Available online at www.sciencedirect.com
ScienceDirect
Gene 389 (2007) 107–113
GENE
SECTION
FUNCTIONAL GENOMICS
www.elsevier.com/locate/gene

An exploration of 3′-end processing signals and their tissue distribution in *Oryza sativa*

Haitao Dong [a,*,1], Ye Deng [a,1], Jack Chen [b], Sheng Wang [a], Sihua Peng [c], Cheng'en Dai [a], Yongqi Fang [a], Jing Shao [a], Yichun Lou [a], Debao Li [a,*]

[a] *Bioinformatics and Gene Network Research Group, College of Agriculture and Biotechnology, Zhejiang University, Hangzhou 310029, China*
[b] *Biocomputation Unit, National Institute of Child Health and Human Development (NICHD), National Institute of Health (NIH), Bethesda 20892, USA*
[c] *College of Mechanical and Energy Engineering, Zhejiang University, Hangzhou 310027, China*

Received 5 April 2006; received in revised form 20 July 2006; accepted 13 October 2006
Available online 10 November 2006

Abstract

The 3′ untranslated regions deeply affect many properties of eukaryotic mRNA. In plants, the polyadenine control signals contained in these regions seem to be more variable than of mammals. Three cDNA libraries derived from the leaf, endosperm and stem tissues of rice were sequenced from the 3′-end. Of the 9911 transcripts analyzed, 5723 unique transcripts were identified from the leaf sequences, 2934 from the endosperm and 1254 from the stem. The information entropy and two statistical methods were used to compile a list of rice poly(A) control signals. Based on their distribution, these signals can be roughly grouped into far-upstream element (FUE), near-upstream element (NUE), T-rich region (TRE) and downstream element (DE). The distribution of rice conserved regions is similar to the previous model from *Arabidopsis* and yeast, with a few differences in word constructions. Interestingly, we also found the word distributions were diverse in the cleavage site of downstream sequences of different rice tissues. The signal bias in downstream sequences may lead mRNA to be differently cleaved in different rice tissues.
© 2006 Elsevier B.V. All rights reserved.

Keywords: 3′-UTR; Rice; Downstream sequences; Cleavage site

1. Introduction

Most eukaryotic mRNAs contain long 3′ untranslated regions (UTR) spanning hundreds of nucleotides in front of polyadenine [poly(A)] tails. These regions affect mRNA stability (Dreyfus and Regnier, 2002), assist in translation (Knirsch and Clerch, 2000), determine intracellular localization (Kislauskis et al., 1994) and control polyadenylation (Mazumder et al., 2003). In mammals, the control region of the poly(A) signal in 3′-end-processing sequences has been widely investigated and two components are found: 1) the canonical AATAAA hexamer, located 10 to 30 bases upstream of the cleavage site (CS) in 90% of all sequenced polyadenylation elements, and 2) a T- or GT-rich element 20 to 40 bases downstream of the CS in approximately 70% of mammalian pre-mRNAs (Colgan and Manley, 1997; Zhao et al., 1999). In yeast and plants, a predictive model of the 3′-end-processing signal is more elusive because there is no universally conserved sequence (word) such as the AATAAA hexamer. The hypothetical 3′-end-processing signal in plants consists of a far-upstream element (FUE), one or more A-rich regions known as near-upstream elements (NUE), and U-rich regions located upstream of, downstream of, or flanking a CS (Graber et al., 1999; Rothnie, 1996). Recently, bioinformatic analysis identified cleavage elements, a new set of poly(A) signals in *Arabidopsis* that flank both sides of the CS (Loke et al., 2005).

Abbreviations: UTR, untranslated region; poly(A), polyadenine; CS, cleavage site; FUE, far-upstream element; NUE, near-upstream elements; EST, expressed sequence tag; TRE, T-rich elements; CE, cleavage element.

* Corresponding authors. Tel.: +86 571 86971183; fax: +86 571 86961525.
E-mail addresses: htdong@zjuem.zju.edu.cn (H. Dong), lidb@mail.hz.zj.cn (D. Li).
URLs: http://www.estarray.org (H. Dong), http://www.estarray.org (D. Li).
[1] These authors contributed equally to this work.

