Compilation of mRNA Polyadenylation Signals in Arabidopsis Revealed a New Signal Element and Potential Secondary Structures[1[w]]

Johnny C. Loke[2], Eric A. Stahlberg, David G. Strenski, Brian J. Haas, Paul Chris Wood, and Qingshun Quinn Li*

Department of Botany, Miami University, Oxford, Ohio 45056 (J.C.L., P.C.W., Q.Q.L.); Ohio Supercomputer Center, Columbus, Ohio 43212 (E.A.S.); Cray, Inc., Brighton, Michigan 48116 (D.G.S.); and The Institute of Genomic Research, Rockville, Maryland 20850 (B.J.H.)

Using a novel program, SignalSleuth, and a database containing authenticated polyadenylation [poly(A)] sites, we analyzed the composition of mRNA poly(A) signals in Arabidopsis (*Arabidopsis thaliana*), and reevaluated previously described cis-elements within the 3′-untranslated (UTR) regions, including near upstream elements and far upstream elements. As predicted, there are absences of high-consensus signal patterns. The AAUAAA signal topped the near upstream elements patterns and was found within the predicted location to only approximately 10% of 3′-UTRs. More importantly, we identified a new set, named cleavage elements, of poly(A) signals flanking both sides of the cleavage site. These cis-elements were not previously revealed by conventional mutagenesis and are contemplated as a cluster of signals for cleavage site recognition. Moreover, a single-nucleotide profile scan on the 3′-UTR regions unveiled a distinct arrangement of alternate stretches of U and A nucleotides, which led to a prediction of the formation of secondary structures. Using an RNA secondary structure prediction program, mFold, we identified three main types of secondary structures on the sequences analyzed. Surprisingly, these observed secondary structures were all interrupted in previously constructed mutations in these regions. These results will enable us to revise the current model of plant poly(A) signals and to develop tools to predict 3′-ends for gene annotation.

Messenger RNA polyadenylation is a crucial step during the maturation of most eukaryotic mRNA, in which a polyadenine [poly(A)] tract is added to the cleaved 3′-end of a precursor mRNA (pre-mRNA) posttranscriptionally. Such a modification of mRNA has been shown to affect its stability, translatability, and nuclear-to-cytoplasmic export (Zhao et al., 1999). The posttranscriptional processing of mRNA is an event that has also been found tightly coupled with splicing and transcription termination (Proudfoot et al., 2002; Proudfoot, 2004). Thus, it is an essential processing event and the integral part of gene expression.

The polyadenylation process requires two major components: the cis-elements or poly(A) signals of the pre-mRNA, and the trans-acting factors that carry out the cleavage and addition of the poly(A) tail at the 3′-end. These trans-acting factors are a complex of about 25 to 30 proteins involved in signal recognition, cleavage, and polyadenylation (Proudfoot, 2004). These proteins seem to be conserved among eukaryotic organisms. However, the poly(A) signals have been found to differ widely among yeast (*Saccharomyces cerevisiae*), animals, and plants in terms of signal locations and sequence content. The highly conserved AAUAAA element in mammals becomes a minor signal in plant and yeast genes, and the ubiquitous downstream elements of mammalian pre-mRNAs are nowhere to be found in yeast and plants. The latter two possess an enhancing element located far upstream of the cleavage site (Zhao et al., 1999).

Previous understanding of these signal elements was derived mostly through conventional genetic and some biochemical analyses, which are both tedious and time consuming to perform. The availability of genomic, full-length cDNA and expressed sequence tag (EST) sequences through large-scale genome-sequencing projects makes it possible to search for poly(A) signals using bioinformatics tools (Graber et al., 1999b; Beaudoing et al., 2000; Hajarnavis et al., 2004). The efficacy of this approach has been demonstrated in the recent stream of publications revisiting the poly(A) cis-elements in different organisms in which such signal models were established based on conventional genetic analysis. Although still prominent, the status of the canonical AAUAAA pattern in mammals has been challenged (MacDonald and Redondo, 2002). The most detailed information comes

[1] This work was supported in part by Miami University (Shoupp Award and Botany Department Academic Challenge Grant), the Ohio Plant Biotechnology Consortium, and the National Science Foundation (grant no. MCB–0313472 to Q.Q.L.).

[2] Present address: Department of Medicine, Mount Sinai School of Medicine, 1425 Madison Ave., New York, NY 10029–6574.

* Corresponding author; e-mail liq@muohio.edu; fax 513–529–4243.

[w] The online version of this article contains Web-only data.

Article, publication date, and citation information can be found at www.plantphysiol.org/cgi/doi/10.1104/pp.105.060541.

from the analysis of yeast poly(A) signals (Graber et al., 1999b, 2002; Van Helden et al., 2000) in which a large number of variants of efficiency elements were found at the same position as AAUAAA. It has been proposed that there are potential secondary or higher ordered structures that may be formed within these sequence elements recognizable by protein factors (Zarudnaya et al., 2003). However, the existence of such structures has not been established by wet lab experiments.

Conventional genetic mutagenesis studies revealed that plant poly(A) signals are composed of three major groups: far upstream elements (FUE), near upstream elements (NUE; an AAUAAA-like element), and cleavage sites (CSs; Rothnie, 1996; Li and Hunt, 1997; Rothnie et al., 2001). The composition of plant consensus signals, such as CSs, is a YA (CA or UA) dinucleotide situated within a U-rich region. The NUE region is A rich and spans about 6 to 10 nucleotides (nt) located between 13 and 30 nt upstream of the CS (referred to as locations −13 to −30; Hunt, 1994; Li and Hunt, 1995). FUE, the control or enhancing element, is a combination of rather ambiguous UG motifs and/or the sequence UUGUAA (Hunt, 1994). These findings were from detailed molecular analysis of only a few genes of different plant species and viruses, the exception being a few thousand ESTs initially examined by statistical modeling (Graber et al., 1999b).

The full scope of the prominent patterns of plant poly(A) signals has not been revealed previously, and this has been an obstacle toward using such information for gene annotation and for better understanding of how the plant polyadenylation machinery operates. The major principles of gene annotations are based on the identification of functional RNA and coding sequences (Arabidopsis Genome Initiative, 2000). A better understanding of the 3′-untranslated region (UTR) cis-elements will enhance principles of gene predictions that ultimately improve the accuracy of gene annotation (Graber et al., 2002). The major obstacle in identification of unique genes through cDNA analysis is the prediction of the terminal exon, misidentification of which may cause either two genes to be fused together or one gene to split into two. Improvement in 3′-UTR annotation will diminish such problems (Hajarnavis et al., 2004). The large-scale bioinformatics approach that we deployed to study polyadenylation signals will not only decipher primordial information of gene constructs in relation to regulation of gene expression via polyadenylation mechanisms, but also reveal higher ordered structures of poly(A) signals that will ultimately open a new frontier in gene annotation technology by predicting the ends of the genes.

With the advancement of genomic research and availability of large numbers of plant ESTs, particularly of the model species Arabidopsis (*Arabidopsis thaliana*), we will be able to collect large-scale datasets for genome-wide poly(A) signal analysis. In this article, we report on efforts to characterize regions of significance in which poly(A) signals reside. Our database consists of two datasets of 3′-UTR sequences covering about 17,000 independent genes, one with 8,160 ESTs with authenticated poly(A) sites, the other with 16,211 full-length cDNA downloaded from The Arabidopsis Information Resource (TAIR). Both datasets were searched independently with supercomputers to probe for the signal pattern locations based on a working model built with conventional genetic analyses of plant poly(A) signals (Hunt and Messing, 1998). We also describe the potential of secondary structure formation within these 3′-UTR regions. Our results will be the basis of building a new algorithm to search for regular and alternative poly(A) sites at a genome level to be integrated into genome annotation programs.

RESULTS

The NUE

To compile plant poly(A) signals using a computer program, it is necessary to generate a numeric model or location of the cis-elements that are sought. To this end, we constructed a working model based on the characterized plant poly(A) signals by conventional genetic or biochemical approaches on a few genes, including the cauliflower mosaic virus (CaMV) 35S transcript, the pea (*Pisum sativum*) small subunit of Rubisco (*rbc*S), the Agrobacterium T-DNA *ocs* gene, and the maize (*Zea mays*) 27-kD protein gene (Rothnie, 1996; Li and Hunt, 1997). There are only a few genes from which poly(A) signals were analyzed in detail through mutagenesis. According to this model, the locations of the elements are at a relative distance from the CS, which is a dinucleotide most likely to be YA. Using the CS as a reference point, the NUE (6–10 nt in length) is located 10 to 40 nt upstream and the FUE (60–100 nt in length) starts from at least 29 nt upstream.

