the majority of the cases, and thus can be reasonably suspected to not correspond to a functional motif. The motif found in *Secale cereale* (Sc*Trx4*) 5′-UTR had a much lower and negative fitness score was clearly not conserved, and structurally different (stem-loop with two unpaired bases) to three of the other motifs (plain stem-loop).

The results in Figure 2.2B show that adding a contaminating sequence (i.e., IRE sequence seq_D15071.1) changed the *Trx4* 5′-UTR motif instances reported by RNAProfile. For example, 5 (out of 7) motif instances reported by RNAProfile had a positive fitness score (up to 3.372) resembling stem-loop motifs with one or two internal loops. These five positive scoring motifs had much higher values than those five reported in Figure 2.2A, indicating that these motifs are more highly conserved. The two other motifs, one belonging to the contaminating sequence, and the other PcTrx4 5′-UTR, both had negative fitness scores, -1.807 and -143.955 respectively. The contaminating sequence was correctly identified by RNAProfile based on its low and negative fitness score. The contaminant motif (IRE) was structurally different (stem-loop with an internal loop and bulge) to PcTrx4 5′-UTR (plain stem-loop), which again was different to the predicted conserved *Trx4* 5′-UTR motif (stem-loop with one or two internal loops). The predicted conserved *Trx4* 5′-UTR motif was confirmed to be located within the conserved 5′-UTR regions (Figure 2.1).

2.3.5 Features of the conserved *Trx4* 5′-UTR motif

Figure 2.1 shows that there are two similar types of the predicted conserved *Trx4* 5′-UTR motif, which will be referred to as Type I and Type II. Type I is a 22 nt stem-loop with an internal loop located near the 5′-end of the 5′-UTR, and is predicted to be conserved in HvTrx4 and TaesTrx4. Type II is a 24 nt stem-loop with two internal loops located closer to the start codon (AUG), and is predicted to be conserved in ScTrx4, HbTrx4, and LpTrx4. The Type II motif was not found in HvTrx4, TaesTrx4, and PcTrx4 due to one or two point mutations (T to C and/or C to T) in the corresponding sequences. The Type II motif is predicted to be more conserved in sequence and structure than the Type I motif, possibly indicating that it is more likely to be functional.

2.3.6 Detection of orthologues for comparative analysis

The purpose of this section was to evaluate the performance of the reciprocal best hit (rbh) method in detecting orthologues (Tatusov et al. 1997; Thompson et al. 1994). This was important as the major aim of this chapter was to use a comparative approach for discovering conserved secondary structures in 5′-UTRs of cereal plants, and therefore orthologous sequences were needed. The principle of rbh is that a pair of sequences are orthologues if they are each others best hit, and therefore paralogues are eliminated. The rbh method was used to detect rice genes that had orthologues in barley by using the full-length rice cDNAs and barley assembled ESTs as test datasets, because barley has a less complicated genome (diploid) than other cereals (e.g., wheat). The number of rbh orthologues that were detected was low (less than 50%) compared to the one-directional BLAST method for finding homologues (data not shown).

To investigate the possibility that factors, other than paralogue sequences, can decrease the detection of rbh orthologues, an analysis was performed comparing plots of the number of hits identified in the forward BLAST against a range of expectation values (E-value 0 to -25) (Figure 2.3A) to a plot of the percentage of rbh orthologues identified over the same range of expectation values (Figure 2.3B). The results show that a majority of homologues (1,632) identified in the forward BLAST had an expect value between 0 and -199. The number of homologues is much lower (mostly between 218 and 311) and relatively consistent in the other E-value intervals. Approximately 20-50% rbh orthologues can be found in each E-value interval. Interestingly, homologues with an expect value of 0 also had the same expect value in the reverse BLAST, but surprisingly only 45% of these were identified as rbh orthologues, suggesting there was redundancy in the rice dataset. The redundant sequences were alternative splice forms, and prevented the top hit in the forward and reverse BLASTs from matching. Alternative splice forms of a gene were distinguished by changes in gene length while still maintaining high sequence identity. For example, alternative splice forms that are longer in length had a better score and e-value, and therefore were listed higher in the hit list. Modifying the rbh method to examine not only the top hit but other similar hits corrected the assignment of orthologues, and resulted in an increase in the number of orthologues identified.

The results in Figure 2.3C show that the average sequence coverage of rbh orthologues (the percentage of the query that aligns over the target sequence) over the range of expect intervals (0 to -25) was generally between 40-56%. This was within the expected range as the query sequence contained translations of the untranslated regions (5′ and 3′-UTR) and the coding region. A higher sequence coverage would be expected if the untranslated regions were removed.

2.3.7 Pipeline for discovering 5′-UTR stem-loops in cereals

The pipeline used rice full-length cDNAs (Kikuchi et al. 2003) and their orthologues in other cereals (wheat, barley, and maize TIGR assembled ESTs) to find conserved 5′-UTR stem-loop motifs (Figure 2.4). In the first step of the pipeline, selective measures were used to focus on rice sequences that had full-length 5′-UTRs that were long (200 to 1200 bp, average 5′-UTR is 259.83 nt), as longer 5′-UTRs have greater potential to form secondary structures (Niepel et al. 1999). Rice sequences with 5′-UTR longer than 1200 bp were excluded as they frequently contained frame-shift and sequencing errors (data not shown). A comparative analysis was used to validate the rice cDNAs against an independent rice dataset (i.e., TIGR rice gene indices), and select for sequences that had a complete 5′-UTR. This was necessary as rice full length sequences can be incomplete (Tran et al. 2008) because of failure of the 5′ capping method (Kikuchi et al. 2003). Rice sequences that were shorter than their corresponding sequence in the TIGR rice gene indices were excluded. However, to account for alternative promoter usage, a 100 bp difference in sequence length was accepted. This validation procedure reduced the number of rice sequences from the initial 32,127 to 12,850, indicating that up to 60% of the KOME rice-sequences may not be full-length. To further enrich for stem-loops, sequences with a long 5′-UTR (200 to 1200 bp) (Figure 2.5) were selected, which now reduced the number of rice sequences to 6,217.

In step 2, rice sequences with complete and long 5′-UTRs were used to find putative orthologues in the three other cereals: wheat, barley, and maize, by using the modified version (Section 2.2.2) of the reciprocal best hit method (rbh) (Tatusov et al. 1997; Thompson et al. 1994). The adaption to the rbh method allows orthologues from alternative splice forms to be kept while at the same time eliminating paralogues. Alternative splice forms of a gene were distinguished by changes in gene length while still maintaining high sequence identity. We found that the modified reciprocal best hit method eliminated 22-50% of paralogue sequences. For example, in the one directional BLAST against the barley assembled EST database 3,370 sequences were identified, however this number was reduced to 1,655 (~51%) sequences when the modified reciprocal best hit method was used (Figure 2.4, Step 2).

In step 3, a novel approach was used to predict the translation start site (TSS) of orthologues, and define the 5′-UTR for finding conserved stem-loop motifs. This approach used CLUSTALW (Thompson et al. 1994) alignments of nucleotide translations of orthologues, in the same frame as the coding regions identified by the modified rbh method, for finding conserved methionine positions representing potential translation initiation sites (Section 2.2.3). Furthermore, to report on the effects of missing, incomplete, or erroneous assembled EST data that can result in suboptimal alignments, the CLUSTALW alignment annotation was extended (Table 2.4, Figure 2.6) to describe insertions and deletions (INDELS), mismatches, and stop and start codons in place of blank spaces (default). The results show that 329 orthologue groups had conserved methionine positions in all four cereals (Figure 2.4, Step 3). This number was slightly lower in 3 out of 4 cereals (335), and significantly lower in 2 out of 4 cereals (42) and those not conserved (14). Table 2.5 and Table 2.6 shows a subset of orthologues that have both conserved methionine positions representing potential translation start sites and later predicted to be conserved 5′-UTR stem-loop motifs (see next paragraph).

The final step (Figure 2.4, Step 4), RNAProfile program (Pavesi et al. 2004) was used to find conserved stem-loop motifs in four cereals. RNAProfile was run based on the following recommended parameters (Section 2.3.2): a region length between 20 to 40 bp, 12 iterations, and with one IRE contaminating sequence (Section 2.3.7). The program reported 56 conserved 5′-UTR secondary structure motifs (Tables 2.7 and 2.8) from 664 orthologue groups that had reliable 5′-UTR definitions (according to Step 3).

2.3.8 RNAProfile identified conserved cereal 5′-UTR secondary structures

To begin the identification of conserved 5′-UTR secondary structure predictions, the 5′-UTR sequences of 664 cereal orthologue groups were used as the input data for RNAProfile. The 5′-UTR sequences of each orthologue group also contained one spurious sequence, an IRE contaminating sequence. As mentioned earlier, the IRE is an ideal motif to use for the feasibility of secondary structure prediction by RNAProfile (Section 2.3.3). In brief, the IRE is well described, commonly used in RNA motif prediction, and will allow for specificity in secondary structure prediction by RNAProfile. Moreover, the apparent absence of IREs in the plant transciptome further adds to the robustness in the predictions, as only predictions with a negative fitness score for the IRE contaminant are considered.

The pipeline identified 38 secondary structure motifs conserved in all four cereals (Table 2.7) from 329 orthologue groups that had an identifiable translation initiation site based on their conserved start methionine position (Table 2.5). A further 18 secondary structure motifs were conserved in three out of four cereals (Table 2.8) from 335 orthologue groups (Table 2.6). Therefore, a total of 56 conserved 5′-UTR secondary structure motifs were identified from 664 orthologue groups that had reliable 5′-UTRs.

All 56 conserved secondary structure motifs predicted by RNAProfile were more conserved in structure (from 85% to 98%) than in sequence (from 50% to 88%) (Tables 2.7 and 2.8). Approximately 12.5% (7/56) of the predicted conserved secondary structure motifs are plain stem-loops, 37.5% (21/56) contained an internal loop, 19.6% (11/56) contained multiple internal loops, 19.6% (11/56) contained an internal bulge, and 10.7% (6/56) contained both an internal loop and a bulge. Therefore, the predominant feature of the conserved stem-loops is the internal loop (67.8%). The average length of the stem-loops is 26 nt. Interestingly, the stem part (2 to 8 nt) of the stem-loops is less variable than the loop itself (4 to 17 nt). However, internal loops and bulges (2 to 7 nt) showed similar length variation to the stem (2 to 8 nt).

Five (out of 56) conserved 5′-UTR secondary structure motifs were identified in chloroplast precursor genes (Tables 2.7 and 2.8). These genes encode aspartate aminotransferase (AK072426), thylakoid luminal protein (AK063551), pyruvate kinase isoenzyme A (AK070512), pheophorbide a oxygenase (AK120554), and RuBisCO large subunit binding protein (AK100602). All but one chloroplast precursor gene were predicted to contain a 5′-UTR stem-loop motif with an internal loop whereas chloroplast precursor gene, AK120554, contains a 5′-UTR plain stem-loop.

The results in Tables 2.7 and 2.8 rank the secondary structure predictions in descending order based on their motif fitness score. Also, each prediction provides information on the average structure and sequence conservation of the identified 5′-UTR motif, and the putative function of the gene containing the motif. The 5′-UTR motif itself is described in string notation where the opening and closing parentheses indicate paired nucleotides and full stops indicate unpaired nucleotides. Other information on the identified gene including the translation start site position, positions of upstream start and stop codons (if available), and the highest scoring anomaly identified in the alignments are shown in Tables 2.5 and 2.6. For example, the Ankyrin-2 gene (AK103103) contains a conserved 5′-UTR motif (91.4% average structure conservation in four cereals) that is listed first in Table 2.7 because its motif had the highest motif fitness score (8.0). The gene containing the motif had a predicted translation start site and an upstream stop codon positioned at 164 and 151 nt respectively downstream from its 5′-end (Table 2.5).

2.3.9 Rice genes with conserved 5′-UTR stem-loops have different functions

Biological functions for the 56 rice genes that are predicted to contain a conserved 5′-UTR stem-loop motif could be inferred from their mORF description, as determined by BLAST homology searches in the SwissProt database (Table 2.9 and 2.10). Approximately 20% (11/56) of these rice genes encode proteins that are involved in DNA or nucleic acid binding, as predicted by the gene ontology (GO) molecular function. In addition, ~11% (6/56) and ~7% (4/56) encode proteins predicted to be involved in protein binding, and zinc ion and ATP binding, respectively. There appears to not be a dominant functional category for genes with 5′-UTR stem-loops, but instead are spread across different functions.