0378-1119/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2006.10.015

The genomic sequences of the two main subspecies of domesticated rice (*Oryza sativa*), *indica* cv. 9311 and *japonica* cv. Nipponbare, are now available (Goff et al., 2002; Yu et al., 2002). However, gene identifications and transcript catalogs covering polyadenylation are still far from complete. Because the control signals of polyadenylation in plants are still unclear, polyadenylic sites cannot be predicted from the genomic sequence alone. *In silico* experiments combining genome sequences with reliable expressed sequence tags (ESTs) from the 3′-end of cDNA clones could help to identify a general structure of polyadenylation control regions (Beaudoing and Gautheret, 2001). Here, a large non-redundant 3′ EST dataset was generated from three rice cDNA libraries and subjected to a statistical analysis to investigate the 3′-end sequence features of rice.

2. Materials and methods

2.1. Preparation of 3′-end processing sequence datasets

Three cDNA libraries from the leaf, endosperm and stem tissues of rice were sequenced from the 3′-end with a MegaBASE™ 1000 automated sequencer (MD). Details can be found at www.estarray.org.

Base calling and quality control of the EST sequences were performed with the Phred program (Ewing et al., 1998). Of 28,324 ESTs, only those sequences containing poly adenines (A) and a part of vector sequences in 3′-end were analyzed. After trimming vectors and poly(A) tails, 24,845 ESTs (13,293 from leaf, 9070 from endosperm and 2482 from stem cDNA libraries) had more than 100 nucleotides (nt). Only the last 100 nt of each 3′ EST were used for alignment with genomic sequences of *O. sativa* ssp. *indica* cv. 9311 (downloaded from http://rice.genomics.org.cn/rice2/link/download.jsp) with the MegaBlast program (Zhang et al., 2000). A strict threshold was set with the following parameters: percentage of identity, >95%; alignment length, >95 nt; gap openings and mismatches, <5 nt; and the end of the aligned query sequence (q. end) had to be the 100th nucleotide. ESTs that aligned genomic sequences in the same position were regarded as the same transcript. The search identified 5723, 2934 and 1254 unique transcripts in the leaf, endosperm and stem libraries, respectively. The 150 nucleotides upstream and 100 nucleotides downstream of the cleavage sites were used for further word detection. These data can be downloaded at http://www.estarray.org/est/download.asp.

2.2. Entropy of the sequence segments of the sliding window

Given a probability distribution $P=(p_1,p_2..p_n)$, its entropy $H(p)$ could be defined as: $H(p) = E(-\log p) = -\sum_{i=1}^{n}(p_i * \log p_i)$, where E is the expectation (Shannon, 1948).

A sliding window was employed to intercept the segment in all of the tested sequences. For example, if the length of the window was 10 nt, 241 segments could be obtained along the 250 nt in each sequence. For a segment j, we could compute its probability distribution as $p(j)=[P_A(j), P_T(j), P_C(j), P_G(j)]$. Here $P_A(j) = \frac{\mathrm{Sum}_A(j)}{\mathrm{Sum}_{A,T,C,G}(j)}$, where $\mathrm{Sum}_A(j)$ is the sum of base A in the segment j of all tested sequences, and $\mathrm{Sum}_{A,T,C,G}(j)$ is the total number of A, T, C and G in this segment j. Therefore, the entropy of sequence segment j is:

$$H(j) = -[p_A(j)*\log p_A(j) + p_T(j)*\log p_T(j) + p_C(j)*\log p_C(j) + p_G(j)*\log p_G(j)]$$

Two was used as the fundus of the logarithm

2.3. Detection of putative cis-elements

Initially, significant word detection was performed through R'MES, a set of programs used to find words that occur in a given DNA sequence with an unexpected frequency based on Markov chain models (Reinert et al., 2000; Schbath, 2000). As suggested by van Helden et al. (2000), 3 was chosen as the Markov chain order. Because a higher Z-score in the program results would lead to the inclusion of words with a low count (Graber et al., 1999), only words with a Z-score more than 3.49, which corresponds to a first-order risk $\alpha = 1/4^6$, and a raw count frequency more than 3 times the random frequency (*e.g.* the random frequency of a 6 nt word is $1/4^6 = 2.44e^{-4}$) were regarded as an unexpected frequency element.