We started with the NUE because it is a region that is slightly better understood from the literature and expected to be more conserved than the FUE. The NUE is defined as a signal element located between −13 nt to −30 nt upstream of the CS (position −1 anchored at the last nucleotide of the 3′-end of each cDNA sequence; thus the upstream sequence and the downstream sequence will have a "−" or "+" designation, respectively; Rothnie, 1996; Li and Hunt, 1997). The first approach was to expand the NUE region and to search for all possible patterns and signals in the subregion from −1 to −50 nt for all the sequences. An example of the results (top 50 patterns of 6-nt pattern length) in the form of a two-dimensional (2-D) matrix output, showing pattern count versus location, is illustrated in Figure 1A. The patterns are ranked based on the deviation of the maximum count from the median value, as described in "Materials and Methods." Pattern sizes of 3 to 11 nt in length were exhaustively searched, but only the 6-nt data are

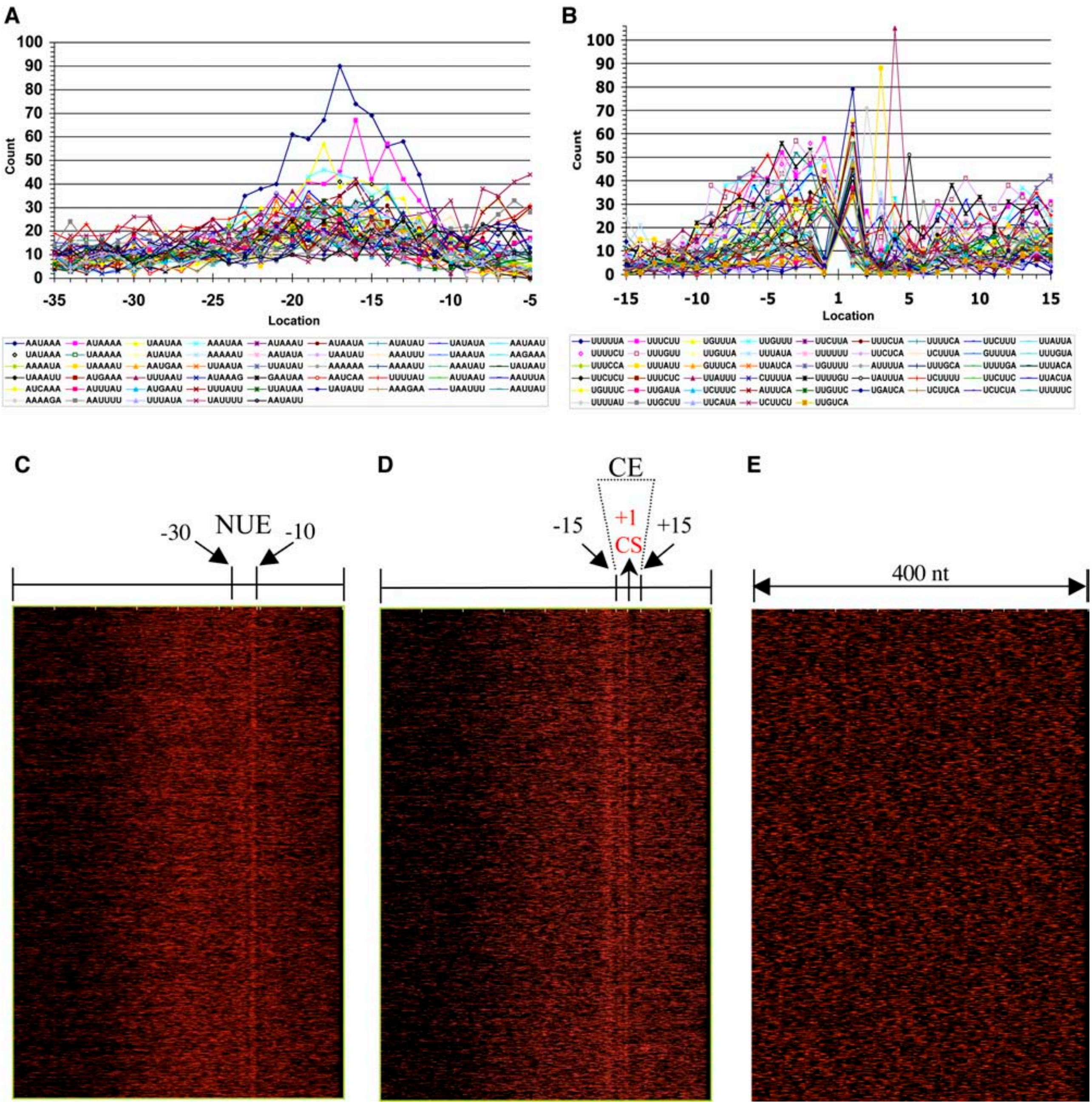


Figure 1. NUE and CE top patterns and their distributions. A, The 8-K dataset scanned for both the NUE and CE 6-nt signal parameter. Top 50 NUE patterns sorted according to counts at upstream position −30 to −1. B, Six-nucleotide scan results (top 50 patterns) yield the distinct peaks of the CE flanking the CS. C to E, Pixel images showing the location of high-count patterns in the NUE, CE regions, and a random dataset as a control. The original images were 400 × 8,160 pixels and were resized here to fit in the figure. C, Top 50 NUE (6 nt) visual alignment as in the sequence graphic view. Each sequence is presented as a single pixel on a horizontal line, and the white bars (marked on the rightmost position of the 6 nt) represent each occurrence of the signal patterns with respect to their location on each sequence. The continuous vertical band of lines from top to bottom indicates the common locations of the signal element. D, Top 50 CE (6 nt) alignment where +1 is the CS position. The two bandings occurring in between the −15 and +15 positions denote the CE. E, Results from a computer-generated DNA dataset comprising 10,000 random sequences as a control showing no significant pattern formation in this 2-D view.

presented here. A full list of results for the top 1,000 most common patterns for sizes from 3 to 11 nt are listed in Supplemental Table I.

A few striking features were found when searching the NUE region of the sequences, namely, that a few patterns had a much larger deviation than the rest of the top 50 patterns. The pattern AAUAAA came out at the top of the list in this region, followed by other less dominant ones. However, in this region, AAUAAA can only account for about 10% of the signals. Second, AAUAAA and related patterns are located at the expected position of the working model at about −13 to −30 nt upstream from the CS (Rothnie, 1996; Li and Hunt, 1997). Surprisingly, there is a second unexpected peak immediately before the CS (Fig. 1A), which was previously unaccounted for in the working model. Moreover, the patterns in this region (−1 to −10) are very different from those of the NUEs, and the peak seemed to distribute across the CS, leaving some patterns after the CS, indicative of possible independent signal elements.