2.4 DISCUSSION

2.4.1 8% of cereal transcripts contain conserved stem-loops in long 5′-UTRs

A total of 56 conserved 5′-UTR secondary structure motifs were identified by the stem-loop search pipeline (Section 2.3.6) from select 664 orthologue groups that had reliable 5′-UTRs based on their conserved methionine positions. Using the 664 orthologue groups as a representative sample of the plant transcriptome, since not every gene had an identifiable orthologue, then these results suggest that approximately 8.4% (56/664) of the cereal transcriptome contain a conserved stem-loop motif in their long 5′-UTRs. This would be an under-estimate based on the conservative approach that was used to find these conserved secondary structures (i.e., the search was limited to the identification of secondary structures in long 5′-UTRs from rbh orthologues).

2.4.2 Conserved cereal 5′-UTR stem-loop motifs may have a regulatory role

It can be reasonably suspected that the predicted conserved 5′-UTR stem-loop motifs may have a regulatory role because they are conserved more in structure than in sequence, they are conserved in up to four cereal species, and are at least weakly stable ($\Delta G < -1$ kcal/mol). Some of the ways that 5′-UTR stem-loop motifs can control translation include the pausing of the 40S ribosomal subunit over weak start codons for improved recognition (Kozak 1987b), block ribosome entry near the 5′-end (Kozak 1989), and attenuate ribosome scanning and reduce translational efficiency (Short and Pfarr 2002). In addition, 5′-UTR stem-loop motifs can regulate translation via binding to proteins. The best example of this can be seen with the iron-responsive protein (IRP), which binds to a defined structure (iron-responsive element, IRE, $\Delta G \sim -3$ kcal/mol) to block entry of the 40S ribosomal subunit and thereby reduce translation of the ferritin mRNA (Muckenthaler et al. 1998).

In this study, approximately 67% of the predicted conserved 5′-UTR stem-loop motifs have one or more internal loops. These stem-loops, although structurally different to the IRE, do contain at least one internal loop as seen in the IRE (canonical form type II). Moreover, most of the mORF associated with these predicted conserved 5′-UTR stem-loop motifs are predicted to encode proteins involved in binding. Therefore it is possible, for at least some of these predicted 5′-UTR stem-loop motifs, that they could regulate translation in a similar manner to the IRE-IRP. It is not expected that all predicted 5′-UTR stem-loop motifs will bind to proteins given the limited reports of RNA-binding proteins in controlling translation, and as such other mechanisms of regulation (mentioned above) are expected. In any case, experimental studies are required to support the predicted claim. For example, site-directed mutagenesis can be used to disrupt stem-loop features and their potential binding partners (Hulzink et al. 2002; Klinkert et al. 2006; Wang and Wessler 2001), RNA-protein interactions can be detected using affinity chromatography (Walker et al. 2008) and terminal transferase-dependent PCR (Chen et al. 2000a; Chen et al. 2008), and the amino acids in RNA binding proteins can be identified using mass spectrometry (Kvaratskhelia and Grice 2008; Urlaub et al. 2008) and nuclear magnetic resonance (Petros and Fesik 1994).

Stem-loop structures can be part of larger secondary structures whereby the conserved part represents the inner core structure and the rest represents the outer shell. For example, two adjacent stem-loop structures representing the core structure of ribonuclease P (RNase P) are conserved in bacteria and throughout eukaryotes (Pavesi et al. 2004). It is unclear whether some of the predicted conserved 5′-UTR motifs have unconserved counterparts since the motifs are identified from closely related putative orthologous sequences. One approach to test for larger secondary structures, although not definitive, would be to sequentially extend the conserved 5′-UTR stem-loop motif and see if it remains stable.

2.4.3 Chloroplast precursor transcripts may be translationally regulated

The stem-loop search pipeline identified conserved 5′-UTR stem-loop motifs in chloroplast precursor transcripts that may be post-translationally imported into the chloroplast (Manuell et al. 2007). It is unknown whether these predicted 5′-UTR motifs in chloroplast precursor transcripts have similar biological roles as those previously described for chloroplast 5′-UTR motifs in transcripts encoded in the chloroplast genome. For genes encoded in the chloroplast genome, it is known that chloroplast gene expression is controlled primarily through the regulation of translation, and requires both chloroplast-specific and nucleus encoded factors (Manuell et al. 2007; Marin-Navarro et al. 2007). In particular, chloroplast translation is regulated mainly at the step of initiation, and often involves secondary structures in the 5′-UTR that cause ribosomal stalling. For example, ribosomal stalling has been reported for tobacco (Zou et al. 2003) and Arabidopsis chloroplast *psbA* (Shen et al. 2001; Zou et al. 2003) and Chlamydomonas chloroplast *psbD* (Klinkert et al. 2006). It is not possible to propose an exact model of translational control for these chloroplast precursor 5′-UTR stem-loops without proper experimental studies, similar to those already described in Section 2.4.2.

2.4.4 Some previously reported plant 5′-UTR stem-loops are not conserved

There are plant 5′-UTR stem-loops previously identified in maize *Lc* (Wang and Wessler 2001) and tobacco *ntp303* (Hulzink et al. 2002) transcripts, but these reports do not mention that these stem-loops are conserved in other plant species. Also, searches in the literature do not report that maize *Lc* and *ntp303* stem-loops are conserved. This study supports this view as no conserved cereal

Lc and *ntp303* 5′-UTR stem-loops were reported. In the case of the maize *Lc* 5′-UTR stem-loop, no clear orthologue could be identified in the rice dataset by the modified rbh method, and as such no comparative analysis could be made. However, the closest rice (AK111704), wheat (TC304458), and barley (TC184361) homologues to maize *Lc*, which could be manually retrieved from the datasets (Section 2.2.1), was not predicted by RNAProfile to contain a conserved 5′-UTR stem-loop. Since the maize *Lc* is translationally regulated in a dual manner by both a 5′-UTR stem-loop and a uORF, which has not been reported before, and appears to be not conserved in orthologous genes, this suggests that the *Lc* 5′-UTR stem-loop/uORF dual mechanism may be a recently evolved translational control mechanism. As for the *ntp303* 5′-UTR stem-loop, the folding length (20 to 40 nt) used by RNAProfile was not long enough to completely fold the large H-I (71 nt) and H-II (46 nt) stem-loops. However, increasing the folding length to 75 nt did not report a conserved 5′-UTR stem-loop (data not shown).

2.4.5 Two types of predicted conserved motifs in 5′-UTR of *Trx4*

RNAProfile was unable to identify a conserved motif in the cereal *Trx4* 5′-UTR in the absence of a contaminating sequence, indicating a lack of structural variation among the sequences that is required to discriminate between highly conserved motifs. Indeed, adding an IRE contaminant among the *Trx4* 5′-UTR sequences allowed RNAProfile to report a conserved stem-loop motif. This was the case for the highly conserved rRNA stem-loops, where adding one or more contaminating sequences helped improve both the sensitivity and specificity of RNAProfile.

Further examinations of the conserved cereal *Trx4* 5′-UTR motifs revealed different structural features. *Secale cereale* (Sc*Trx4*_5′UTR), *Lolium perenne* (Lp_*Trx4*_5′UTR), and *Hordeum bulbosum* (Hbtrx_5′UTR) are predicted to form a stem-loop with two-internal loops (referred to as Type II); and *Triticum aestivum* (Taes_trx_UA69309_4565) and *Hordeum vulgare*

(Hvtrx_5′UTR) is predicted to form a stem-loop with one-internal loop (referred to as Type I). On the other hand, *Phalaris coerulescens* (PC*Trx4*_5′UTR) is predicted to form a plain stem-loop, which is not conserved in the other *Trx4* cereal species according to RNAProfile. It is therefore evident that the *Phalaris coerulescens* 5′-UTR stem-loop shows the greatest structural difference among the *Trx4* cereal species, where it no longer contains any internal loops. This is consistent with the view that changes in secondary structures tends to occur in the stem and internal loop regions, as seen in IREs (Ke et al. 1998) and in miRNAs (Lai et al. 2003).

It is difficult to predict the function of the conserved *Trx4* 5′-UTR motif (Type I and Type II) without proper experimental studies. However, it is known that internal loops and bulges of stem-loops allow for binding to RNA-binding proteins (Ke et al. 1998). Moreover, the best characterised example of RNA-binding proteins in regulating mRNA translation is the iron responsive protein, which binds to the iron responsive element to block entry of the 40S ribosomal subunit, and thus inhibit downstream translation (Proudhon et al. 1996). Therefore, the *Trx4* 5′-UTR Type I motif (stem-loop with an internal loop) and Type II (stem-loop with two internal loops) could function in a similar manner to the iron responsive element-iron responsive protein mechanism.

2.4.6 Conserved methionine positions as indicators of translation initiation

The approach for finding conserved methionine positions as potential translation initiation sites developed from the need to define the 5′-UTR. Although 5′-UTR annotations are available from the KOME full-length rice cDNA database (Kikuchi et al. 2003) and the TIGR plant gene indices databases (Lee et al. 2005), they are less reliable. In the case of KOME, the 5′-UTR annotation is based on the longest open reading frame as the coding region, which is not always correct for short proteins less than 100 amino acids long (Frith et al. 2006). On the other hand, TIGR uses several programs (ESTScan, FrameFinder, and ORFfinder) to predict the main open reading frame, and each program has its own advantages and disadvantages. For example, the ESTScan program uses a hidden Markov model (HMM) to identify coding regions based on oligonucleotide frequencies (Iseli et al. 1999). However, ESTScan does not include a model for TIS (Iseli et al. 1999), and as such coding regions do not always start with an ATG (Nadershahi et al. 2004).

Nadershahi et al. (2004) showed that current computational methods on identifying TISs in EST data have an accuracy between 50 to 75%, which is far from optimal. These computational methods based the TIS predictions on a single sequence, and therefore can fail when sequences are incomplete and/or are of poor quality (Adams et al. 1991). One approach to improve the accuracy of TIS predictions is to use orthologous genes and find methionines in similar positions (1-5 amino acids) (Zhang and Dietrich 2005). This study used a more conservative approach in finding TIS, and is based on conserved methionine positions in orthologous genes from closely related cereal species.

It is difficult to determine the accuracy of TIS predictions based on comparative approaches in finding conserved methionine positions in closely related cereal sequences, as many of these sequences have unknown TISs. However, the results in Figure 2.4 (step 3) show that 92% (664 out of 720) orthologue groups had a conserved methionine position in at least 3 out of 4 cereals, suggesting that methionine positions are highly conserved and could represent potential TISs. Correspondingly, only a small number (8%) of orthologue groups either did not have a TIS or could not be reliably identified, indicating that the putative orthologue group sequences are generally of good quality.

2.4.7 Extending CLUSTALW annotation helps identify problematic sequences

Incomplete and erroneous sequence data can affect the alignment of the nucleotide translations of orthologue groups in Figure 2.4 (step 3): translation start site prediction. To further limit the effects of poor sequences, the CLUSTALW annotation was extended to not only report on the level of sequence conservation but also on the presence or absence of insertions and/or deletions (INDELS), and the extent of particular mismatches. The additional information in the extended annotation helps identify those sequences that are problematic, as well as, allow for their removal so that the remaining sequences can be re-aligned for a more optimal alignment. For example, 14 (out of 720) orthologue groups did not have a conserved methionine position. Also, 42 (out of 720) orthologue groups had conserved methionine positions in only 2 (out of 4) other cereals. Therefore, these “less reliable” orthologue groups contain sequences that are likely to be quite dissimilar to each other, and as such could be investigated with the help of the extended annotation to determine the type of sequence anomalies (e.g., a maize deletion).

The extended CLUSTALW annotation is currently limited for a multiple alignment of four sequences, as the code has been written to report specific anomalies (Table 2.4). It may be possible to generalise some of the anomalies, but this would result in a loss of specific information that may be useful. For example, the program can just report that there are INDELs and mismatches instead of a specific anomaly (e.g., rice insertion).