Then, χ^2 test was performed to analyze positional bias. Every sequence was divided into 25 classes (from 1 to 250, with a class interval of 10). For each word, the number of occurrences in each class was compared to the expected number of occurrences, calculated on the basis of a flat distribution hypothesis (van Helden et al., 2000). Different χ^2 values were chosen based on the length of the word, such as 50 for a 6 nt word. The Perl script for the entire process can be downloaded at www.estarray.org. Finally, Cluster3.0 software (Eisen et al., 1998) was used to cluster significant words.

3. Results

3.1. General statistics of EST datasets of three cDNA libraries

Of a total 28,324 ESTs sequenced by ourselves (Genbank accession No.: ***BI794850–BI804839, BI807369–BI813841, BM037978–BM039065, BM418528–BM422313, BQ905903–BQ909262***), 24,845 ESTs with poly(A) tails were found in the leaf (13,293), endosperm (9070) and stem (2482) cDNA

Table 1
Sequence processing data

	Leaf dataset	Stem dataset	Endosperm dataset	General dataset
The number of EST	15,396	2683	10,245	28,324
The number of poly(A) EST	13,293	2482	9070	24,845
Poly(A) EST percent (%)	86.3	92.5	88.5	87.7
Average length (nt)	488	522	462	482
Genome mapped ESTs	8511	1380	5677	15,568
Genome mapped unique transcripts	5723	1254	2934	9911