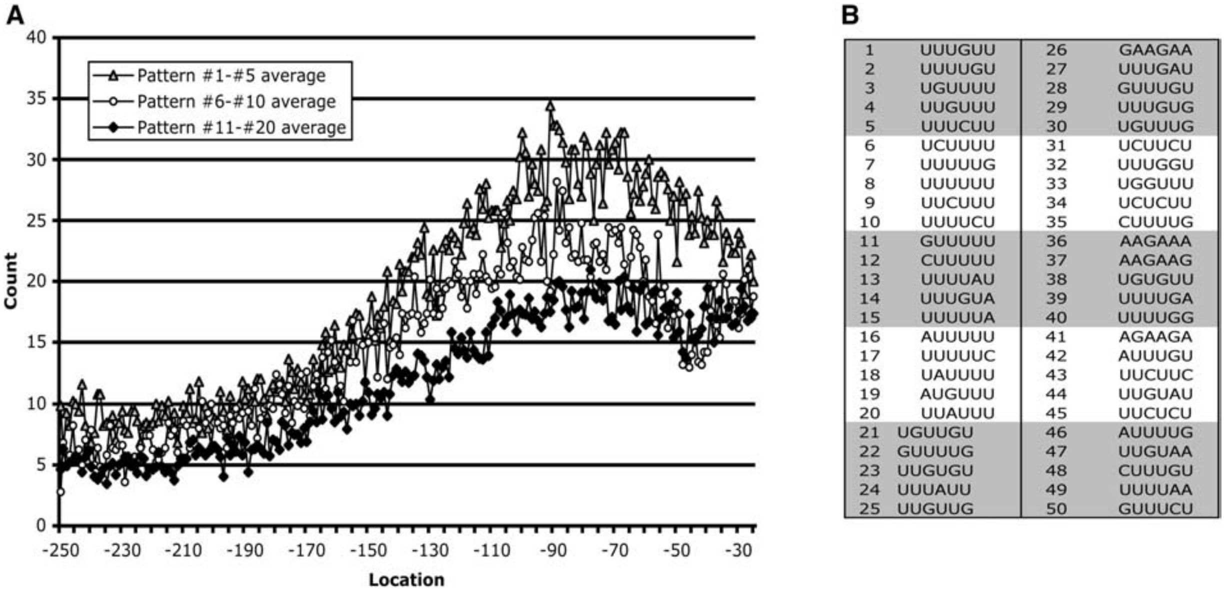
A New Poly(A) Signal: Cleavage Element

To determine whether there is a new signal element around the CS, another dataset (UTR + downstream) was created to include 100 nt of genomic sequence downstream of the CS for each of the sequences in the 3′-UTR 8-K dataset. When the region of the sequences (−15 to +20) was scanned for predominant patterns, the full peaks were seen collectively, with the CS in the center (Fig. 1B). Due to the nature of the SignalSleuth program, the patterns are counted from the position of the rightmost nucleotide (from 3′ to 5′). Thus, considering the 6-nt size patterns, most of the patterns with the highest counts are across the CS. Moreover, the region −10 to +15 is highly saturated with such patterns (the top 1,000 lists for 3- to 11-nt scan results are available in Supplemental Table I). It is clear that this region of the RNA consists of a signal element that was not previously documented in plants. The new signal element is termed cleavage element (CE).

The SignalSleuth program also has the capability to generate an output in the form of a 2-D image, as explained in ''Materials and Methods,'' where the width represents the full length of the 400-nt sequence, and each red pixel represents one of the top 50 patterns. Thus, the locations of the top 50 patterns on the sequences can be marked on the image. As shown in Figure 1, C and D, these 2-D images clearly demonstrate the existence of the top signal patterns located within the NUE or CE regions (vertical bands) in the 3′-UTR dataset, while no significant signals can be seen in the random DNA sequence dataset (Fig. 1E). Note that these top 50 patterns were reranked in narrower regions manually corresponding only to the NUE (−10 to −30) or the CE (−10 to +10), respectively. This stipulation ensures visual evidence that the top patterns on the NUE or CE lists are representative of each element rather than a sum of the two elements. It seems that the CE patterns are located mainly in two regions, one at the CS, the other a few nucleotides apart at the right of the CS.

The FUE

From mutagenesis analysis of the FUE region, it has been defined that this is a region with low conservation for cis-acting element patterns. Molecular evidence suggested that the FUE region should span about 60 to 100 nt with combinations of motifs from 6 to 18 nt in length (Sanfacon et al., 1991; Mogen et al., 1992). With the current version of SignalSleuth, we limit our search in the FUE region to a pattern with



1	UUUGUU	26	GAAGAA
2	UUUUGU	27	UUUGAU
3	UGUUUU	28	GUUUGU
4	UUGUUU	29	UUUGUG
5	UUUCUU	30	UGUUUG
6	UCUUUU	31	UCUUCU
7	UUUUUG	32	UUUGGU
8	UUUUUU	33	UGGUUU
9	UUCUUU	34	UCUCUU
10	UUUUCU	35	CUUUUG
11	GUUUUU	36	AAGAAA
12	CUUUUU	37	AAGAAG
13	UUUUAU	38	UGUGUU
14	UUUGUA	39	UUUUGA
15	UUUUUA	40	UUUUGG
16	AUUUUU	41	AGAAGA
17	UUUUUC	42	AUUUGU
18	UAUUUU	43	UUCUUC
19	AUGUUU	44	UUGUAU
20	UUAUUU	45	UUCUCU
21	UGUUGU	46	AUUUUG
22	GUUUUG	47	UUGUAA
23	UUGUGU	48	CUUUGU
24	UUUAUU	49	UUUUAA
25	UUGUUG	50	GUUUCU

Figure 2. The top FUE patterns and their distributions in 3′-UTR. A, The profile of the top 20 FUE signals (6 nt) indicates the frequency of the patterns in the FUE region (−30 to −170). To simplify the graph, the counts of the top 20 patterns were divided into three groups (see boxed legend in A): Pattern numbers 1 to 5 average are the average counts of the first five patterns at each location; pattern numbers 6 to 10 average are the average of the second five patterns, and so on. B, A list of the top 50 FUE patterns based on the total counts in the 8-K dataset. A full list and the counts for each pattern are available in the supplemental data.

word length of 11 nt. Figure 2A represents the output from the FUE search of the top 20 patterns of 6 nt in length, presenting a cluster of pattern spikes with no defined individual signal peaks, as seen in the NUE. However, as the search approaches the coding regions, there is a sudden shift in the combinations of nucleotides. The FUE patterns do not occupy one small domain, but rather span across an approximately 125-nt region, as indicated by the drops of peak density flanking this FUE region (Fig. 2A). The top 50 patterns are listed in Figure 2B, while a full list of the top 1,000 patterns of size 3 to 11 nt is available in Supplemental Table I.

The NUE, CE, and FUE compilations were also done with the 16-K dataset. Similar results, including rankings of the signals, were found (see Supplemental Table II).

Nucleotide Composition of 3′-UTRs

The NUE and CE patterns seemed to be notably rich in A and U nucleotides. This prompted the need to analyze the nucleotide composition profile of the 3′-UTR sequences in the databases. A full-scan sweep of the 8-K dataset from −250- to +100-nt positions revealed intriguing findings (Fig. 3A). First, the distributions of A and U are clearly distinct, where the ups and downs of the curves complement each other. This is true between −200 to +60 nt, covering a span of a 260-nt region. The only exception is at the CS, where C seemed to have a spike (for the previously known YA dinucleotide). Second, distinct A and U profiles are also seen in different signal elements. The FUE region has a high U content, while the NUE region has a high A content, with a clear transition between each. We