The limitation of using EST datasets in the stem-loop search pipeline is the variable quality and completeness of the tentative contigs, and as such some genes may not be represented, error prone, and may not be full-length. Therefore, the number of putative orthologue groups that can be detected by the rbh method is reduced, which in turn limits the comparative secondary structure analysis to a smaller set of related sequences. Although some measures were introduced to limit the effects of using EST data (i.e., incorporating a rice FL-cDNA dataset and restricting the rbh method to a “2-way comparison” between rice and one other cereal, its use is still a major limitation, which cannot be avoided at this time.

2.5 CONCLUSION

In summary, work described in this chapter on identifying conserved secondary structure motifs in the 5′-UTR of cereal genes shows that at least 8% (56/664) of the cereal transcriptome is predicted to contain conserved stem-loop motifs. The sequence and structure conservation of stable stem-loop motifs in up to four cereal species suggests that these motifs may have a regulatory function. The genes themselves that contain conserved 5′-UTR stem-loop motifs are spread across different functions. Further experimental work is required to confirm the importance of the identified conserved 5′-UTR stem-loops in gene regulation (e.g., controlling translation), and can include deletion and mutational analyses of 5′-UTR stem-loop motifs. It is also important to determine whether the identified 5′-UTR stem-loop motifs interact with RNA-binding proteins (RBPs), as some genes are regulated coordinately by stem-loops and RBPs (Bailey-Serres et al. 2009). Finally, it would be of interest to determine if some 5′-UTR stem-loop motifs are conserved in more distantly related plant species (e.g., *Arabidopsis thaliana* and *Physcomitrella patens*).

Table 2.1. Comparison of predicted motifs from 24 16S rRNA sequences

Identifier	RNAProfile predicted motif [a]			European rRNA database predicted motif [b]
	Fitness score	Energy value (ΔG)	Sequence and structure	
Conserved motifs [c]: RNAProfile structure prediction agrees with European rRNA structural annotation				
D85038	26.136	-18.100	gaGGCUUCCUCCuucgGGAGGGAGUCga ..((((((((((....))))))))))..	GGCUUCCUCCUUCGGGAGGGAGUC ((((((((((....))))))))))
M36474	11.842	-19.800	gaGGCCCGGCCCcuuGGGUCGGGUCga ..((((((((((...))))))))))..	GGCCCGGCCCCUUGGGUCGGGUC ((((((((((...))))))))))
X05567	8.038	-17.900	gaGGGGGUGGCCcuaGGCCACCUUCga ..((((((((((...))))))))))..	GGGGGUGGCCCUAGGCCACCUUC ((((((((((...))))))))))
AJ002946	0.545	-7.300	gaGGUCGACGcaaCGUCGGUCga ..((((((((...))))))))..	GGUCGACGCAACGUCGGUC ((((((((...))))))))
AF028690	13.235	-14.200	gaGGCCUUAGUcuuagGCUAAGGUCga ..(((((((((.....)))))))))..	GGCCUUAGUCUUAGGCUAAGGUC (((((((((.....)))))))))
AE000940	12.343	-14.900	gaGGCCACAACauucuGUUGUGGUCga ..(((((((((.....)))))))))..	GGCCACAACAUUCUGUUGUGGUC (((((((((.....)))))))))
M32222	11.015	-18.800	gaGGCCAUGGCCucuGGCCAUGGUCga ..((((((((((...))))))))))..	GGCCAUGGCCUCUGGCCAUGGUC ((((((((((...))))))))))
M36507	13.852	-16.000	gaGGCCUUGGCCuuuGGCUAGGGUCga ..((((((((((...))))))))))..	GGCCUUGGCCUUUGGCUAGGGUC ((((((((((...))))))))))
AF028693	28.349	-16.900	gaGGCUGCGGUuguuGCCGCAGUCga ..(((((((((....)))))))))..	GGCUGCGGUUGUUGCCGCAGUC (((((((((....)))))))))
M60880	14.009	-6.300	gaGGAUGUAUCauuGAUAUGUUCga ..(((((((((...)))))))))..	GGAUGUAUCAUUGAUAUGUUC (((((((((...)))))))))
X16932	13.541	-13.300	gaGAGCGCUUUcuuugGAGGCGUUCga ..(((((((((.....)))))))))..	GAGCGCUUUCUUUGGAGGCGUUC (((((((((.....)))))))))
Z75218	12.416	-16.800	gaGGCUCCGUCCucuGGGCGGGGUCga ..((((((((((...))))))))))..	GGCUCCGUCCUCUGGGCGGGGUC ((((((((((...))))))))))
AJ224936	7.557	-10.800	gaGGUUUAAUUCgagaGGGUUAAAUCaa ..((((((((((....))))))))))..	GGUUUAAUUCGAGAGGGUUAAAUC ((((((((((....))))))))))
M38637	10.749	-13.800	gaGGGUCCGUCCucuGGAUGGAUUCga ..((((((((((...))))))))))..	GGUCCGUCCUCUGGAUGGAUUC (((((((((...))))))))))
Not conserved motifs [d]: RNAProfile structure prediction agrees with European rRNA structural annotation				
D85508	-85.704	-23.600	aGGCCUCUugCCCcucgGGGugGGAGGUCg .(((((((..(((....)))..))))))).	GGCCUCUUGCCCCUCGGGGUGGGAGGUC (((((((..(((....)))..)))))))
M21087	-48.374	-21.000	aGGCCCCGuCCCUcgccAGGGCGGGGUCg .(((((((.((((....)))))))))))).	GGCCCCGUCCCUCGCCAGGGCGGGGUC (((((((.((((....)))))))))))
D26490	-2.783	-21.300	gaGGCCCCUugCCuuugGGugGGGGGUCga ..(((((((..((....))..)))))))..	GGCCCCUUGCCUUUGGGUGGGGGGUC (((((((..((....))..)))))))
X14835	-100.182	-21.300	aGGCCUCCugCCgacgaGGugGGAGGUCg .(((((((..((.....))..))))))).	GGCCUCCUGCCGACGAGGUGGGAGGUC (((((((..((.....))..)))))))
X69874	-53.601	-12.100	gaGGGCACGGacuucgugCCGUGUUCga ..((((((((........))))))))..	GGGCACGGACUUCGUGCCGUGUUC ((((((((........))))))))
M35966	-0.868	-16.000	uCGCUUGGGgcaaCCCAGGUGg .((((((((....)))))))).	GGCCCCUCGCUUGGGGCAACCCAGGUG .......((((((((....))))))))
X72495	-0.874	-14.500	gaGGCCCGGuuuaCCGGGUCga ..(((((((....)))))))..	GGCCCGGUUUACCGGGUCGAAUC (((((((....))))))).... .
Z70247	-11.933	-20.700	aGCCCGAUCUCCuucgGGAGGUCGGGUc .(((((((((((....))))))))))).	GCCCGAUCUCCUUCGGGAGGUCGGGU (((((((((((....)))))))))))
Z75240	-0.036	-16.300	uGGUCUCCCuucgGGGAGGCCg .((((((((....)))))))).	GGCCUGGUCUCCCUUCGGGGAGGCCGGGUC .....((((((((....))))))))).....
Not conserved motif [d]: RNAProfile structure prediction does not agree with European rRNA structural annotation				
D13379	-544.820	-3.900	aGUGAGguCCggaugaGGCUUGCc .(((((..((......))))))).	GGCUUGCCACGCACGUCGAAUC (((((((...))))))).... .

[a] RNAProfile settings: 3 iterations, random picking order, and template region (20 to 40 nt, default setting).

[b] Secondary structure determined by comparative sequence analysis that aligns sequences and looks for compensating mutations. Query ssu rRNA database to access alignment: http://bioinformatics.psb.ugent.be/webtools/rRNA/ssu/query/index.html

[c] A conserved motif in the set of rRNA sequences is indicated by a positive fitness score.

[d] A motif that is not conserved in the set of rRNA sequences is indicated by a negative fitness score.

Table 2.2. Effects of adding contaminating sequences

Dataset [a]	Sensitivity (%) [b]	Specificity (%) [c]	False Negative (%) [d]	False Positive (%) [e]
14 16S rRNA	57.1	na	42.9	na
14 16S rRNA + 6 IRE	100.0	100.0	0.0	0.0
4 16S rRNA	75.0	na	25.0	na
4 16S rRNA + 1 IRE (a)	100.0	100.0	0.0	0.0
4 16S rRNA + 1 IRE (b)	100.0	100.0	0.0	0.0
4 16S rRNA + 1 IRE (c)	100.0	100.0	0.0	0.0
4 16S rRNA + 1 IRE (d)	100.0	100.0	0.0	0.0
4 16S rRNA + 1 IRE (e)	100.0	100.0	0.0	0.0
4 16S rRNA + 1 IRE (f)	100.0	100.0	0.0	0.0
4 16S rRNA + 2 IRE (a,b)	100.0	100.0	0.0	0.0
4 16S rRNA + 2 IRE (a,c)	100.0	100.0	0.0	0.0
4 16S rRNA + 2 IRE (a,d)	100.0	100.0	0.0	0.0
4 16S rRNA + 2 IRE (a,e)	100.0	100.0	0.0	0.0
4 16S rRNA + 2 IRE (a,f)	100.0	100.0	0.0	0.0
4 16S rRNA + 3 IRE (a,b,c)	100.0	100.0	0.0	0.0
4 16S rRNA + 3 IRE (a,c,d)	100.0	100.0	0.0	0.0
4 16S rRNA + 3 IRE (a,d,e)	100.0	100.0	0.0	0.0
4 16S rRNA + 3 IRE (a,e,f)	100.0	100.0	0.0	0.0
4 16S rRNA + 4 IRE (a,c,d,f)	100.0	75.0	0.0	25.0
4 16S rRNA + 4 IRE (a,c,d,e)	100.0	100.0	0.0	0.0
3 16S rRNA + 4 IRE (a,b,e,f)	33.3	25	66.7	75

[a] 14 rRNAs: D85038, M36474, X05567, AJ002946, AF028690, AE000940, M32222, M36507, AF028693, M60880, X16932, Z75218, AJ224936, M38637.
4 rRNAs: M36474, AJ002946, AE000940, M60880.
3 rRNAs: M36474, AJ002946, AE000940.
IRE: a - seq_L37082.1, b - seq_D28463.1, c - seq_Y15629.1, d - seq_D15071.1, e - seq_M60170.1, f - seq_AJ251148.1
RNAProfile settings: 3 iterations, random picking order, template region (20 to 29 nt).

[b] Sensitivity = Percentage of rRNAs that have a positive fitness score from the total number of rRNAs.

[c] Specificity = Percentage of IREs/contaminating sequences that have a negative fitness score from the total number of IREs.

[d] False negative = 1 – sensitivity.

[e] False positive = 1 – specificity.

na – not applicable because no contaminating sequence in dataset.

Table 2.3. Effect of varying the region length and the number of iterations

Region length (bp)	4 16S rRNA + 1 IRE		
	Sensitivity[a] (%)	Specificity[b] (%)	Sensitivity for different conserved motifs (%)
10 – 19	0	0	50
20 – 29	100	100	0
30 – 39	0	0	100[c]
10 - 40	0	0	75
20 – 40[d]	0	100	100[c]

No. of iterations	4 16S rRNA + 1 IRE		
	Sensitivity[a] (%)	Specificity[b] (%)	Sensitivity for different conserved motifs (%)
1[d]	0	0	100
3	0	100	100[c]
6	0	0	100[c]
12	100	100	0
24	100	0	0

RNAProfile settings: 3 iterations, random picking order, template region (20 to 40 nt, default setting)
rRNAs: M36474, AJ002946, AE000940, M60880
IRE: a - seq_L37082.1

[a] Sensitivity = Percentage of rRNAs that have a positive fitness score from the total number of rRNAs.

[b] Specificity = Percentage of IREs/contaminating sequences that have a negative fitness score from the total number of IREs.