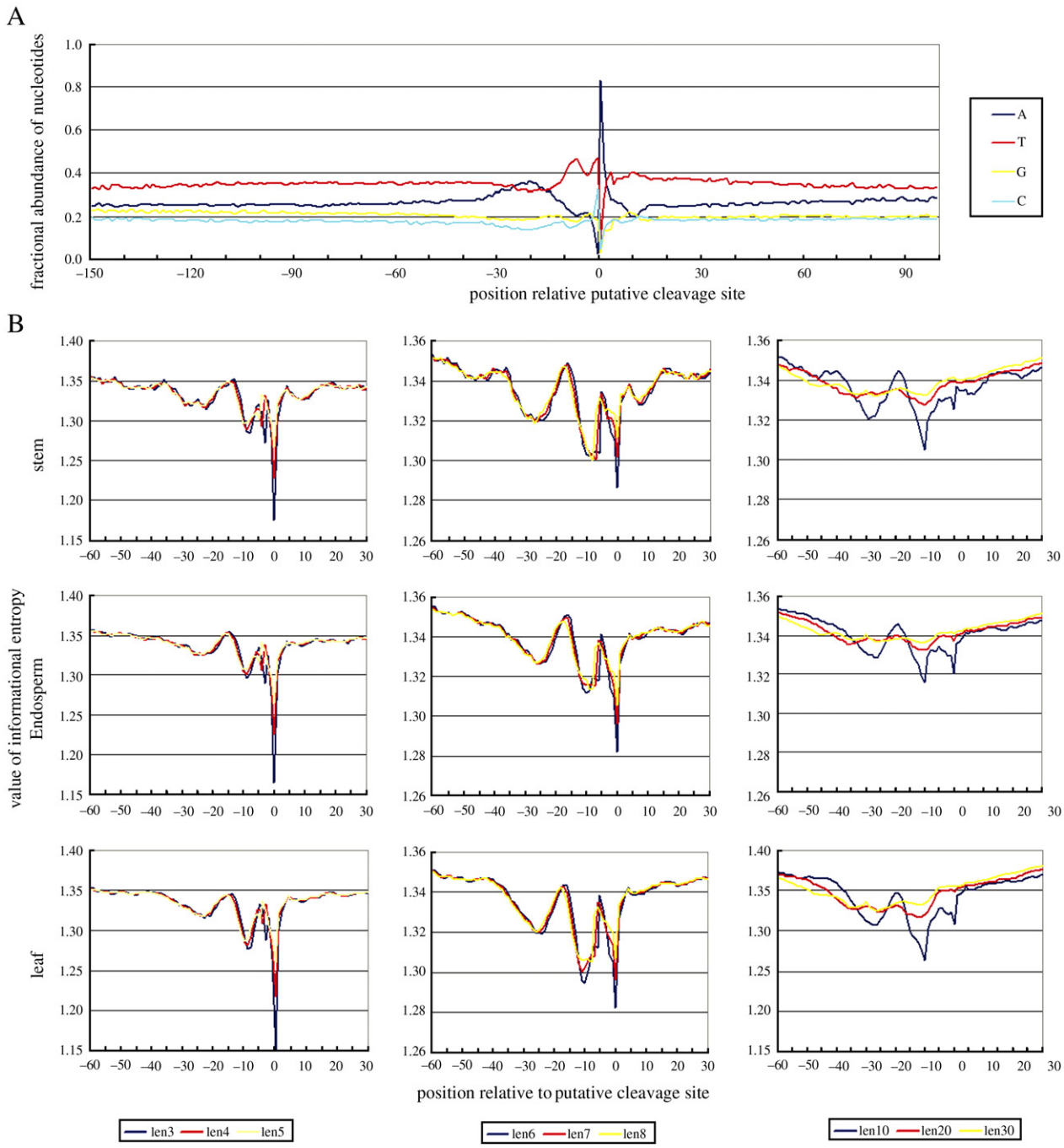


Fig. 1. Distribution of nucleotide and information entropy. A, The nucleotide distribution between 150 nt upstream and 100 nt downstream of CS in all sequences. B, Information entropy of stem, endosperm and leaf sequences from 60 nt upstream to 30 nt downstream with a series of window lengths (*e.g.* Len3 is a window length of 3 nt).

libraries of rice. After mapped to the genomic sequences of *O. sativa* ssp. *indica* cv. 9311 with a very strict threshold, total 9911 cleavage sites (CS) in rice genomic sequences were used for further analysis. A summary of these results is shown in Table 1.