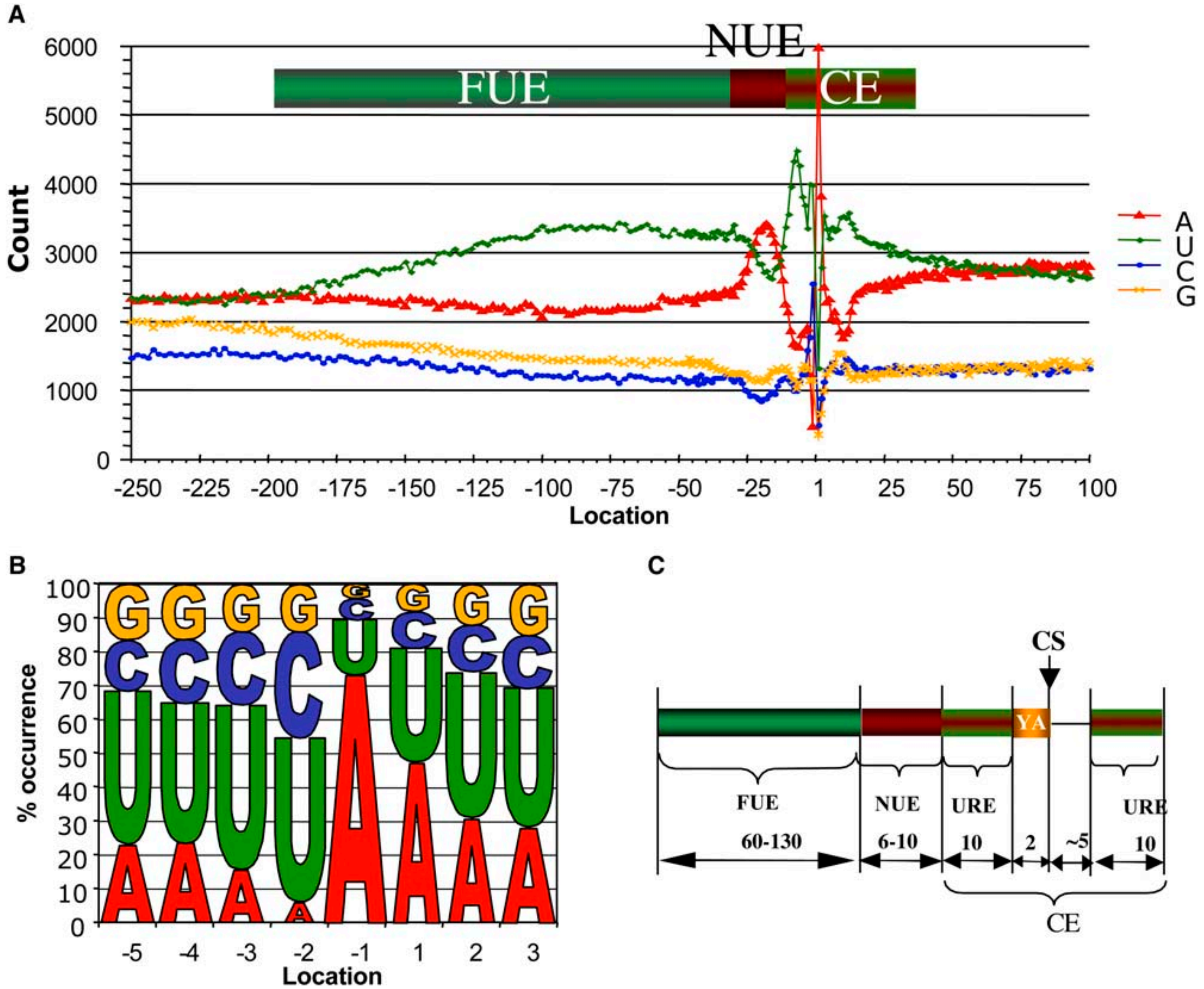


Figure 3. Single-nucleotide profile of 3′-UTR and a current model of plant poly(A) signals. A, Single-nucleotide scan from positions −250 to +100 in the whole UTR + downstream region. Distinct profiles flanking the CS are now named CEs. B, Sequence logo generated from the actual percentage of each of the four nucleotide's occurrence in the 8-K dataset, indicating preferred nucleotides flanking the CS (−5 to +3 nt). C, A current model for Arabidopsis mRNA poly(A) signals. URE, U-rich regions, which are found flanking both up- and downstream of the CS.

noticed that the CE region is composed of a complex, but clear, nucleotide composition with alternating A- and U-rich submotifs. Third, the location of the elements can be clearly identified as previously proposed, e.g. the NUE is located at about 20 nt upstream of the CS and the FUE at about −25 to −160, covering a region of over 100 nt. The occurrence of the NUE and the CE are consistent with the alignment shown in the images in Figure 1, C and D. Finally, the CS −1 position can be recognized by an A in about 73% of the sequences, followed by distinct U or C at the −2 position in a total of 80% of the sequences (see Fig. 3B). The −2 position C was found to be 5.64-fold (from 453 to 2,553 counts) at that location compared with the adjacent −1 position.

To verify whether such a nucleotide distribution profile holds in the broader range of sequences, we also scanned the 16-K dataset. The results showed similar patterns of nucleotide distribution at this region (see Supplemental Fig. 1). This result indicates that both datasets, which cover about 17,000 unique genes of the Arabidopsis genome, possess similar profiles of poly(A) signals.

Based on the information presented here, we propose a new model for Arabidopsis mRNA poly(A) signals (Fig. 3C). From the single-nucleotide scan analysis, there is no obvious spatial separation among the three types of signals, FUE, NUE, and CE. However, the locations of the signals seem to be well positioned, where the transition from one to the other is complete. The proposed CE contains a subset of small cis-elements: two U-rich sequences flanking both sides of the CS.

Potential Secondary Structure Formation of 3′-End Regions

The occurrence of patterns flanking CSs and the order of an alternate arrangement of residues of complementarity for the poly(A) signals (Fig. 3A) indicate the possibility of the formation of higher order structures. To explore this, we used the mFold 2.3 model analysis as described by Zuker (2003). The folding results from the manual analysis of a subset of randomly selected 3′-UTR sequences (n = 128) from the 8-K dataset indicating categorical secondary structures (Fig. 4). Regions where the predicted poly(A) signals may reside were revealed from trends of secondary structures related to their locations, especially flanking the CS. These secondary structures can be categorized into three main groups based on the location of the CS: group I, CS clustered; group II, CS stem loop; and group III, CS flat. Using such classification, a total of 128 3′-UTR foldings were completed, with 57.0% of the foldings falling into group I, 28.1% in group II, and 14.8% in group III. The stem loop structures found around the CS were all from the base pairing of adjacent sequences. For analysis purposes, the input of each sequence was 400 nt for maximal coverage of possible pre-mRNA 3′-UTR sequence length. Similar

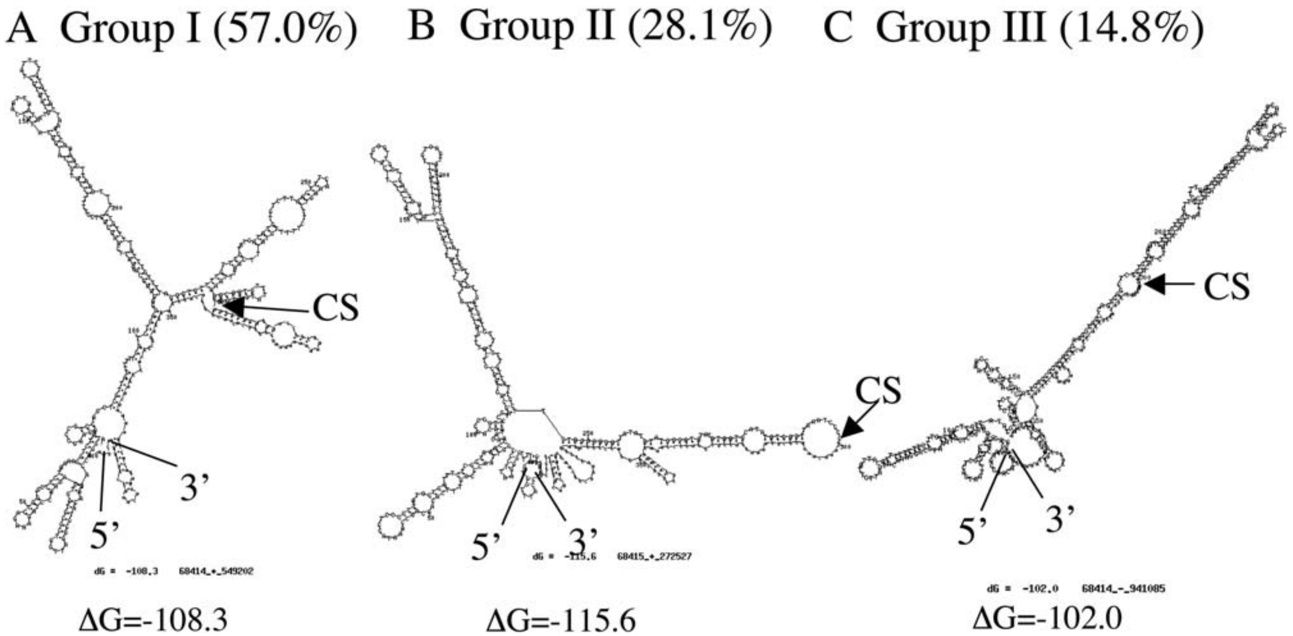


Figure 4. Representative secondary structure models of 3′-UTR predicted by mFold. A, Group I, the CS (all at position 300, total 400 nt used) is situated on a cluster of stem loop structures. B, Group II, the CS is situated on or around the stem loop, but is not flanked with a cluster of secondary structures. C, Group III, the CS is situated in the middle of a long structure. The free energy indicated is for individual structures. Percentages of structures in each group are given, and the ends of the RNA are as marked. The mFold program tends to match both ends of the RNA. This does not interfere with the structure prediction here because the structures around the CSs are formed by the adjacent sequences, not by the end sequences.

results (structures around the CS) have been obtained from folding of shorter sequences of approximately 220 nt, or longer sequences to include some coding sequences (approximately 800 nt).

Mutagenesis Data Support the Existence of the Predicted Secondary Structures

The secondary structure described above was based on in silico prediction. Support of such structures would be strongest from experimental evidence if the alteration of these structures could interrupt the function of the poly(A) signals. Here we analyzed published data based on conventional mutagenesis on a set of genes. It was this kind of classical analysis that contributed to the understanding of the poly(A) signals in plants.