[c] rRNA motif nested within

[d] Default setting of RNAProfile

Table 2.4. Extended CLUSTALW annotation code

Possible combinations	Code	Interpretation
Match type 1	*	Fully conserved amino acid
Match type 2	:	Strongly conserved amino acid
Match type 3	.	Weakly conserved amino acid
Mismatch	blank space	Non conserved amino acid
SAME ‘M’{wheat, barley,maize, rice}	@	All Methionines (Met) aligned
SAME ‘M’{wheat, barley,maize}	#	Rice Met not aligned
SAME ‘M’{barley,maize,rice}	&	Wheat Met not aligned
SAME ‘M’{maize, rice, wheat}	$	Barley Met not aligned
SAME ‘M’{rice, wheat, barley}	%	Maize Met not aligned
SAME ‘M’{wheat, barley}	[	Maize & rice Met not aligned
SAME ‘M’{barley, rice}	]	Maize & wheat Met not aligned
SAME ‘M’{maize, rice}	(	Wheat & barley Met not aligned
SAME ‘M’{barley, maize}	)	Wheat & rice Met not aligned
SAME ‘M’{rice, wheat}	<	Barley & maize Met not aligned
SAME ‘M’{wheat, maize}	>	Barley & rice Met not aligned
SAME ‘-‘ {wheat, barley, maize}	r	Rice insertion
SAME ‘-‘ {barley, maize, rice}	w	Wheat insertion
SAME ‘-‘ {maize, rice, wheat}	b	Barley insertion
SAME ‘-‘ {rice, wheat, barley}	m	Maize insertion
{wheat, barley, maize, -}	R	Rice deletion
{-, barley, maize, rice}	W	Wheat deletion
{wheat, -, maize, rice}	B	Barley deletion
{wheat, barley, -, rice}	M	Maize deletion
Pair of ‘-‘	e	Either insertion or deletion
SAME CHAR{wheat, barley, maize}	0	Rice mismatch
SAME CHAR{barley, maize, rice}	1	Wheat mismatch
SAME CHAR{maize, rice, wheat}	2	Barley mismatch
SAME CHAR{rice, wheat, barley}	3	Maize mismatch
SAME CHAR{wheat, barley}	4	Maize & rice mismatch
SAME CHAR{barley, maize}	5	Wheat & rice mismatch
SAME CHAR{maize, rice}	6	Wheat & barley mismatch
SAME CHAR{rice, wheat}	7	Barley & maize mismatch
SAME CHAR{wheat, maize}	8	Barley & rice mismatch
SAME CHAR{barley, rice}	9	Wheat & maize mismatch
SAME ‘X’ {wheat, barley, maize, rice}	a	Common stop codon
SAME ‘X’{wheat, barley,maize}	b	Wheat, barley, maize stop codon
SAME ‘X’{barley,maize,rice}	c	Barley, maize, rice stop codon
SAME ‘X’{maize, rice, wheat}	d	Maize, rice, wheat stop codon
SAME ‘X’{rice, wheat, barley}	e	Rice, wheat, barley stop codon
SAME ‘X’{wheat, barley}	f	Wheat and barley stop codon
SAME ‘X’{barley, rice}	g	Barley and rice stop codon
SAME ‘X’{maize, rice}	h	Maize and rice stop codon
SAME ‘X’{barley, maize}	i	Barley and maize stop codon
SAME ‘X’{rice, wheat}	j	Rice and wheat stop codon
SAME ‘X’{wheat, maize}	k	Wheat and maize stop codon
SAME ‘X’{wheat}	l	Wheat stop codon
SAME ‘X’{barley}	m	Barley stop codon
SAME ‘X’{maize}	n	Maize stop codon
SAME ‘X’{rice}	o	Rice stop codon

Table 2.5. Rice translation initiation site predictions for rbh orthologue groups (4/4)

Identifier	PCHUS	HSA	TSS (aa)	Upstream start codon position (aa)	Upstream stop codon position (aa)
AK103103	4/4	maize and rice mismatch	164	none	151
AK103173	1/4	rice insertion	239	197,160,146,143,116	113
AK106377	2/4	maize insertion	205	134	130
AK062304	3/4	maize and rice mismatch	268	none	0
AK102399	4/4	barley deletion	184	147	126
AK102057	3/4	maize and rice mismatch	86	none	64
AK069837	2/4	maize mismatch	201	108,81	75
AK065483	1/4	barley deletion	150	95,85	24
AK071857	4/4	maize and rice mismatch	110	105, 24	103
AK072243	1/4	maize mismatch	269	165,153,110,103,72,19	0
AK102267	4/4	wheat deletion	258	139,95,69	94
AK100074	2/4	barley deletion	208	123	108
AK071306	3/4	barley insertion	212	107,6	50
AK072426	3/4	rice mismatch	299	185,182,163,136	262
AK067368	1/4	rice insertion	199	82	55
AK065504	0/4	rice insertion	185	143,96	0
AK068088	3/4	barley deletion	211	111,107,93,85	78
AK063551	1/4	rice insertion	292	201,181,174,153	179
AK067644	2/4	rice mismatch	268	none	181
AK073830	2/4	rice insertion	336	263,228,203,201,107,100	46
AK111816	2/4	maize insertion	72	none	0
AK067177	2/4	rice insertion	288	170,81	64
AK104877	2/4	maize insertion	121	none	52
AK101609	0/4	maize insertion	95	55	0
AK104437	4/4	maize insertion	146	145	109
AK072017	3/4	barley mismatch	323	276,261,165,112	105
AK065629	4/4	wheat deletion	270	81	12
AK102987	4/4	wheat deletion	150	133	75
AK065258	2/4	wheat deletion	148	108,72	0
AK070512	1/4	maize and rice mismatch	108	none	0
AK111060	3/4	maize insertion	113	none	59
AK099454	4/4	maize and rice mismatch	77	none	64
AK120554	1/4	barley deletion	303	96,93	0
AK072267	2/4	maize and rice mismatch	191	174,77,75	68
AK104755	4/4	maize and rice mismatch	90	none	64
AK121775	1/4	barley insertion	113	82	44
AK061748	4/4	maize insertion	99	none	46
AK065035	4/4	barley insertion	197	178,104,96,3	81

PCHUS – Proportion of cereals having upstream stop codons
HSA – Highest scoring anomaly (Section 2.2.3)
TSS – Translation start site of rice cDNA based on conserved methionine positions (Section 2.2.3)

Table 2.6. Rice translation initiation site predictions for rbh orthologue groups (3/4)

Identifier	PCHUS	CLCMP	HSA	TSS (aa)	Upstream start codon positions (aa)	Upstream stop codon position (aa)
AK069040	3/4	barley	maize and rice mismatch	211	169,68	12
AK102349	1/4	maize	wheat insertion	79	none	12
AK101530	2/4	rice	maize and rice mismatch	113	none	106
AK119232	4/4	rice	rice deletion	157	none	149
AK070308	2/4	barley	barley mismatch	102	none	0
AK062845	1/4	maize	maize deletion	85	none	0
AK069526	4/4	rice	wheat deletion	246	147,50	191
AK111883	2/4	maize	rice insertion	334	305,277,117,101,83,80	120
AK058418	2/4	maize	maize mismatch	331	215,200,123,103,93,47,35	182
AK072560	2/4	rice	wheat mismatch	207	187,167,162,147,114,98	0
AK067725	2/4	wheat	maize deletion	195	156,99	105
AK111777	3/4	barley	barley mismatch	234	218,71	51
AK100602	2/4	rice	rice mismatch	127	116,106,56	114
AK059712	0/4	rice	rice mismatch	247	none	0
AK073957	1/4	maize	maize deletion	253	199,110,95	80
AK073091	3/4	maize	barley deletion	214	134	92
AK063364	2/4	maize	maize deletion	81	none	5
AK068213	3/4	maize	maize and rice mismatch	251	179,85	78

PCHUS - Proportion of cereals having upstream stop codons.
CLCMP - Cereal lacking conserved methionine position.
HSA - Highest scoring anomaly (Section 2.2.3).
TSS - Translation start site for rice cDNA based on conserved methionine positions (Section 2.2.3).

Table 2.7. Conserved secondary structures predicted by RNAProfile in 4/4 cereals

Identifier	Average conservation (%)		Motif fitness	Motif	BLAST searches			
	Structure [a]	Sequence			Expect	Percent identity (%)	Alignment length	Putative function [b]
AK103103	91.4	78.2	8.0	(((((((.................))).)))) [c]	3.0e-12	43	42/97	Ankyrin-2
AK103173	94.8	83.4	7.5	(((.((.((.((((......)))).)).)).))) [d]	1.0e-08	28	43/151	Putative methyltransferase NSUN5
AK106377	93.5	67.7	7.3	((((..(((((....))).))..)))) [e]	-	-	-	no hits found
AK062304	94.2	82.8	6.0	.((.(((((((.((.......)).))))))).)). [d]	2.0e-36	31	107/338	Inactive ubiquitin carboxyl-terminal hydrolase 53
AK102399	92.9	80.0	5.8	(((.((.((...........)))).))) [e]	1.0e-155	100	276/276	Probable protein ABIL1
AK102057	85.8	50.2	5.3	(((((((............))))))) [f]	1.0e-11	29	54/185	Melanoma-associated antigen G1
AK069837	94.4	90.5	5.3	(((..(((((((.........)))))..))...))) [e]	7.0e-33	29	110/368	Speckle-type POZ protein-like A
AK065483	95.4	85.1	4.6	.(((((((..(((....)))..))....)).))). [e]	4.0e-63	73	108/146	HVA22-like protein i
AK071857	87.6	57.6	4.3	..(((((.((((((.....)))))).))))).. [g]	2.0e-08	35	28/78	E3 ubiquitin-protein ligase MARCH4 precursor
AK072243	95.1	84.8	4.2	.((((..((...((........)).))....)))). [d]	8.0e-95	47	187/391	6-phosphofructo-2-kinase
AK102267	97.0	84.7	3.8	(((..((((..((.......))..))))...))) [d]	1.0e-107	64	193/299	Probable serine/threonine-protein kinase NAK
AK100074	96.4	85.0	3.8	.((((....(((.....)))...)))). [g]	5.0e-17	37	47/127	Haloacid dehalogenase-like
AK071306	96.7	84.5	3.2	.(((.((((.((.....)).)))).))).[d]	3.0e-08	32	48/149	Dihydrolipoyllysine-residue acetyltransferase
AK072426	96.9	78.4	3.1	.(((((.((......)).))))).[g]	0	87	243/277	Aspartate aminotransferase, chloroplast precursor
AK067368	95.8	83.4	3.1	(((.((((((......))))))..))) [g]	2.0e-85	62	165/264	Pyrroline-5-carboxylate reductase
AK065504	96.1	81.5	3.0	.(((((.((.......))))))). [c]	1.0e-10	36	34/93	Protein SIP5
AK068088	95.3	74.6	3.0	.(((((((......))))))). [f]	e-134	86	237/275	Expansin-like A2 precursor
AK063551	96.4	81.4	2.9	.(((..(((((.....)))))..))). [g]	1.0e-41	69	82/118	Thylakoid lumenal protein 2, chloroplast precursor
AK067644	96.8	75.7	2.9	..((.(((((...........)))))...)).. [g]	1.0e-133	58	161/273	Vacuolar protein sorting-associated protein 4
AK073830	96.1	83.8	2.8	.((((((((.....))).))))). [c]	5.0e-24	31	96/307	PAB1-binding protein 1
AK111816	96.0	70.4	2.7	((((((((....)))))))) [f]	-	-	-	no hits found
AK067177	97.5	78.1	2.7	..((((.((((.....))))....)))).. [g]	0	60	359/589	Vacuolar protein sorting-associated protein 33
AK104877	96.9	86.9	2.5	(((.((((.........)))).))) [g]	0	95	396/416	Enolase
AK101609	92.9	77.8	2.5	..((((((..(((......)))....)))))).. [g]	8.0e-9	32	36/110	R3H domain-containing protein 2
AK104437	97.0	84.4	2.5	.(((((...(((....)))..))))). [g]	-	-	-	no hits found
AK072017	96.9	80.7	2.5	((.(((((.........))))).)) [g]	-	-	-	no hits found

More…

Table 2.7. Conserved secondary structures predicted by RNAProfile in 4/4 cereals (Continued)