3.2. Analysis of cis-element distribution and length

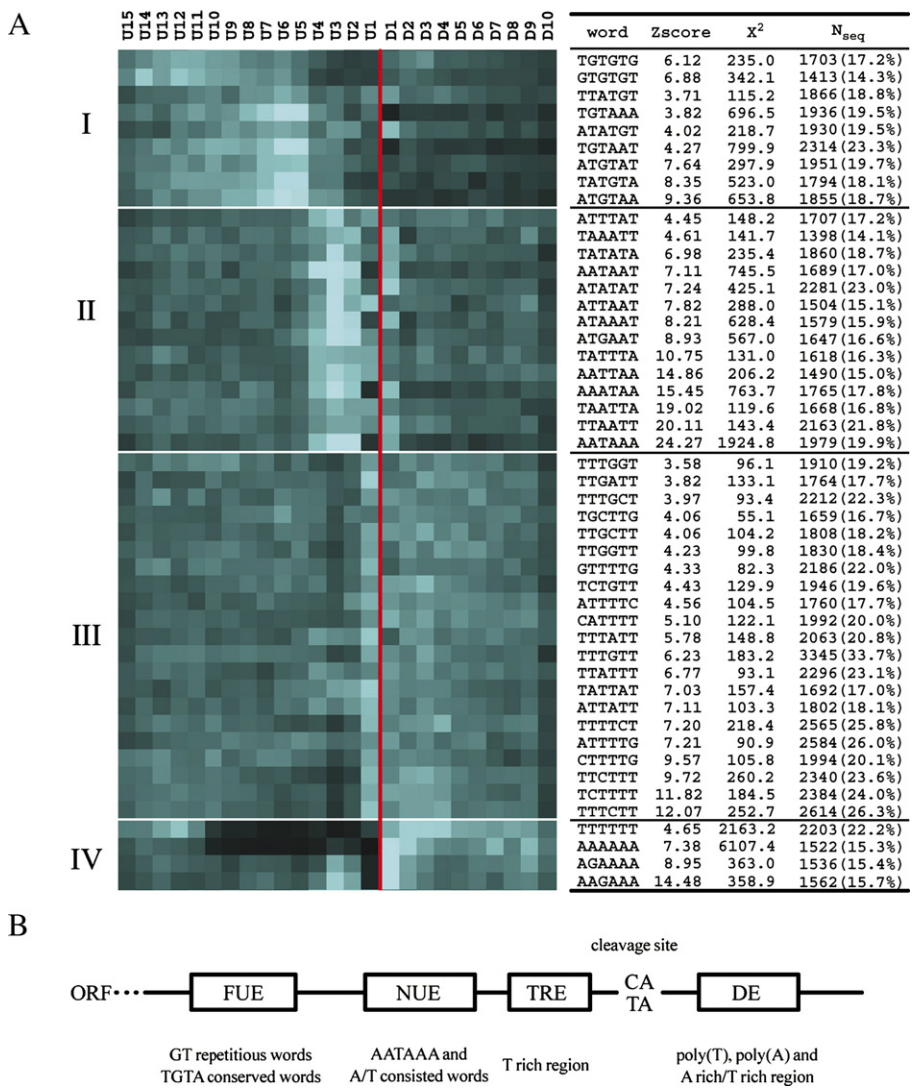
The nucleotide distribution in the interval between 150 nt upstream and 100 nt downstream of the CS is shown in Fig. 1A. It was a region with a relatively low GC content (40% on average). Significant fluctuation appeared in the positions from −35 to +15, indicative of the presence of centralized *cis*-elements. Notably higher C and T contents at position −1 to −2, and a high abundance of adenines at positions 0 to +1 (80%), were characteristic of the conservative composition of the poly (A) site (CA/TA, 79.4%).