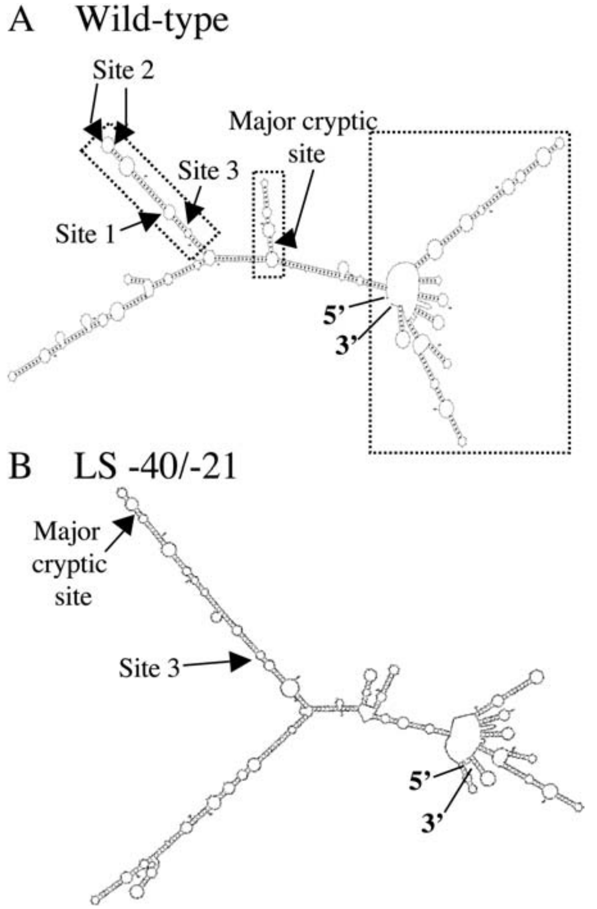


Figure 5. The secondary structure of pea *rbc*S E-9 3′-UTR and a mutation. The structures were derived using the lowest free-energy levels (−190.8 and −194.0, respectively). A, Wild type. The boxes indicate three cleavage regions. The arrows indicate the normal CSs and a main cryptic site (site 4 in the original paper; Mogen et al., 1992). The exact locations of other cryptic sites in the far-right boxed area were not given in the paper. The ends of the sequence are as marked. B, An example of LS mutations that altered the secondary structures, as well as the efficiencies of the CSs (Mogen et al., 1992). Normal sites 1 and 2 were eliminated, and site 3 and the major cryptic site became dominant.

From the predicted secondary structure of pea *rbc*S E-9 3′-UTR (Fig. 5A), it is clear that the primary and cryptic CS (based on Hunt and MacDonald, 1989) are all situated on three stem loop areas. Mutation of the signals for one area would lead to the uses of the CSs in the next area. Thus, deletion of the NUE and FUE responsible for CSs 1, 2, and 3 (on the left) led to the usage of the major cryptic sites (Fig. 5A, middle and right boxed areas; based on C6 mutants in figure 1 of Hunt and MacDonald, 1989), and so on (other mutants in the mentioned figure).

The finer correlation of the secondary structure and CS efficiency can be better illustrated by the linker-scanning (LS) analysis on the same *rbc*S 3′-UTR as described (Mogen et al., 1992). The authors noted that 20-nt LS of the area −220 to −61 did not yield the results of the deletion of the whole region by Hunt and MacDonald (1989). This can be explained at the secondary structure level: The deletion abolished the far left area structure (Fig. 5A), but the LS did not significantly alter the structure (data not shown). However, with LS −60/−41, −40/−21, and −20/−1, the stem loop arrangements around the CSs were drastically altered, respectively (e.g. Fig. 5B). As a result, the efficiency of the normal CS altered accordingly (Mogen et al., 1992).

The mutagenesis analysis of CaMV poly(A) signals also offers a clue pertaining to the importance of the secondary structures. Deletion of the CaMV NUE pattern AAUAAA almost abolished the use of a corresponding poly(A) site (only 15% that of the wild type; see figure 9 of Rothnie, 1996). However, in the same set of experiments, a single-nucleotide change of most of the nucleotides in AAUAAA had only a subtle impact on the poly(A) selection (80%–100% activity remained). This can be elaborated using the secondary structure model of the CaMV 3′-UTR, which falls into group III in our structure models (Fig. 4), as depicted in Figure 6A. There is no visible secondary structural change in the single-nucleotide mutations of the CaMV NUE, except that the structure wobbles slightly near the NUE region (Fig. 6D). However, when AAUAAA is deleted, there is no observed secondary structure at the NUE region; hence the dramatic reduction in the use of the CS (Fig. 6B). Moreover, when 3 of the 6 nt were changed (from AAUAAA to UAGAAU), the efficiency of the signal was reduced to 51% calculated from the band intensity of the S1 nuclease protection assay (figure 3 in Mogen et al., 1990), and the structural change is visible (Fig. 6C).

DISCUSSION

Using in silico analysis tools, we compiled Arabidopsis nuclear mRNA poly(A) signals from two independently produced 3′-UTR datasets covering about 17,000 independent genes. Beyond confirming the previous working model on the NUE and FUE, we revealed complex nucleotide distribution patterns

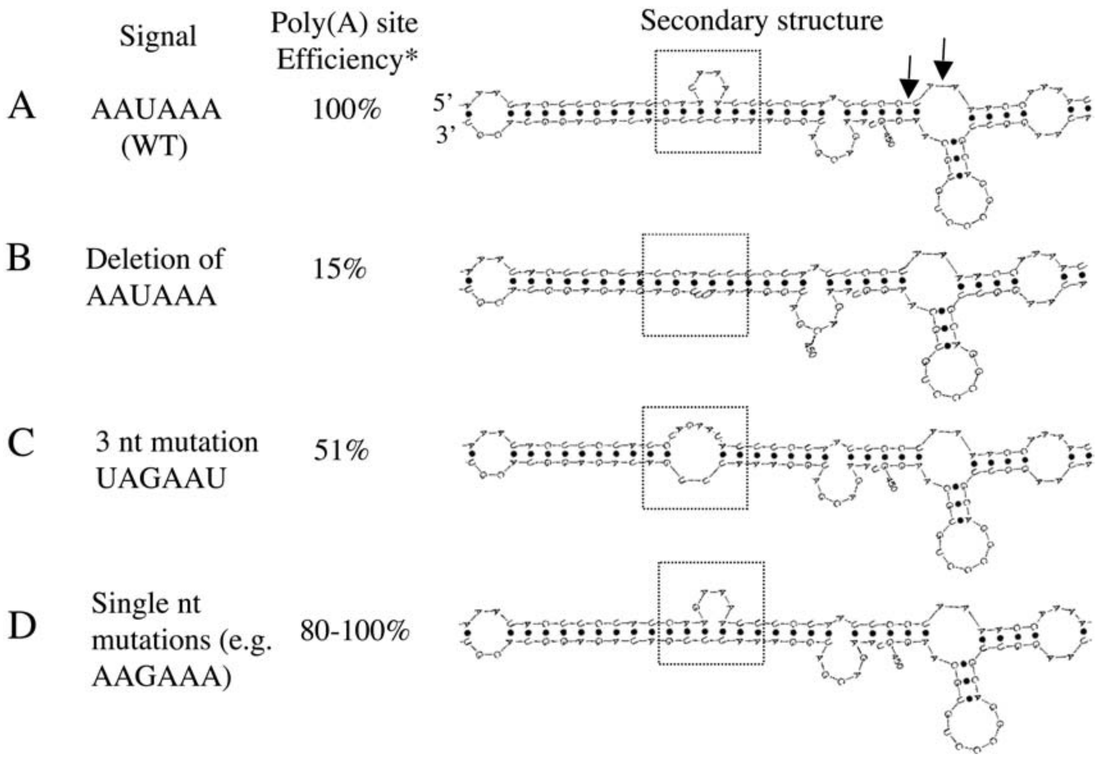


Figure 6. The relationship of CaMV 3′-UTR structure, mutation, and poly(A) signal efficiency. A, Wild type; its efficiency is defined as 100%. The arrows mark the CSs (Rothnie et al., 1994). The orientation of the sequence is as marked. B, Deletion of the NUE AAUAAA completely disrupts the secondary structure, which is also consistent with the reduction of efficiency of the signal. C, If 3 of the 6 nt altered, the structure changed from a small loop to a big one, and the signal efficiency was somewhat reduced. D, Some single-nucleotide mutations resulted in little structural changes, as well as signal efficiencies. *, The mutations and signal usage efficiencies were measured from the data presented by Rothnie et al. (1994) for A, B, and D and Mogen et al. (1990) for A and C, respectively. Free energy is −187 for all structures.

around the CS and poly(A) site. The signal surrounding the CS is named CE here. A set of prevailing, although not highly conserved, patterns that are potentially poly(A) signals for each of the three elements are presented. Conserved secondary structures surrounding the CSs were also predicted using the RNA secondary structure prediction program, mFold. Using data from the literature, it is confirmed that these structures are important for the functionality of the signals because only those mutations that altered secondary structures had impact on the efficiency of the signals. These findings should serve as a new starting point for plant poly(A) signal study, e.g. the basis for mutagenesis tests of CE, the design of a program to predict poly(A) sites for genome annotation purposes, and for finding alternative poly(A) sites.