Identifier	Average conservation (%)		Motif fitness	Motif	BLAST searches			
	Structure [a]	Sequence			Expect	Percent identity (%)	Alignment length	Putative function [b]
AK065629	95.8	88.2	2.3	.((((((.(((......))).)))).)). [e]	1.0e-82	62	173/278	Uncharacterized protein At1g03900
AK102987	93.3	82.6	1.4	.((((......((((....)))).)))). [g]	1.0e-52	37	125/334	Uncharacterized membrane protein At1g06890
AK065258	96.6	77.5	1.2	.(((((((....)))).))). [c]	1.0e-32	27	94/345	Transcription elongation factor A protein 3
AK070512	97.2	86.2	1.1	.((.(((.((......))....))).)). [d]	0	81	410/501	Pyruvate kinase isozyme A, chloroplast precursor
AK111060	98.3	84.1	1.1	(((((((.........))))))) [f]	2.0e-50	44	97/217	Menaquinone biosynthesis methyltransferase ubiE
AK099454	87.3	60.0	0.3	.(((((((.....))))))). [f]	5.0e-19	57	44/76	U6 snRNA-associated Sm-like protein LSm1
AK120554	91.7	76.5	0.2	..(((((((.......))))))).. [f]	0	70	343/485	Pheophorbide a oxygenase, chloroplast precursor
AK072267	91.3	77.3	0.1	((((((..((.....))..)))))) [g]	9.0e-24	40	86/210	Uncharacterized membrane protein At4g06598
AK104755	91.3	57.3	0.1	.(((.((((.......))))))). [c]	1.0e-18	27	81/297	SEC14-like protein 1
AK121775	96.4	77.1	0.02	.(((..((((.....))))..))). [g]	4.0e-26	23	93/389	Glutelin type-B 1 precursor
AK061748	90.2	57.7	0.01	.((((..(((.........)))..)))). [g]	-	-	-	no hits found
AK065035	90.3	75.8	0.01	((.(((..((....)).))).)) [d]	6.0e-10	30	48/157	StAR-related lipid transfer protein 7

[a] Average structure conservation is based on the percentage of gaps introduced into the alignment of secondary structures.
[b] Functional annotation based on “SwissProt” database.
[c] Stem-loop with an internal bulge.
[d] Stem-loop with multiple internal loops.
[e] Stem-loop with an internal loop and bulge.
[f] Plain stem-loop.
[g] Stem-loop with an internal loop.

Table 2.8. Conserved secondary structures predicted by RNAProfile in 3/4 cereals

Identifier	Average conservation (%)		Motif fitness	Motif	BLAST searches			
	Structure [a]	Sequence			Expect	Percent identity (%)	Alignment length	Putative function [b]
AK069040	90.0	63.1	12.8	..((((((((....)).......)))))).. [c]	1.0e-116	60	205/341	Probable mannitol dehydrogenase
AK102349	92.4	53.1	11.9	(((.((((((.....))))))))) [c]	2.0e-16	27	77/276	Transmembrane protein 115
AK101530	90.2	69.7	8.5	..(((..((.(((.....))).)).))).. [d]	2.0e-74	45	122/266	Tyrosyl-tRNA synthetase
AK119232	92.6	79.5	6.5	.((((((((....((.........))...)).)))))). [e]	0	95	325/340	UTP--glucose-1-phosphate uridylyltransferase
AK070308	97.2	85.0	3.5	.(((.((...(((.......)))..))...))). [d]	-	-	-	no hits found
AK062845	96.2	82.9	3.0	.((((.(((.......))).)))). [f]	-	-	-	no hits found
AK069526	97.0	80.0	2.6	.(((..((((........))))))). [c]	1.0e-146	65	252/385	Cyclin-dependent kinase F-4
AK111883	95.3	70.7	2.4	.((((.(((.....))).)))). [f]	1.0e-109	35	249/692	WD repeat-containing protein 48
AK058418	91.3	63.7	2.4	..(((....((((...))))....))).. [f]	8.0e-47	56	93/164	Acetolactate synthase small subunit
AK072560	96.5	80.0	2.3	((..(((.(((....))).)))...)) [d]	1.0e-28	38	69/178	Protein quaking-A
AK067725	93.5	74.2	1.6	(((((((........)))).))) [c]	3.0e-47	40	101/249	RNA-binding protein Musashi homolog 2
AK111777	98.2	83.4	1.4	.((((((....)).)))). [c]	1.0e-107	49	191/384	SEC12-like protein 1
AK100602	95.0	79.2	1.2	.((.((((((.....)))))).)). [f]	0	91	378/413	RuBisCO large subunit-binding, chloroplast precursor
AK059712	99.1	88.6	0.9	(((((((........))))))) [g]	0	93	272/290	Adenosylhomocysteinase
AK073957	88.5	61.1	0.6	.((((...((((......))))....)))). [f]	5.0e-63	37	131/349	F-box/kelch-repeat protein SKIP4
AK073091	92.8	78.9	0.1	..((((..(((..(((....))).)))...)))).. [d]	0	70	292/414	Branched-chain-amino-acid aminotransferase-like
AK063364	93.3	58.6	0.05	..((((.((((....)))))))).. [c]	-	-	-	no hits found
AK068213	94.1	76.4	0.0	.(((.((((((.....)))).))))). [f]	3.0e-90	67	181/268	Auxin-responsive protein IAA30

[a] Average structure conservation is based on the percentage of gaps introduced into the alignment of secondary structures.
[b] Functional annotation based on “SwissProt” database.
[c] Stem-loop with an internal bulge.
[d] Stem-loop with multiple internal loops.
[e] Stem-loop with an internal loop and bulge.
[f] Stem-loop with an internal loop.
[g] Plain-stem loop.

Table 2.9. Function of rice clones with predicted conserved (4/4) 5′-UTR stem-loop motifs

Identifier	Clone name	mORF description[a]	Gene ontology molecular function[b]
AK103103	J033118O11	Ankyrin-2	Protein binding
AK103173	J033121G21	Putative methyltransferase NSUN5	Unknown
AK106377	002-102-D05	no hits found	Unknown
AK062304	001-101-A04	Inactive ubiquitin carboxyl-terminal hydrolase 53	DNA binding
AK102399	J033092I23	Probable protein ABIL1	Protein kinase activity
AK102057	J033082D03	Melanoma-associated antigen G1	Unknown
AK069837	J023031F07	Speckle-type POZ protein-like A	Protein binding
AK065483	J013008G03	HVA22-like protein i	Unknown
AK071857	J023118I21	E3 ubiquitin-protein ligase MARCH4 precursor	Protein binding
AK072243	J023003N10	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 1	Fructose-2,6-bisphosphate 2-phosphatase activity
AK102267	J033088N21	Probable serine/threonine-protein kinase NAK	Protein kinase activity
AK100074	J013170E11	Haloacid dehalogenase-like	Phosphoglycolate phosphatase activity
AK071306	J023088C20	Dihydrolipoyllysine-residue acetyltransferase	Hydrolase activity
AK072426	J023091M09	Aspartate aminotransferase, chloroplast precursor	Aspartate transaminase activity
AK067368	J013104L18	Pyrroline-5-carboxylate reductase	Pyrroline-5-carboxylate reductase activity
AK065504	J013027C16	Protein SIP5	Zinc ion binding
AK068088	J013131C11	Expansin-like A2 precursor	Unknown
AK063551	001-117-E02	Thylakoid lumenal protein 2, chloroplast precursor	Unknown
AK067644	J013114F09	Vacuolar protein sorting-associated protein 4	DNA binding
AK073830	J033068A12	PAB1-binding protein 1	Protein binding
AK111816	J013147K05	no hits found	DNA binding
AK067177	J013097L24	Vacuolar protein sorting-associated protein 33	Unknown
AK104877	001-044-E09	Enolase	Phosphopyruvate hydratase activity
AK101609	J033052B11	R3H domain-containing protein 2	DNA binding
AK104437	006-209-A02	no hits found	Unknown
AK072017	J013104N20	no hits found	Unknown
AK065629	J013028C23	Uncharacterized protein At1g03900	Transporter activity
AK102987	J033116C12	Uncharacterized membrane protein At1g06890	Unknown
AK065258	J013002J17	Transcription elongation factor A protein 3	DNA binding, transcription factor
AK070512	J023055B17	Pyruvate kinase isozyme A, chloroplast precursor	Pyruvate kinase activity
AK111060	002-175-D07	Menaquinone biosynthesis methyltransferase ubiE	Methyltransferase activity
AK099454	J013022L10	U6 snRNA-associated Sm-like protein LSm1	Unknown
AK120554	J013130M21	Pheophorbide a oxygenase, chloroplast precursor	Oxidoreductase activity
AK072267	J023004I18	Uncharacterized membrane protein At4g06598	DNA binding, transcription factor
AK104755	001-038-G03	SEC14-like protein 1	Binding
AK121775	J033092N18	Glutelin type-B 1 precursor	Nutrient reservoir activity
AK061748	001-038-G04	no hits found	Unknown
AK065035	J013001H07	StAR-related lipid transfer protein 7	Unknown

[a] Based on “SwissProt” database.

[b] Retrieved from the Knowledge-based Oryza Molecular biological Encyclopedia (KOME) online report pages (http://cdna01.dna.affrc.go.jp/cDNA/).

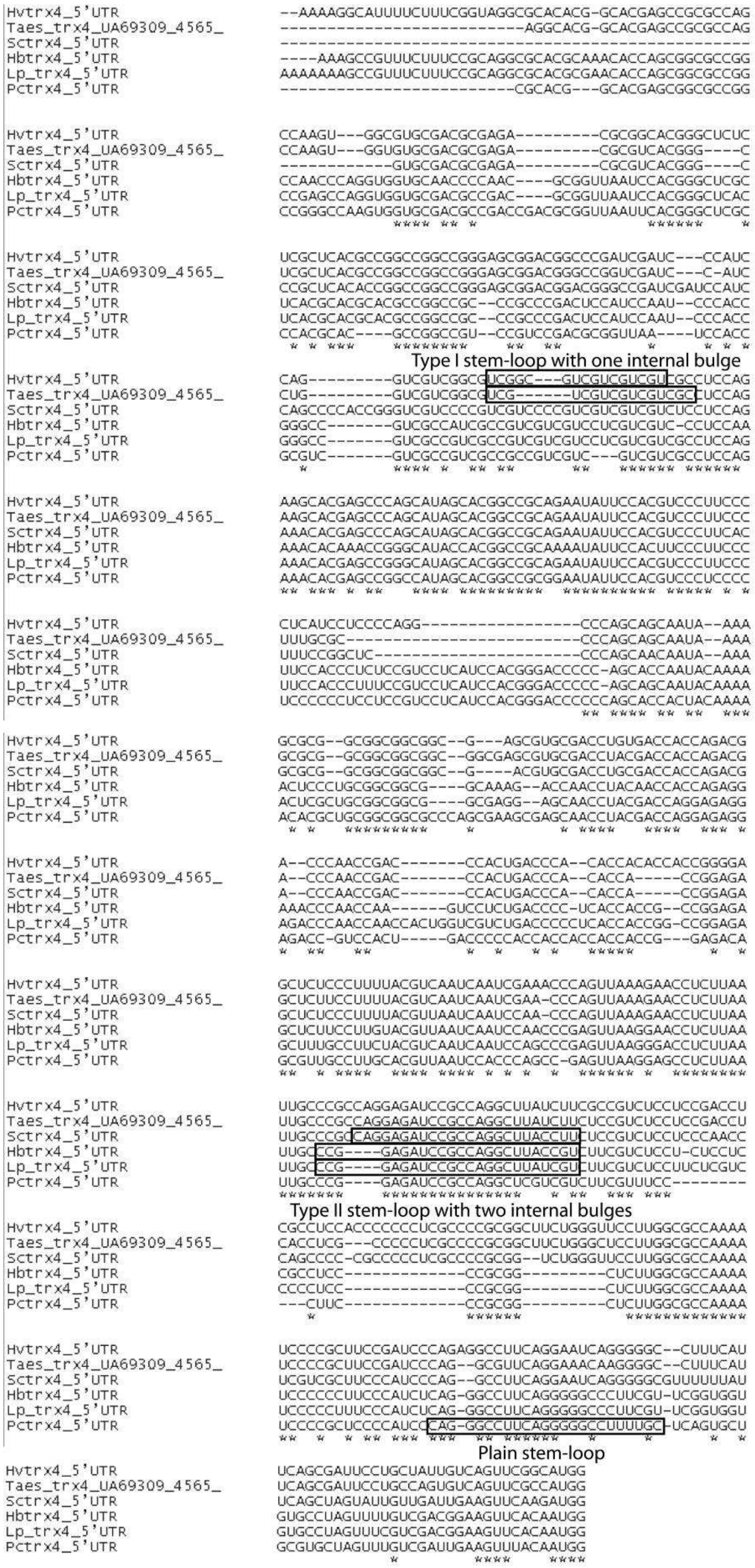
Table 2.10. Function of rice clones with predicted conserved (3/4) 5′-UTR stem-loop motifs