A series of sliding windows of different lengths were used to show the segment entropy variation within the sequences (Fig. 1B). After increasing the width of the windows, the line of entropy became flatter, signifying that the continuous

conservation of nucleotides was no more than 10 nt. In general, there were three peaks in the region between −60 to 30 of the CS, which may represent the existence of a specific sequence composition. At position 0, the highest peak represented the conservation of sequence at the poly(A) site. In this position, within a window of 10 nt, the trend of the entropy variation remained unchanged, however the values changed greatly. The smallest value was 3 nt in width, meaning that the longest possible length of conserved sequences is less than 3 nt. The peak at position −10 represented the value from a T-rich region located at about −10 nt upstream of the poly(A) site. The third peak at −30 nt upstream of poly(A) site, characterized an A-rich region, which was similar to an NUE. At the −10 and −30 positions, the entropy values displayed only minor changes within 10 nt, and the width of region scales increased. However, a peak analogous to the FUE was not found, implying a higher degree of polymorphism in element composition and position. Interestingly, only the entropy curves of the stem library had a fourth peak at downstream position 8.

3.3. Detection of potential elements

The Markov chain model and χ^2 test were used to measure the over-representation and position biases of words. These significant words were screened as putative *cis*-elements. Twenty-five classes with 10-nt intervals were used to determine the distribution of significant words and to perform *K*-mean ($K=4$) clustering (Fig. 2A). The words in cluster I were putative FUEs, which mainly distributed 40 nt upstream of the CS. Two repetitive GT words were located approximately 100 nt upstream, while the remaining seven appeared 50–70 nt upstream and had a core motif of TGTA. The words in cluster



word	Zscore	X^2	N_{seq}
TGTGTG	6.12	235.0	1703(17.2%)
GTGTGT	6.88	342.1	1413(14.3%)
TTATGT	3.71	115.2	1866(18.8%)
TGTAAA	3.82	696.5	1936(19.5%)
ATATGT	4.02	218.7	1930(19.5%)
TGTAAT	4.27	799.9	2314(23.3%)
ATGTAT	7.64	297.9	1951(19.7%)
TATGTA	8.35	523.0	1794(18.1%)
ATGTAA	9.36	653.8	1855(18.7%)
ATTTAT	4.45	148.2	1707(17.2%)
TAAATT	4.61	141.7	1398(14.1%)
TATATA	6.98	235.4	1860(18.7%)
AATAAT	7.11	745.5	1689(17.0%)
ATATAT	7.24	425.1	2281(23.0%)
ATTAAT	7.82	288.0	1504(15.1%)
ATAAAT	8.21	628.4	1579(15.9%)
ATGAAT	8.93	567.0	1647(16.6%)
TATTTA	10.75	131.0	1618(16.3%)
AATTAA	14.86	206.2	1490(15.0%)
AAATAA	15.45	763.7	1765(17.8%)
TAATTA	19.02	119.6	1668(16.8%)
TTAATT	20.11	143.4	2163(21.8%)
AATAAA	24.27	1924.8	1979(19.9%)
TTTGGT	3.58	96.1	1910(19.2%)
TTGATT	3.82	133.1	1764(17.7%)
TTTGCT	3.97	93.4	2212(22.3%)
TGCTTG	4.06	55.1	1659(16.7%)
TTGCTT	4.06	104.2	1808(18.2%)
TTGGTT	4.23	99.8	1830(18.4%)
GTTTTG	4.33	82.3	2186(22.0%)
TCTGTT	4.43	129.9	1946(19.6%)
ATTTTC	4.56	104.5	1760(17.7%)
CATTTT	5.10	122.1	1992(20.0%)
TTTATT	5.78	148.8	2063(20.8%)
TTTGTT	6.23	183.2	3345(33.7%)
TTATTT	6.77	93.1	2296(23.1%)
TATTAT	7.03	157.4	1692(17.0%)
ATTATT	7.11	103.3	1802(18.1%)
TTTTCT	7.20	218.4	2565(25.8%)
ATTTTG	7.21	90.9	2584(26.0%)
CTTTTG	9.57	105.8	1994(20.1%)
TTCTTT	9.72	260.2	2340(23.6%)
TCTTTT	11.82	184.5	2384(24.0%)
TTTCTT	12.07	252.7	2614(26.3%)
TTTTTT	4.65	2163.2	2203(22.2%)
AAAAAA	7.38	6107.4	1522(15.3%)
AGAAAA	8.95	363.0	1536(15.4%)
AAGAAA	14.48	358.9	1562(15.7%)