A new working model for Arabidopsis mRNA poly(A) signals has emerged. As shown in Figure 3, the location of the FUE and the NUE has been updated based on this large-scale analysis, where the FUE region spans 60 to 125 nt, the NUE region 6 to 10 nt, but the CE is clearly expanded from the original CS (only 2 nt) to include two U-rich regions before and after the CS, both spanning about 5 to 10 nt. A closer view of the CS indicates a sharp nucleotide composition change where the U before the CS is highly desirable and a few Us also follow (Fig. 3B). Such a model could serve well in designing a computer algorithm to scan genomic sequences for possible poly(A) sites.

Conventional genetic analysis of plant poly(A) signals was not able to reveal the significance of the sequence elements surrounding the CS. This may be due partially to the signal element not being strong enough to be readily detected. The CE contains a long stretch of sequence to confirm its existence (although such a hypothesis is subject to further testing). It was postulated that there may be a U-rich region surrounding the YA dinucleotide (Hunt, 1994), but it was neither tested nor confirmed. A part of the CE, the sequence after the CS, was sometimes called the downstream element in the early literature. For example, a downstream element was found to affect the precision of cleavage, but did not influence the processing efficiency (Sanfacon et al., 1991). More dramatic effect was noted in the analysis of *ocs* and *rbc*S 3′-UTR, in which the deletion of the downstream element alters or eliminates the use of the poly(A) site (Hunt and MacDonald, 1989; MacDonald et al., 1991). These notions were not pursued further, and hence remained unresolved, but are revisited in this article.

The U-rich domain after the CE described here differs from the downstream elements found in animal systems, which are disrupted by about 15 nt of nonconserved sequences after the CS (Zhao et al., 1999). In Arabidopsis, it seems to be interrupted, only approximately 5-nt spacing right after the CS followed by a small U-rich element. To distinguish this, we designate it as a part of the CE. The U-rich sequence before the CS has been demonstrated by Graber et al. (1999b), with a few thousand ESTs from Arabidopsis. Although the U-rich sequence after the CS was suggested on their model, no evidence was presented (Graber et al., 1999b). Interestingly, the single-nucleotide profile around the CS of Arabidopsis 3′-UTRs is strikingly similar to that of yeast 3′-UTR, as reported (Graber et al., 1999a). This could be another indication of the similarity of the poly(A) signals between plants and yeast. Both UA and CA dinucleotides at the CS seem to be more prevalent in plants, while yeast seems to use UA much more than CA, according to the analysis of 1,352 unique genes (Graber et al., 1999a).

Comparing the FUE and NUE signal patterns we compiled to those characterized in pea *rbc*S, CaMV, figwort mosaic virus, rice tungro bacilliform virus (RTBV), *nos*, *ocs*, and maize 27-kD protein gene (Sanfacon and Hohn, 1990; Mogen et al., 1992; Wu et al., 1993; Hunt, 1994; Rothnie et al., 1994; Rothnie et al., 2001), we found that the signals from these genes possess high delta character and are ranked high among the top patterns in the Arabidopsis cDNA datasets of this study (data not shown). For example, AAUAAA, the NUE for CaMV, RTBV, and *nos*, is listed first of 50; AAUGAA, the NUE for *rbc*S E-9 and the maize 27-kD protein gene, is number 20. The same is true for the FUE signals, e.g. UUUGUA widely found in CaMV and RTBV, is number 14; and UUGUA, UUGUU, UGUGUA for *rbc*S and *ocs* are in the top 50. This information validates the compiled patterns.

Actual mutagenesis studies done in virus, yeast, and humans (Shen et al., 1999; Zarudnaya et al., 2003) indicate the presence of a higher order structure in 3′-UTR of mRNA, and its importance in the functionality of the mRNA. The analysis of the RNA structure by mFold has also demonstrated the presence of similar secondary structures flanking the CS (Figs. 4–6). The fidelity of such predicted secondary structures warrants further experimental testing, e.g. mutagenesis, and binding by poly(A)-related proteins. However, the validity of the mFold program has been proven, e.g. in a recent publication by Teixeira et al. (2004). The stem observed in both group I and group II is produced by A- and U-rich residues. Although, within these two groups, the secondary structures formed by the NUE and the CS signals vary, the range of the variation is reasonably within the predicted signal regions. The stem loop structure has been observed in relation to many poly(A) events, where the CSs are situated on the loop, and mutation of such by base substitution or deletion severed cleavage activities, leading to decreased biological activity as characterized in the IgM secretory transcript (Phillips et al., 1999). This phenomenon may be explained by loss of recognition and formation of the CPSF/CstF complex due to mutation or signal loss (Phillips et al., 1999). Many examples involving hairpins, internal loops, and globular circles represent target sites for RNA-interacting protein (Ruff et al., 1991). Similar events have also been observed in R2 RNA transcripts, where the stem loop found in the 3′-UTR of R2 was the prime target for reverse transcription complexes involved in targeting prime reverse transcription (Ruschak et al., 2004).

The deletion of these regions that contain relevant stem loops has been demonstrated to accompany the loss of poly(A) activity. This may be due to disruption of recognition of the higher order structures by protein factors. As mentioned in Zarudnaya et al. (2003), these loops collectively orchestrate the formation of a certain conformation grove for the trans-acting factors to recognize and bind. Disruption of either region, leading to a change in the structures, leads to changes in the poly(A) profile. Mutagenesis of CaMV mutants with single-nucleotide AAUAAA for the NUE mutation did not abolish NUE-processing efficiency, but almost lost its efficiency upon complete AAUAAA deletion (Rothnie et al., 1994). In a similar context, Hajarvanis and colleagues (2004) have proposed possible classes of 3′-ends that are recognized by specific regulatory factors that may direct different positioning of other factors. Secondary structure predictions provide a possible explanation for such phenomena; although nucleotides of NUEs have changed, it is crucial that the protein can still recognize the signal by conformation targeting. All mFold results indicate that NUE loops are observed in all mutants, except the deletion mutant, and hence trans-acting factors could no longer be targeted by structural recognition at the site where poly(A) signals are present. Our results should be the basis for further analysis of secondary structures on these and other genes.

MATERIALS AND METHODS

Compiling the 8-K Poly(A) Sites within Arabidopsis Genome Sequences

All Arabidopsis (*Arabidopsis thaliana*) transcript sequences, including ESTs and partial or complete cDNA sequences, were downloaded from GenBank on September 1, 2004. Using the trimpoly program, included in The Institute of Genomic Research (TIGR) Gene Indices seqclean software (http://www.tigr.org/tdb/tgi/software), transcripts containing terminal poly(A) sequences were identified and trimmed. The terminal transcript nucleotide of each trimmed polyadenylated transcript was classified as a poly(A) site. Since the trimpoly tool trims low-quality regions from transcript sequence ends in addition to poly(A) sequences, our analysis included only those trimmed poly(A) site transcript ends identified by trimpoly, which were followed by a stretch of 8 to 15 nt with at least 80% adenine content. This criteria proved sufficient to differentiate the presumed genuinely polyadenylated sequences from those of low-quality sequence ends, disregarding other sequences trimmed by trimpoly due to low sequence quality rather than based on terminal poly(A) content. Checking the set of poly(A) sites identified in the genome, and limiting a sequence composition analysis to the 8 bp beyond the CS, there are a maximum of 6.8% of the sequences that could be falsely