Identifier	Clone name	mORF description[a]	Gene ontology molecular function[b]
AK069040	J023003B20	Probable mannitol dehydrogenase	zinc ion binding
AK102349	J033091C24	Transmembrane protein 115	Unknown
AK101530	J033047I05	Tyrosyl-tRNA synthetase	ATP binding, tRNA ligase activity
AK119232	001-114-E10	UTP--glucose-1-phosphate uridylyltransferase	nucleotidyltransferase activity
AK070308	J023049A10	no hits found	Unknown
AK062845	001-107-H09	no hits found	Unknown
AK069526	J023025B11	Cyclin-dependent kinase F-4	ATP binding, kinase activity
AK111883	J033060M06	WD repeat-containing protein 48	DNA binding
AK058418	001-015-D05	Acetolactate synthase small subunit	Protein binding
AK072560	J023133B04	Protein quaking-A	Nucleic acid binding
AK067725	J013116G21	RNA-binding protein Musashi homolog 2	RNA binding, nucleic acid binding
AK111777	J023077D02	SEC12-like protein 1	DNA binding
AK100602	J023107D12	RuBisCO large subunit-binding, chloroplast precursor	Protein binding
AK059712	001-032-F05	Adenosylhomocysteinase	Adenosylhomocysteinase activity
AK073957	J033076O03	F-box/kelch-repeat protein SKIP4	Unknown
AK073091	J033022E01	Branched-chain-amino-acid aminotransferase-like	catalytic activity
AK063364	001-114-D07	no hits found	Unknown
AK068213	J013149K13	Auxin-responsive protein IAA30	Transcription factor

[a] Based on "SwissProt" database.

[b] Retrieved from the Knowledge-based Oryza Molecular biological Encyclopedia (KOME) online report pages (http://cdna01.dna.affrc.go.jp/cDNA/)

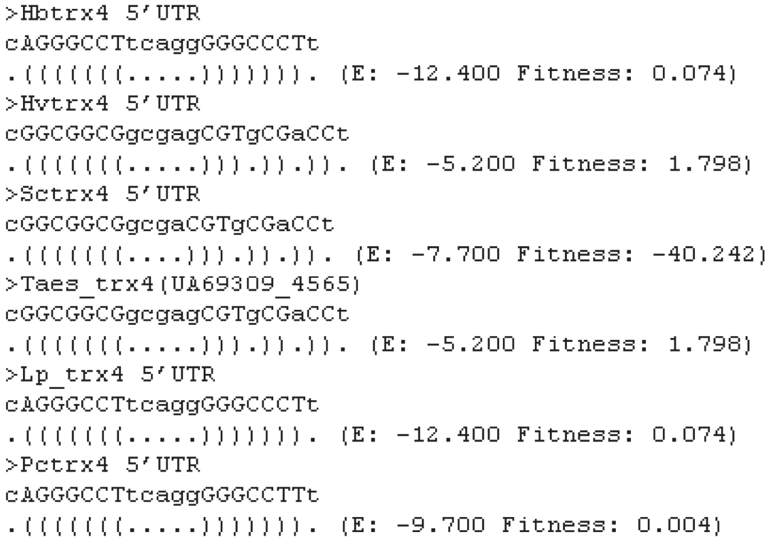
Figure 2.1 CLUSTALW alignment of cereal thioredoxin-h4 5′-UTRs. Alignment includes barley (*Hordeum vulgare*, Hv_trx4_5′UTR), wheat (*Triticum aestivum*, Taes_trx4_UA69309_4565_), perennial ryegrass (*Lolium perenne*, Lp_thrx4_5′UTR, rye (*Secale cereale*, Sc_trx4_5′UTR), bulbous barley (*Hordeum bulbosum*, Hb_trx4_5′UTR), sunolgrass (*Phalaris coerulescens*, PC_trx4_5′UTR). RNAProfile predicted structures are boxed (Figure 2.2B).



```
Hvtrx4_5'UTR              --AAAAGGCAUUUUCUUUCGGUAGGCGCACACG-GCACGAGCCGCGCCAG
Taes_trx4_UA69309_4565_   --------------------------AGGCACG-GCACGAGCCGCGCCAG
Sctrx4_5'UTR              --------------------------------------------------
Hbtrx4_5'UTR              ----AAAGCCGUUUCUUUCCGCAGGCGCACGCAAACACCAGCGGCGCCGG
Lp_trx4_5'UTR             AAAAAAAGCCGUUUCUUUCCGCAGGCGCACGCGAACACCAGCGGCGCCGG
Pctrx4_5'UTR              -------------------------CGCACG---GCACGAGCGGCGCCGG


Hvtrx4_5'UTR              CCAAGU---GGCGUGCGACGCGAGA---------CGCGGCACGGGCUCUC
Taes_trx4_UA69309_4565_   CCAAGU---GGUGUGCGACGCGAGA---------CGCGUCACGGG----C
Sctrx4_5'UTR              ------------GUGCGACGCGAGA---------CGCGUCACGGG----C
Hbtrx4_5'UTR              CCAACCCAGGUGGUGCAACCCCAAC----GCGGUUAAUCCACGGGCUCGC
Lp_trx4_5'UTR             CCGAGCCAGGUGGUGCGACGCCGAC----GCGGUUAAUCCACGGGCUCAC
Pctrx4_5'UTR              CCGGGCCAAGUGGUGCGACGCCGACCGACGCGGUUAAUUCACGGGCUCGC
                                      **** ** *                  ******    *

Hvtrx4_5'UTR              UCGCUCACGCCGGCCGGCCGGGAGCGGACGGCCCGAUCGAUC---CCAUC
Taes_trx4_UA69309_4565_   UCGCUCACGCCGGCCGGCCGGGAGCGGACGGGCCGGUCGAUC---C-AUC
Sctrx4_5'UTR              CCGCUCACACCGGCCGGCCGGGAGCGGACGGACGGGCCGAUCGAUCCAUC
Hbtrx4_5'UTR              UCACGCACGCACGCCGGCCGC--CCGCCCGACUCCAUCCAAU--CCCACC
Lp_trx4_5'UTR             UCACGCACGCACGCCGGCCGC--CCGCCCGACUCCAUCCAAU--CCCACC
Pctrx4_5'UTR              CCACGCAC----GCCGGCCGU--CCGUCCGACGCGGUUAA----UCCACC
                           * * ***    ********    **  **         *     * * *
                                        Type I stem-loop with one internal bulge
Hvtrx4_5'UTR              CAG---------GUCGUCGGCGUCGGC---GUCGUCGUCGUCGCCUCCAG
Taes_trx4_UA69309_4565_   CUG---------GUCGUCGGCGUCG------UCGUCGUCGUCGCCUCCAG
Sctrx4_5'UTR              CAGCCCCACCGGGUCGUCCCCGUCGUCCCCGUCGUCGUCGUCUCCUCCAG
Hbtrx4_5'UTR              GGGCC-------GUCGCCAUCGCCGUCGUCGUCCUCGUCGUC-CCUCCAA
Lp_trx4_5'UTR             GGGCC-------GUCGCCGUCGCCGUCGUCGUCCUCGUCGUCGCCUCCAG
Pctrx4_5'UTR              GCGUC-------GUCGCCGUCGCCGCCGUCGUC---GUCGUCGCCUCCAG
                            *         **** *  ** **      **   ****** ******

Hvtrx4_5'UTR              AAGCACGAGCCCAGCAUAGCACGGCCGCAGAAUAUUCCACGUCCCUUCCC
Taes_trx4_UA69309_4565_   AAGCACGAGCCCAGCAUAGCACGGCCGCAGAAUAUUCCACGUCCCUUCCC
Sctrx4_5'UTR              AAACACGAGCCCAGCAUAGCACGGCCGCAGAAUAUUCCACGUCCCUUCAC
Hbtrx4_5'UTR              AAACACAAACCGGGCAUACCACGGCCGCAAAAUAUUCCACUUCCCUUCCC
Lp_trx4_5'UTR             AAACACGAGCCGGGCAUAGCACGGCCGCAGAAUAUUCCACGUCCCUUCCC
Pctrx4_5'UTR              AAACACGAGCCGGCCAUAGCACGGCCGCGGAAUAUUCCACGUCCCUCCCC
                          ** *** * **   **** *********  ********** ***** * *

Hvtrx4_5'UTR              CUCAUCCUCCCCAGG-----------------CCCAGCAGCAAUA--AAA
Taes_trx4_UA69309_4565_   UUUGCGC-------------------------CCCAGCAGCAAUA--AAA
Sctrx4_5'UTR              UUUCCGGCUC----------------------CCCAGCAACAAUA--AAA
Hbtrx4_5'UTR              UUCCACCCUCUCCGUCCUCAUCCACGGGACCCCC-AGCACCAAUACAAAA
Lp_trx4_5'UTR             UUCCACCCUUUCCGUCCUCAUCCACGGGACCCCC-AGCAGCAAUACAAAA
Pctrx4_5'UTR              UCCCCCCUCCUCCGUCCUCAUCCACGGGACCCCCCAGCACCACUACAAAA
                                                          ** **** ** **  ***

Hvtrx4_5'UTR              GCGCG--GCGGCGGCGGC--G---AGCGUGCGACCUGUGACCACCAGACG
Taes_trx4_UA69309_4565_   GCGCG--GCGGCGGCGGC--GGCGAGCGUGCGACCUACGACCACCAGACG
Sctrx4_5'UTR              GCGCG--GCGGCGGCGGC--G----ACGUGCGACCUGCGACCACCAGACG
Hbtrx4_5'UTR              ACUCCCUGCGGCGGCG----GCAAAG--ACCAACCUACAACCACCAGAGG
Lp_trx4_5'UTR             ACUCGCUGCGGCGGCG----GCGAGG--AGCAACCUACGACCAGGAGAGG
Pctrx4_5'UTR              ACACGCUGCGGCGGCGCCCAGCGAAGCGAGCAACCUACGACCAGGAGAGG
                           * *   *********    *         * ****   ****  *** *

Hvtrx4_5'UTR              A--CCCAACCGAC-------CCACUGACCCA--CACCACACCACCGGGGA
Taes_trx4_UA69309_4565_   A--CCCAACCGAC-------CCACUGACCCA--CACCA-----CCGGAGA
Sctrx4_5'UTR              A--CCCAACCGAC-------CCACUGACCCA--CACCA-----CCGGAGA
Hbtrx4_5'UTR              AAACCCAACCAA------GUCCUCUGACCCC-UCACCACCG--CCGGAGA
Lp_trx4_5'UTR             AGACCCAACCAACCACUGGUCGUCUGACCCCCUCACCACCGG-CCGGAGA
Pctrx4_5'UTR              AGACC-GUCCACU-----GACCCCCACCACCACCACCACCG---GAGACA
                          *  **    **          *  *   * *   *****        *  *

Hvtrx4_5'UTR              GCUCUCCCUUUUACGUCAAUCAAUCGAAACCCAGUUAAAGAACCUCUUAA
Taes_trx4_UA69309_4565_   GCUCUUCCUUUUACGUCAAUCAAUCGAA-CCCAGUUAAAGAACCUCUUAA
Sctrx4_5'UTR              GCUCUCCCUUUUACGUUAAUCAAUCCAA-CCCAGUUAAAGAACCUCUUAA
Hbtrx4_5'UTR              GCUCUUCCUUGUACGUUAAUCAAUCCAACCCGAGUUAAGGAACCUCUUAA
Lp_trx4_5'UTR             GCUUUGCCUUCUACGUCAAUCAAUCCAGCCCGAGUUAAGGGACCUCUUAA
Pctrx4_5'UTR              GCGUUGCCUUGCACGUUAAUCCACCCAGCC-GAGUUAAGGAGCCUCUUAA
                          **  * ****  **** **** * * *  *  ****** *  ********

Hvtrx4_5'UTR              UUGCCCGCCAGGAGAUCCGCCAGGCUUAUCUUCGCCGUCUCCUCCGACCU
Taes_trx4_UA69309_4565_   UUGCCCGCCAGGAGAUCCGCCAGGCUUAUCUUCUCCGUCUCCUCCGACCU
Sctrx4_5'UTR              UUGCCCGCCAGGAGAUCCGCCAGGCUUACCUUCUCCGUCUCCUCCCAACC
Hbtrx4_5'UTR              UUGCCCG----GAGAUCCGCCAGGCUUACCGUCUUCGUCUCCU-CUCCUC
Lp_trx4_5'UTR             UUGCCCG----GAGAUCCGCCAGGCUUAUCGUCUUCGUCUCCUUCUCGUC
Pctrx4_5'UTR              UUGCCCG----GAGAUCCGCCAGGCUCGUCGUCUUCGUUUCC--------
                          *******    ***************   * **  *** ***
                            Type II stem-loop with two internal bulges
Hvtrx4_5'UTR              CGCCUCCACCCCCCCUCGCCCCGCGGCUUCUGGGUUCCUUGGCGCCAAAA
Taes_trx4_UA69309_4565_   CACCUCG---CCCCCUCGCCCCGCGGCUUCUGGGCUCCUUGGCGCCAAAA
Sctrx4_5'UTR              CAGCCCC-CGCCCCCUCGCCCCGCGG--UCUGGGUUCCUUGGCGCCAAAA
Hbtrx4_5'UTR              CGCCUCC-------------CCGCGG---------CUCUUGGCGCCAAAA
Lp_trx4_5'UTR             CCCCUCC-------------CCGCGG---------CUCUUGGCGCCAAAA
Pctrx4_5'UTR              ---CUUC-------------CCGCGG---------CUCUUGGCGCCAAAA
                             *                ******           *************

Hvtrx4_5'UTR              UCCCCGCUUCCGAUCCCAGAGGCCUUCAGGAAUCAGGGGGC--CUUUCAU
Taes_trx4_UA69309_4565_   UCCCCGCUUCCGAUCCCAG--GCGUUCAGGAAACAAGGGGC--CUUUCAU
Sctrx4_5'UTR              UCGUCGCUUCCCAUCCCAG--GCCUUCAGGAAUCAGGGGGCGUUUUUUAU
Hbtrx4_5'UTR              UCCCCCCUUCCCAUCUCAG-GGCCUUCAGGGGGCCCUUCGU-UCGGUGGU
Lp_trx4_5'UTR             UCCCCCUUUCCCAUCUCAG-GGCCUUCAGGGGGCCCUUCGU-UCGGUGGU
Pctrx4_5'UTR              UCCCCGCUCCCCAUCCCAG-GGCCUUCAGGGGGCCUUUUGC-UCAGUGCU
                          **  * *  ** *** *** ** ******   *     *      *  *
                                               Plain stem-loop
Hvtrx4_5'UTR              UCAGCGAUUCCUGCUAUUGUCAGUUCGGCAUGG
Taes_trx4_UA69309_4565_   UCAGCGAUUCCUGCCAGUGUCAGUUCGCCAUGG
Sctrx4_5'UTR              UCAGCUAGUAUUGUUGAUUGAAGUUCAAGAUGG
Hbtrx4_5'UTR              GUGCCUAGUUUUGUCGACGGAAGUUCACAAUGG
Lp_trx4_5'UTR             GUGCCUAGUUUCGUCGACGGAAGUUCACAAUGG
Pctrx4_5'UTR              GCGUGCUAGUUUGUCGAUUGAAGUUUACAAUGG
                                      *        ****    ****
```