Fig. 2. Distribution of conserved regions in rice sequences. A, The distribution of significant words was performed by *K*-mean ($K=4$) clustering. Significant words were generated through the scanning methods of Markov chain model and the χ^2 test. Sequences 150 nt upstream and 100 nt downstream of the CS were divided into 25 classes with 10-nt intervals (*e.g.* U1 means the 1st class from 1 to 10 nt of upstream sequences.). N_{seq} is the number of sequences in which the word could be found. N_{seq} divided by the number of total sequences (9911) with the percentage of the total in parentheses. B, The model of for rice mRNA poly(A) signals. FUE, far-upstream element; NUE, near-upstream element; TRE, T-rich region; DE, downstream element.

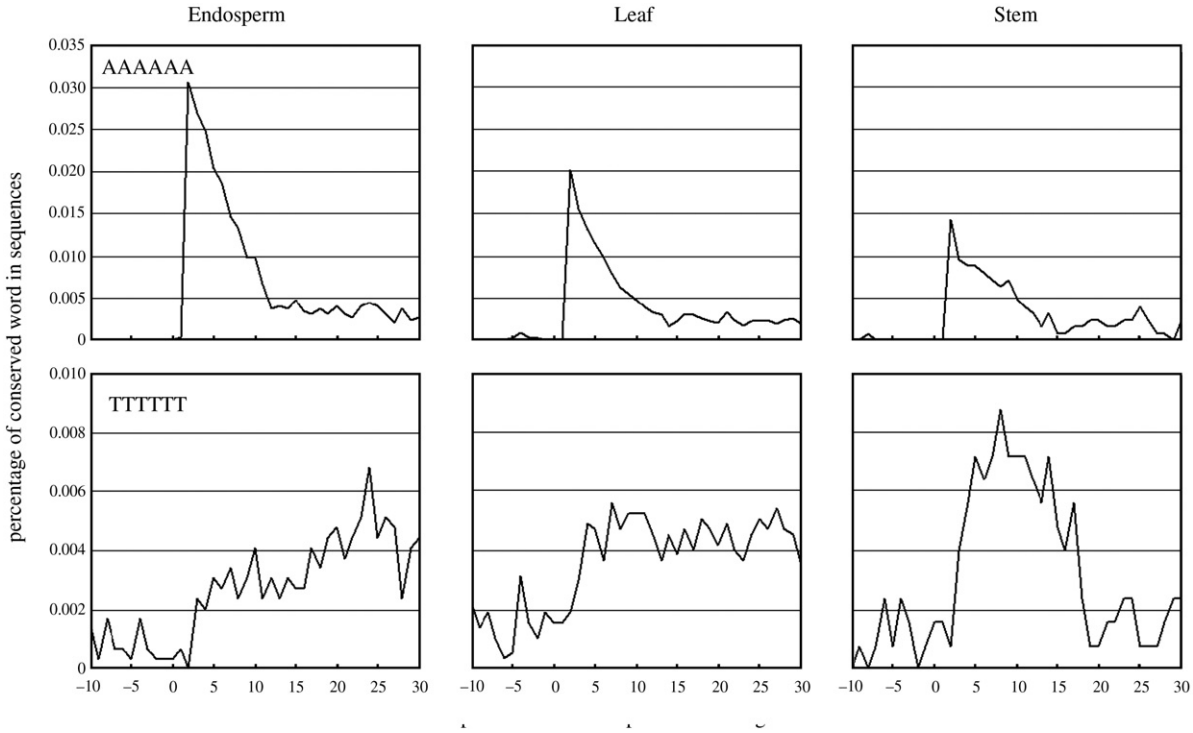


Fig. 3. Two typical downstream elements were scanned in stem, leaf and endosperm sequences from 10 nt upstream to 30 nt downstream of a CS. The *Y*-coordinate was the percentage of sequences in which the word was found in this position.

II, putative NUEs, were distributed 10 to 40 nt upstream of a CS. The most significant word was AATAAA, which had the highest *Z*-score and χ^2 value of all of the words found in this analysis. Of the 1979 sequences that could be scanned using this word, 854 of them were found in the predicted positions and accounted for 8.6% of the 9911 3′-UTR sequences. ATATAT appeared in 2281 sequences, but only 568 (5.7%) of them were 10–40 nt upstream of a CS. The other words in this cluster also consisted exclusively of A and T letters, with the sole exception of the ATGAAT hexamer. Cluster III appeared to be more dispersive than the other three clusters, with words mainly found in a region that spanned 10 nt upstream to 40 nt downstream of cleavage sites. Since all of these words had more than four T letters, they were designated as T-rich elements (TRE). The AAAAAA hexamer, its two substitutions in which a letter A changed to G, and the TTTTTT hexamer were mainly distributed downstream of poly(A) sites and were included in cluster IV. These four words were denoted as downstream elements (DE).

In summary, based on the generated position distribution profile and base frequency statistic, a general model of the conservative elements of 3′-end processing and their distribution in rice is displayed in Fig. 2B.

3.4. Tissue specific distribution of downstream elements

The entropy results reveal that there is a fourth peak in the downstream of CS only observed in the curve generated from stem-derived sequences. Thus, the tissue distributions of two typical downstream elements, AAAAAA and TTTTTT, were analyzed from 10 nt upstream to 30 nt downstream of a CS (Fig. 3). The AAAAAA hexamer was found in more than 3% of endosperm UTRs after a CS, 123 times more than a random distribution. However, this signal was weaker in the analyses of leaf and stem tissues. Only 1.5% of stem 3′-UTR, less than half that was found in endosperm, contained the AAAAAA hexamer. The poly(T) signal was stronger in the downstream region in all three tissues, but was stronger and more concentrated in the stem. It indicated the distributions and intensity of conserved signal were diverse in the downstream sequences of different rice tissues.

4. Discussion

4.1. The structure of conserved regions in the 3′-UTR of rice

Using computational methods to analyze entropy and conserved regions, a compilation of rice poly(A) control signals was produced from 3′-UTR datasets covering 9911 unique cleavage sites. The distribution of conserved regions in rice is similar to the previous working models in *Arabidopsis* and yeast, but there are still small differences in word constructions.

Information entropy analysis revealed that the strongest signal appeared at the CS and a conservative region with a length of less than 3 nt. Based on the statistical result, 79% of rice 3′-UTRs have CA or TA in the CS, therefore it is the most probable signal at this position. This result is consistent with the findings from six eukaryotic species (Graber et al., 1999), which

indicates that the CS signal is the most conserved positioning signal in all eukaryotes.

The function of an FUE is to enhance the formation efficiency of plant mRNA (Rothnie, 1996). Two FUE element clusters were identified in the rice genome. One is a word cluster covering the core TGTA motif, generally 50–70 nt upstream of the CS, which was reported to contribute to the efficiency of 3′-processing and has been correlated with transcript stability in yeast (Graber, 2003). The second FUE is a GT-repeat element around 100 nt upstream of a CS. In *Arabidopsis*, FUEs are usually T-rich elements that are found approximately 100 nt upstream of a CS (Loke et al., 2005; Zhao et al., 1999). Therefore, these result show that different words and positions have been selected in rice and *Arabidopsis* to control poly(A) efficiency.