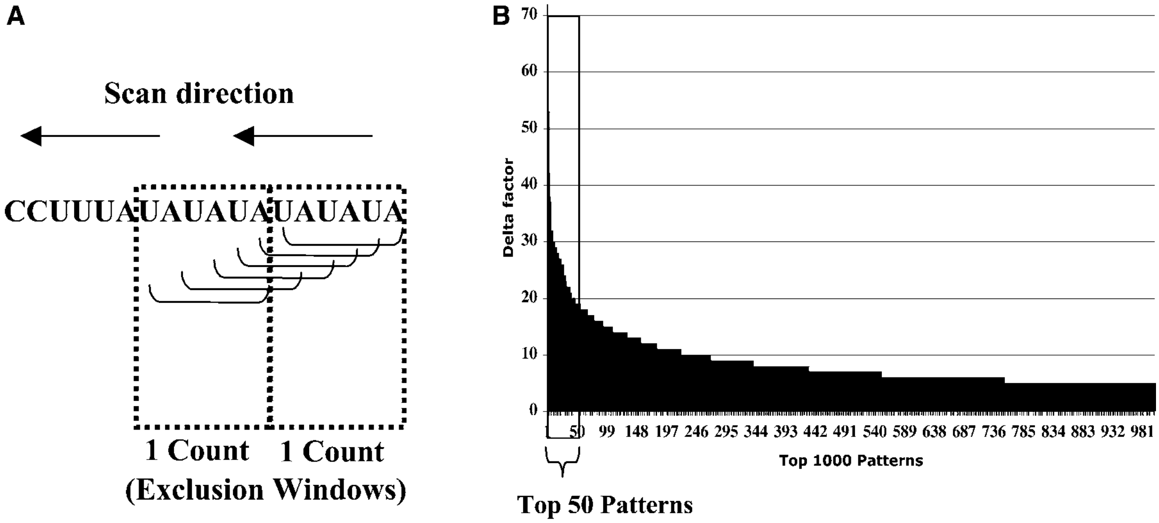


Figure 7. Scanning methods and selection of the top 50 patterns. A, Scanning method. Multiple-count algorithm will result in overcounting for a particular repetitive signal, resulting in overrepresentation, e.g. pattern UAUAUA total count is 6, as indicated by the solid half-frames. An algorithm with an exclusion window allows for only those frames that fall within the scan window (dotted boxed lines) of a particular signal length to be counted once, but allows counts of the recurring same pattern throughout each of the sequences (total count of 2). B, Top 50 patterns of 6-nt pattern size for the NUE of the 8-K dataset shows a more significant difference in the delta factor among the top 1,000 patterns.

classified by using these criteria alone. Using the 15-bp maximal window size, we found the maximal false-positive rate drops to 5.6%.

From the sequences of 191,301 Arabidopsis ESTs and 35,557 mRNA sequences obtained from GenBank, we found 10,735 sequences containing poly(A) sequences, which align almost perfectly to the Arabidopsis genome. Approximately one-half of the polyadenylated sequences were derived from full-length cDNAs. The final assembly of 9,298 ESTs was further filtered through methods described by Beaudoing and colleagues (2000) in which sequences containing stretches of As within 10 bp after the CS may denote internal priming contamination. The genome alignments of the trimmed poly(A)-containing sequences provide the identity of 8,160 poly(A) sites within all five chromosomes. The genomic sequence position corresponding to the poly(A) site of each relevant transcript sequence was identified via sequence alignment. The Program to Assemble Spliced Alignments (PASA) pipeline (Haas et al., 2003) was used to align the trimmed transcript sequences to the Arabidopsis Release 5.0 chromosome pseudomolecule sequences (available at ftp://ftp.tigr.org/pub/data/a_thaliana). A total of 8,160 sequences 300 nt upstream of the poly(A) sites (including assumed 3′-UTRs) were collected as this 8-K dataset. The coordinates of each poly(A) site based on the genome coordinate reference were calculated based on the corresponding transcript alignment coordinates. To include the study of downstream sequences, a program was developed to extract sequence regions (100 nt downstream from the CSs) from the chromosome sequences based on the poly(A) site and annotated gene coordinates. FASTA formatted sequence files (as cDNA) were created to serve as input for the remaining sequence analysis.

Compiling the 16-K Arabidopsis 3′-UTR Sequences

The 16-K dataset consists of the 3′-UTR terminal 300 nt, from the assembled 16,211 Arabidopsis full-length cDNAs as described (Haas et al., 2003). A total of 177,973 Arabidopsis ESTs, 27,414 cDNAs, and 3,217 partial cDNAs, were examined in building this cDNA database, using PASA, which integrates any unaligned sequences into at least one of the maximal assembly. All sequences were selected for quality regions and poly(A) tail by using a SeqClean tool, and realigned defined sequence regions, using the BLAST-like Alignment Tool (BLAT) against the complete genomic sequences.

Comparing the two datasets used here, one containing 8,160 ESTs (8 K) with authenticated poly(A) sites and the other with 16,211 full-length cDNAs (16 K), the 8-K dataset contains 584 EST sequences that are not found in the 16-K dataset totaling 442 unique genes in the 8-K dataset. There are also 10,474 genes that are unique to the 16-K data set and 5,737 genes are common in both datasets. Thus, the combined total number of genes being analyzed is about 17,000. Both the 8-K and 16-K datasets are available at http://www.users.muohio.edu/liq.

The Pattern Compilation Program

A program, SignalSleuth, was created to perform an exhaustive search of varying size patterns within a subregion of a large set of sequences. (The code can be downloaded at http://www.users.muohio.edu/liq.) The program was developed, installed, and run on Cray computers located at both Cray facilities in Wisconsin and the Ohio Supercomputer Center (OSC). With the use of the Cray Bioinformatics Library (CBL; Cray, Inc., 2004), the program was quickly implemented and achieved unprecedented performance by accessing special bit manipulation hardware instructions on the machine.

The algorithm used in the program starts out by reading the sequence data from a FASTA file using the CBL routine cb_read_fasta. The program then enters a triply nested loop, looping over pattern size, the number of sequences, and the location within a given sequence. After entering the outermost loop, pattern size, the program allocates enough memory to hold all possible combinations of the four unique characters {A, C, G, T} in *n* locations, where *n* is the size of the pattern for this trip through the loop. The program then begins at the starting location for the first nucleotide, in the subregion of interest, within the first sequence. The program copies the first pattern length worth of characters from the sequence into a temporary variable and compresses it, using the CBL routine cb_compress, into a 2-bit compress form by picking out the second and third bits from each character in the variable. Since there are only four possible characters, only 2 bits of information are needed {00, 01, 10, and 11} to store this information. Shifting the bits in this 2-bit compress variable to the rightmost bits of the pattern, the variable can then be used as an integer to index into the all-possible combination array and increment that location. With this location now tallied, the code shifts to the right one character in the input sequence and repeats the process. When all the characters within the subregion for this sequence are processed, the code advances to the next sequence and repeats the process for the subregion in the next sequence. The program continues in this fashion until all sequences have been processed.

At the end of this search process, the all-possible combination arrays now contain a histogram of how frequently each combination was found within the target regions of all the sequences. The next step is to search this array to find the largest number, or set of largest numbers, such as the top 50 most common patterns in the target regions of the sequences. The program then converts the indices into this array from its 2-bit compressed form back into its full 8-bit ASCII characters, and the characters associated with the index are printed out. This scanning algorithm took on several different variations, based on user-defined parameters. With these parameters, the count of a particular pattern can be counted once or multiple times per sequence, and if multiple counts are allowed, a gap can be defined regarding when to start counting the particular pattern again. This helps to prevent short repeated patterns from being overly represented. If the single count option is used, the count of a particular pattern is only counted once per sequence and may result in an underrepresented count of a given pattern. For example, if a tract of UAUAUAUAUAUA were encountered for a 6-nt window size pattern, each frame of the UAUAUA will be counted as the same pattern, resulting in an overrepresented count for this particular pattern. This algorithm will allow the signal for a particular repeated pattern to be counted again on the same sequence only if it falls outside a particular exclusion window size on the sequence (Fig. 7A). These repetitive patterns can be observed in Figure 1, described in the following section.

The ranking of the counts in the array of patterns of all-possible combinations is based on a deviation factor from the median value, which is termed delta, and is defined as the difference between the maximal count and the median count of a respective pattern. Pattern counts deviating the farthest from the median are ranked the highest. The selection of the top 50 is justified by the reduction of this deviation among the top 1,000 signals because this deviation drops sharply after the first 50 patterns, as seen in Figure 7B.