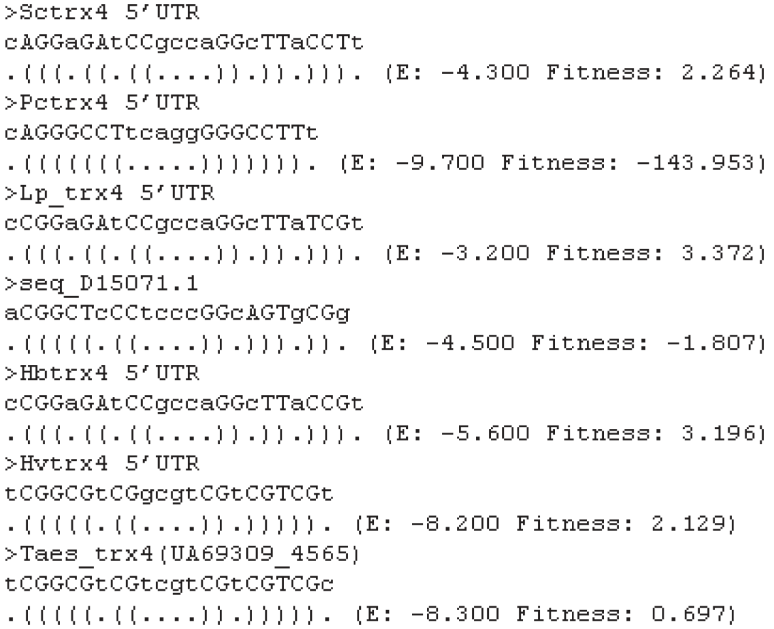
Figure 2.2 Highest scoring motifs predicted by RNAProfile on the *Trx4* 5′-UTR dataset with their respective energy and fitness value. A) Output without contaminating sequences B) Output with one IRE contaminating sequence (seq_D15071.1).

A) Trx4 stem-loops motifs



```
>Hbtrx4 5'UTR
cAGGGCCTtcaggGGGCCCTt
.(((((((.....))))))). (E: -12.400 Fitness: 0.074)
>Hvtrx4 5'UTR
cGGCGGCGgcgagCGTgCGaCCt
.(((((((.....))).)).)). (E: -5.200 Fitness: 1.798)
>Sctrx4 5'UTR
cGGCGGCGgcgaCGTgCGaCCt
.(((((((....))).)).)). (E: -7.700 Fitness: -40.242)
>Taes_trx4(UA69309_4565)
cGGCGGCGgcgagCGTgCGaCCt
.(((((((.....))).)).)). (E: -5.200 Fitness: 1.798)
>Lp_trx4 5'UTR
cAGGGCCTtcaggGGGCCCTt
.(((((((.....))))))). (E: -12.400 Fitness: 0.074)
>Pctrx4 5'UTR
cAGGGCCTtcaggGGGCCTTt
.(((((((.....))))))). (E: -9.700 Fitness: 0.004)
```

B) Trx4 stem-loop motifs and one IRE contaminant



```
>Sctrx4 5'UTR
cAGGaGAtCCgccaGGcTTaCCTt
.(((.((.((....)).)).))). (E: -4.300 Fitness: 2.264)
>Pctrx4 5'UTR
cAGGGCCTtcaggGGGCCTTt
.(((((((.....))))))). (E: -9.700 Fitness: -143.953)
>Lp_trx4 5'UTR
cCGGaGAtCCgccaGGcTTaTCGt
.(((.((.((....)).)).))). (E: -3.200 Fitness: 3.372)
>seq_D15071.1
aCGGCTcCCtcccGGcAGTgCGg
.(((((.((....)).))).)). (E: -4.500 Fitness: -1.807)
>Hbtrx4 5'UTR
cCGGaGAtCCgccaGGcTTaCCGt
.(((.((.((....)).)).))). (E: -5.600 Fitness: 3.196)
>Hvtrx4 5'UTR
tCGGCGtCGgcgtCGtCGTCGt
.(((((.((....)).))))). (E: -8.200 Fitness: 2.129)
>Taes_trx4(UA69309_4565)
tCGGCGtCGtcgtCGtCGTCGc
.(((((.((....)).))))). (E: -8.300 Fitness: 0.697)
```

Figure 2.3 Analysis of detected rbh orthologues. A) E-value distribution of barley homologues identified by the one-directional BLAST. B) Percentage of barley rbh orthologues in a particular E-value interval. C) Sequence coverage (%) of barley rbh orthologues in a particular E-value interval.

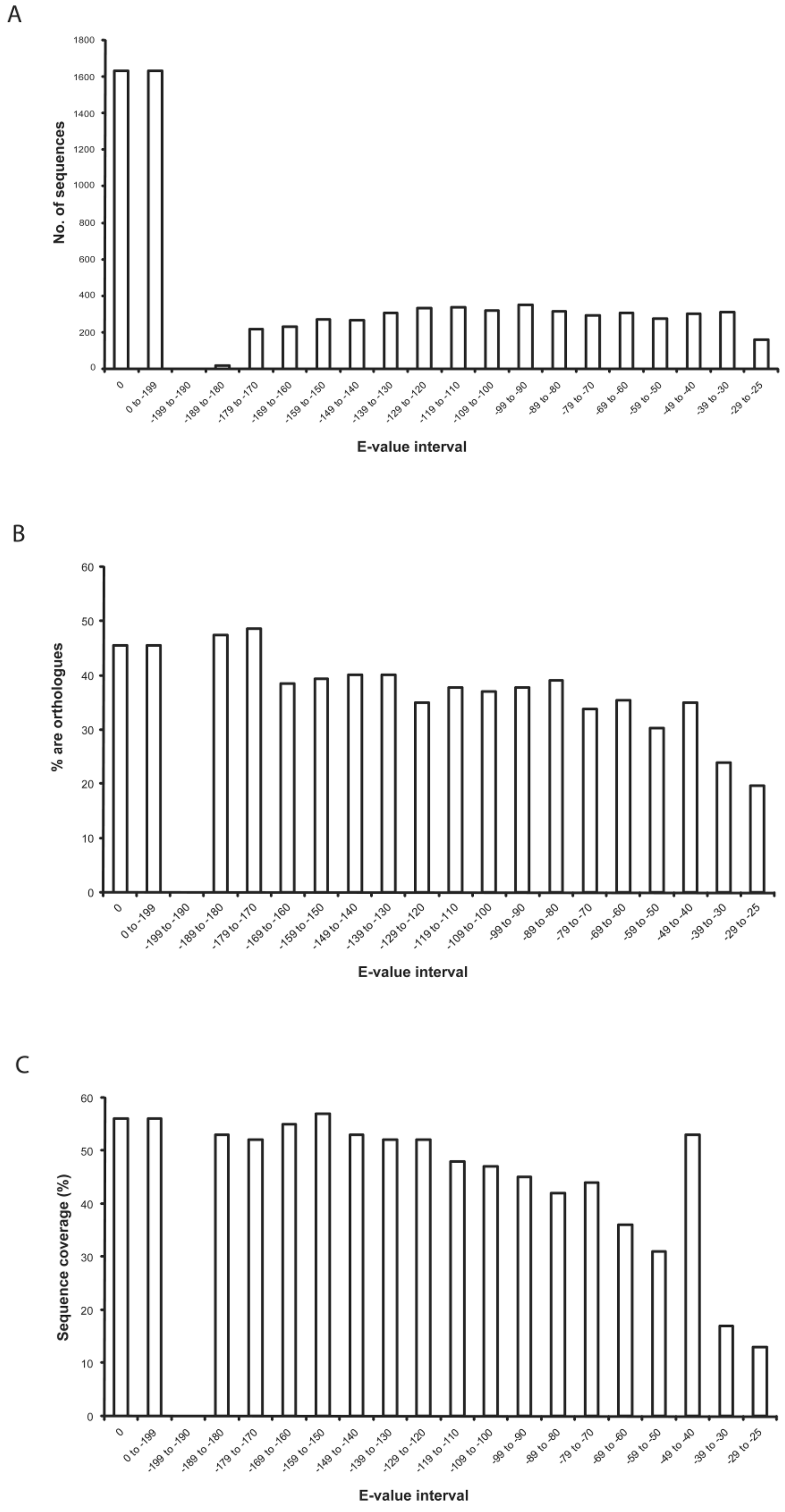
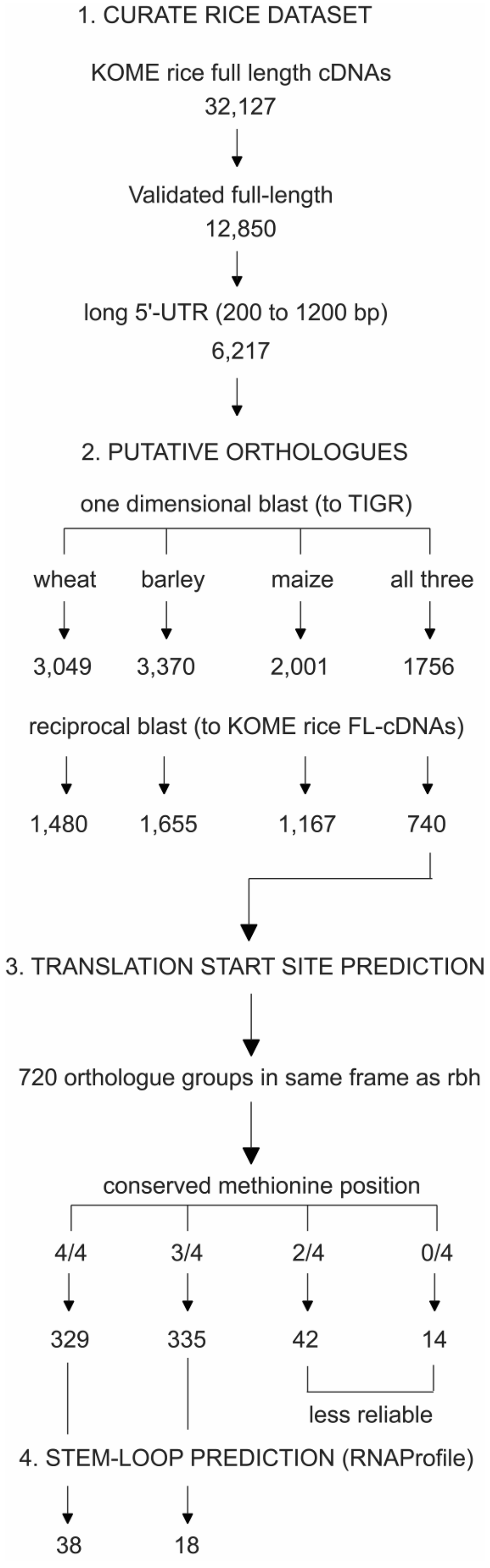

A
No. of sequences
0
200
400
600
800
1000
1200
1400
1600
1800
E-value interval
B
% are orthologues
0
10
20
30
40
50
60
E-value interval
C
Sequence coverage (%)
0
10
20
30
40
50
60
E-value interval
0
0 to -199
-199 to -190
-189 to -180
-179 to -170
-169 to -160
-159 to -150
-149 to -140
-139 to -130
-129 to -120
-119 to -110
-109 to -100
-99 to -90
-89 to -80
-79 to -70
-69 to -60
-59 to -50
-49 to -40
-39 to -30
-29 to -25

Figure 2.4 Overview of the stem-loop discovery pipeline. Step 1) Curate the rice dataset to select for full-length rice clones with long 5′-UTRs. Step 2) Find putative orthologues using rbh method. Step 3) Predict translation start site in each orthologue based on the conserved methionine position. Step 4) Predict conserved stem-loop motif in the 5′-UTR of each orthologue using the RNAProfile program.


1. CURATE RICE DATASET
KOME rice full length cDNAs
32,127
Validated full-length
12,850
long 5'-UTR (200 to 1200 bp)
6,217
2. PUTATIVE ORTHOLOGUES
one dimensional blast (to TIGR)
wheat
barley
maize
all three
3,049
3,370
2,001
1756
reciprocal blast (to KOME rice FL-cDNAs)
1,480
1,655
1,167
740
3. TRANSLATION START SITE PREDICTION
720 orthologue groups in same frame as rbh
conserved methionine position
4/4
3/4
2/4
0/4
329
335
42
14
less reliable
4. STEM-LOOP PREDICTION (RNAProfile)
38
18

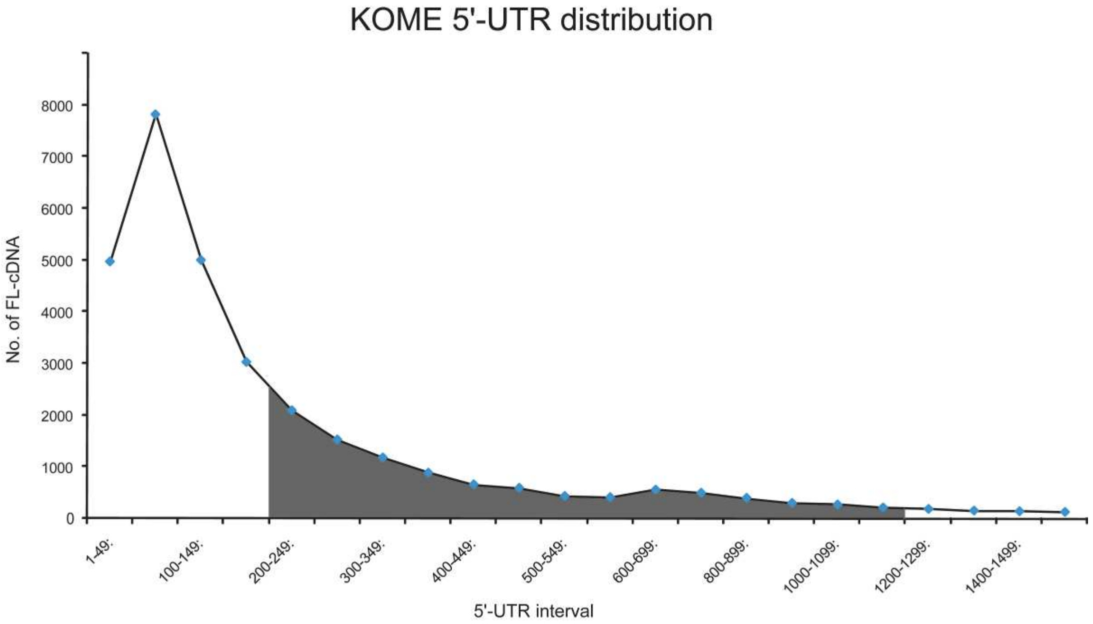
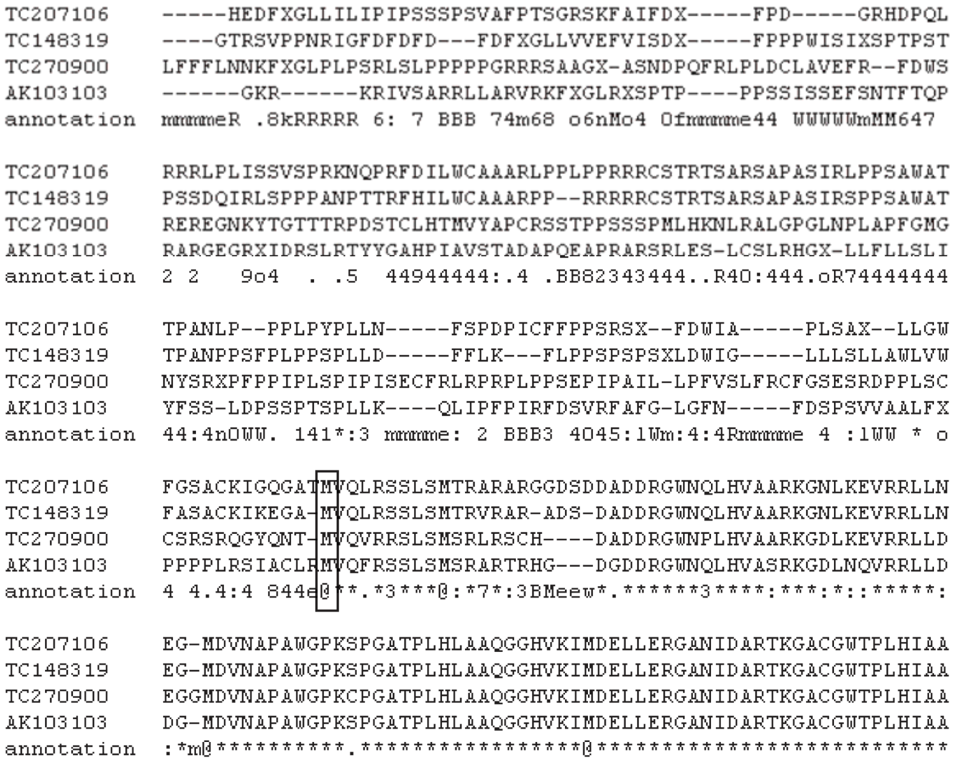


Figure 2.5 5′-UTR distribution of KOME full-length cDNA clones. Grey region represents cDNA clones with long 5′-UTRs between 200 and 1200 bp.



```
TC207106    -----HEDFXGLLILIPIPSSSPSVAFPTSGRSKFAIFDX-----FPD-----GRHDPQL
TC148319    ----GTRSVPPNRIGFDFDFD---FDFXGLLVVEFVISDX-----FPPPWISIXSPTPST
TC270900    LFFFLNNKFXGLPLPSRLSLPPPPPGRRRSAAGX-ASNDPQFRLPLDCLAVEFR--FDWS
AK103103    ------GKR------KRIVSARRLLARVRKFXGLRXSPTP----PPSSISSEFSNTFTQP
annotation  mmmmeR .8kRRRRR 6: 7 BBB 74m68 o6nMo4 0fmmmme44 WWWWWmMM647

TC207106    RRRLPLISSVSPRKNQPRFDILWCAAARLPPLPPRRRCSTRTSARSAPASIRLPPSAWAT
TC148319    PSSDQIRLSPPPANPTTRFHILWCAAARPP--RRRRRCSTRTSARSAPASIRSPPSAWAT
TC270900    REREGNKYTGTTTRPDSTCLHTMVYAPCRSSTPPSSSPMLHKNLRALGPGLNPLAPFGMG
AK103103    RARGEGRXIDRSLRTYYGAHPIAVSTADAPQEAPRARSRLES-LCSLRHGX-LLFLLSLI
annotation  2 2   9o4  . .5  44944444:.4 .BB82343444..R40:444.oR74444444

TC207106    TPANLP--PPLPYPLLN-----FSPDPICFFPPSRSX--FDWIA-----PLSAX--LLGW
TC148319    TPANPPSFPLPPSPLLD-----FFLK---FLPPSPSPSXLDWIG-----LLLSLLAWLVW
TC270900    NYSRXPFPPIPLSPIPISECFRLRPRPLPPSEPIPAIL-LPFVSLFRCFGSESRDPPLSC
AK103103    YFSS-LDPSSPTSPLLK----QLIPFPIRFDSVRFAFG-LGFN-----FDSPSVVAALFX
annotation  44:4n0WW. 141*:3 mmmme: 2 BBB3 4045:1Wm:4:4Rmmmme 4 :1WW * o

TC207106    FGSACKIGQGATMVQLRSSLSMTRARARGGDSDDADDRGWNQLHVAARKGNLKEVRRLLN
TC148319    FASACKIKEGA-MVQLRSSLSMTRVRAR-ADS-DADDRGWNQLHVAARKGNLKEVRRLLN
TC270900    CSRSRQGYQNT-MVQVRRSLSMSRLRSCH----DADDRGWNPLHVAARKGDLKEVRRLLD
AK103103    PPPPLRSIACLRMVQFRSSLSMSRARTRHG---DGDDRGWNQLHVASRKGDLNQVRRLLD
annotation  4 4.4:4 844e@**.*3***@:*7*:3BMeew*.******3****:***:*::*****:

TC207106    EG-MDVNAPAWGPKSPGATPLHLAAQGGHVKIMDELLERGANIDARTKGACGWTPLHIAA
TC148319    EG-MDVNAPAWGPKSPGATPLHLAAQGGHVKIMDELLERGANIDARTKGACGWTPLHIAA
TC270900    EGGMDVNAPAWGPKCPGATPLHLAAQGGHVKIMDELLERGANIDARTKGACGWTPLHIAA
AK103103    DG-MDVNAPAWGPKSPGATPLHLAAQGGHVKIMDELLERGANIDARTKGACGWTPLHIAA
annotation  :*m@**********.*****************@***************************
```

Figure 2.6 CLUSTALW alignment of the nucleotide translation of the 5′ region of rice Ankyrin-2 (AK103103) and its cereal rbh orthologues TC207106 (wheat), TC148319 (barley), and TC270900 (maize). The annotation line shows the extended code (Table 2.4). The rectangle box shows a conserved methionine position as the predicted translation start site.