Pattern Location Images

After the most common patterns were found, the next task was to illustrate where these common patterns fall within the sequences, and to see if they were more common, for example, in the NUE region as opposed to the rest of the sequence. To accomplish this, additional code was added to the program to form a graphic picture of the locations of the patterns within each sequence. Imagine a picture that is 8,160 pixels tall and 400 pixels wide, where a pixel is a dot on the screen or printed on a page. In this picture each pixel represents the starting location of one of the top 50 patterns within the sequence. Referring to Figure 1C, the program was run to search the NUE region for the top 50 most common patterns. With this list, the program then turned on the pixel that corresponds to those patterns as they are located in each sequence. Notice that, even though these patterns can be found throughout all the sequences, they are clearly more common in the NUE region, as marked at the top of the image. Similarly, the program was run to search the region near the CS, and its top 50 most common patterns plotted in Figure 1D. Again, distinct vertical bands can be seen on either side of the CS. Figure 1E is used as a control to show what an image would look like using random data. For these images, no exclusion window was used, so long repeats of short patterns can be seen as small horizontal bars in the images.

Secondary Structure of RNA

The predictions of secondary structure of the RNA region surrounding poly(A) sites were carried out by an RNA secondary structure prediction program, mFold (Zuker, 2003; http://www.bioinfo.rpi.edu/applications/mfold). All analyses were deployed at 25°C, 0.1 M Na^+, 0.002 M Mg^{2+}, and 5% suboptimality number with a maximum of 50 upper bounds on the number of computed folding. Only the top five of all given outputs were selected as justified by the most favorable free-energy conditions.

ACKNOWLEDGMENTS

We thank Mouin Hourani, Mingjun Tang, and Minghui Ding for technical assistance, and Michael Hughes for helping with statistical analysis. We also thank Kim Davis for helping with mFold analysis, and Paula Strenski for her invaluable assistance in editing and reviewing this manuscript. We are grateful for the use of Cray's SV1ex supercomputer, located in Chippewa Falls, Wisconsin, and the Ohio Supercomputer Center's SV1 located in Columbus, Ohio.

Received January 31, 2005; revised April 29, 2005; accepted May 3, 2005; published June 17, 2005.

LITERATURE CITED

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

Beaudoing E, Freier S, Wyatt JR, Claverie JM, Gautheret D (2000) Patterns of variant polyadenylation signal usage in human genes. Genome Res **10:** 1001–1010

Cray, Inc. (2004) Man Page Collection: Bioinformatics Library Procedures. http://www.cray.com/craydoc/manuals/S-2397-21/S-2397-21.pdf

Graber JH, Cantor CR, Mohr SC, Smith TF (1999a) Genomic detection of new yeast pre-mRNA 3′-end-processing signals. Nucleic Acids Res **27:** 888–894

Graber JH, Cantor CR, Mohr SC, Smith TF (1999b) In silico detection of control signals: mRNA 3′-end-processing sequences in diverse species. Proc Natl Acad Sci USA **96:** 14055–14060

Graber JH, McAllister GD, Smith TF (2002) Probabilistic prediction of *Saccharomyces cerevisiae* mRNA 3′-processing sites. Nucleic Acids Res **30:** 1851–1858

Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK Jr, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, et al (2003) Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res **31:** 5654–5666

Hajarnavis A, Korf I, Durbin R (2004) A probabilistic model of 3′ end formation in *Caenorhabditis elegans.* Nucleic Acids Res **32:** 3392–3399

Hunt AG (1994) Messenger RNA 3′ end formation in plants. Annu Rev Plant Physiol Plant Mol Biol **45:** 47–60

Hunt AG, MacDonald MH (1989) Deletion analysis of the polyadenylation signal of a pea ribulose-1,5-bisphosphate carboxylase small-subunit gene. Plant Mol Biol **13:** 125–138

Hunt AG, Messing J (1998) mRNA Polyadenylation in Plants. *In* J Bailey-Serres, DR Gallie, eds, A Look beyond Transcription Mechanisms Determining mRNA Stability and Translation in Plants. American Society of Plant Physiologists, Rockville, MD, pp 29–39

Li QQ, Hunt AG (1995) A near upstream element in a plant polyadenylation signal consists of more than six bases. Plant Mol Biol **28:** 927–934

Li QQ, Hunt AG (1997) The polyadenylation of RNA in plants. Plant Physiol **115:** 321–325

MacDonald CC, Redondo JL (2002) Reexamining the polyadenylation signal: Were we wrong about AAUAAA? Mol Cell Endocrinol **190:** 1–8

MacDonald MH, Mogen BD, Hunt AG (1991) Characterization of the polyadenylation signal from the T-DNA-encoded octopine synthase gene. Nucleic Acids Res **19:** 5575–5581

Mogen BD, MacDonald MH, Graybosch R, Hunt AG (1990) Upstream sequences other than AAUAAA are required for efficient messenger RNA 3′-end formation in plants. Plant Cell **2:** 1261–1272

Mogen BD, MacDonald MH, Leggewie G, Hunt AG (1992) Several distinct types of sequence elements are required for efficient mRNA 3′ end formation in a pea rbcS gene. Mol Cell Biol **12:** 5406–5414

Phillips C, Kyriakopoulou CB, Virtanen A (1999) Identification of a stem-loop structure important for polyadenylation at the murine IgM secretory poly(A) site. Nucleic Acids Res **27:** 429–438

Proudfoot N (2004) New perspectives on connecting messenger RNA 3′ end formation to transcription. Curr Opin Cell Biol **16:** 272–278

Proudfoot NJ, Furger A, Dye MJ (2002) Integrating mRNA processing with transcription. Cell **108:** 501–512

Rothnie HM (1996) Plant mRNA 3′-end formation. Plant Mol Biol **32:** 43–61

Rothnie HM, Chen G, Futterer J, Hohn T (2001) Polyadenylation in rice tungro bacilliform virus: cis-acting signals and regulation. J Virol **75:** 4184–4194

Rothnie HM, Reid J, Hohn T (1994) The contribution of AAUAAA and the upstream element UUUGUA to the efficiency of mRNA 3′-end formation in plants. EMBO J **13:** 2200–2210

Ruff M, Krishnaswamy S, Boeglin M, Poterszman A, Mitschler A, Podjarny A, Rees B, Thierry JC, Moras D (1991) Class II aminoacyl transfer RNA synthetases: crystal structure of yeast aspartyl-tRNA synthetase complexed with tRNA(Asp). Science **252:** 1682–1689

Ruschak AM, Mathews DH, Bibillo A, Spinelli SL, Childs JL, Eickbush TH, Turner DH (2004) Secondary structure models of the 3′ untranslated regions of diverse R2 RNAs. RNA **10:** 978–987

Sanfacon H, Brodmann P, Hohn T (1991) A dissection of the cauliflower mosaic virus polyadenylation signal. Genes Dev **5:** 141–149

Sanfacon H, Hohn T (1990) Proximity to the promoter inhibits recognition of cauliflower mosaic virus polyadenylation signal. Nature **346:** 81–84

Shen LX, Basilion JP, Stanton VP Jr (1999) Single-nucleotide polymorphisms can cause different structural folds of mRNA. Proc Natl Acad Sci USA **96:** 7871–7876

Teixeira A, Tahiri-Alaoui A, West S, Thomas B, Ramadass A, Martianov I, Dye M, James W, Proudfoot NJ, Akoulitchev A (2004) Autocatalytic RNA cleavage in the human beta-globin pre-mRNA promotes transcription termination. Nature **432:** 526–530

Van Helden J, Olmo M, Perez-Ortin JE (2000) Statistical analysis of yeast genomic downstream sequences reveals putative polyadenylation signals. Nucleic Acids Res **28:** 1000–1010

Wu L, Ueda T, Messing J (1993) 3′-end processing of the maize 27 kDa zein mRNA. Plant J **4:** 535–544

Zarudnaya MI, Kolomiets IM, Potyahaylo AL, Hovorun DM (2003) Downstream elements of mammalian pre-mRNA polyadenylation signals: primary, secondary and higher-order structures. Nucleic Acids Res **31:** 1375–1386

Zhao J, Hyman L, Moore C (1999) Formation of mRNA 3′ ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis. Microbiol Mol Biol Rev **63:** 405–445

Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res **31:** 3406–3415