The AATAAA sequence is the most significant NUE, and is equally prevalent in both *Arabidopsis* (Graber et al., 1999) and rice in 8.6% 3′-UTRs which are located in 10–30 nt upstream of a CS. The word TATATA was identified as an FUE, but its location differs by 20–30 nt from that of the NUE in yeast (van Helden et al., 2000). However, its position distribution showed that it is a significant NUE in rice.

In *Arabidopsis*, cleavage elements (CE) are conserved regions 10 nt upstream to 20 nt downstream of cleavage sites that contain a subset of small *cis*-elements: two T-rich sequences flanking both sides of the CS (Loke et al., 2005). The construction and positions of CEs correspond to TREs and DEs in our rice model. The only difference is that the DE in rice also includes the AAAAAA hexamer and some A-rich elements, all of which are as significant as the TRE in downstream sequences.

4.2. Control signal variation in different rice tissues

Herein, we also found that word distributions were diverse in the CS downstream sequences of different rice tissues. In the entropy distribution of sequences from rice stems, there was a special fourth peak at 15 nt downstream of the CS. After scanning the DE above around this position, the intensity of the poly(A) signal and the distribution of the poly(T) signal both differ from those of other tissues. Considering that the CS downstream sequences followed the 3′-end of the cDNA and that this position was so close to CS, we thought that this signal might be related to the choice of cleavage position. Research has indicated that the alternative polyadenylation phenomenon is common in mammalian (Gautheret et al., 1998) and plant genes (Meyers et al., 2004). It has been estimated that at least 49% of human genes and 31% of mouse genes have alternative polyadenylation sites (Yan and Marr, 2005). Moreover, differential poly(A) sites have been shown to occur in a tissue- or disease-specific manner (Edwalds-Gilbert et al., 1997). Statistical biases in the distribution of alternative mRNA forms among EST libraries of different tissues have also been found for the human and mouse genomes (Beaudoing and Gautheret, 2001). In this study, the results suggest that the mRNA transcribed in rice stem cells is preferentially cleaved from the genome before the poly(T) signal. Nevertheless, clarification of the relationship between the tissue-specific regulation of polyadenylation and the signals distributed downstream of the CS requires further analysis with a large number of 3′ ESTs.

Acknowledgements

We thank Dr. Christoper L. Hemme (Oak Ridge National Lab, Tennessee, USA) and Dr. Aifen Zhou (University of Oklahoma, Oklahoma, USA) for critical reading of the manuscript. This work was supported by a grant from the Project of Science and Technology of Zhejiang Province (Grant no: 021107614).

References

Beaudoing, E., Gautheret, D., 2001. Identification of alternate polyadenylation sites and analysis of their tissue distribution using EST data. Genome Res. 11, 1520–1526.

Colgan, D.F., Manley, J.L., 1997. Mechanism and regulation of mRNA polyadenylation. Genes Dev. 11, 2755–2766.

Dreyfus, M., Regnier, P., 2002. The poly(A) tail of mRNAs: bodyguard in eukaryotes, scavenger in bacteria. Cell 111, 611–613.

Edwalds-Gilbert, G., Veraldi, K.L., Milcarek, C., 1997. Alternative poly(A) site selection in complex transcription units: means to an end? Nucleic Acids Res. 25, 2547–2561.

Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D., 1998. Cluster analysis and display of genome-wide expression patterns. PNAS 95, 14863–14868.

Ewing, B., Hillier, L., Wendl, M.C., Green, P., 1998. Base-calling of automated sequencer traces using Phred. I. Accuracy assessment. Genome Res. 8, 175–185.

Gautheret, D., Poirot, O., Lopez, F., Audic, S., Claverie, J.M., 1998. Alternate polyadenylation in human mRNAs: a large-scale analysis by EST clustering. Genome Res. 8, 524–530.

Goff, S.A., et al., 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science 296, 92–100.

Graber, J.H., 2003. Variations in yeast 3′-processing *cis*-elements correlate with transcript stability. Trends Genet. 19, 473–476.

Graber, J.H., Cantor, C.R., Mohr, S.C., Smith, T.F., 1999. In silico detection of control signals: mRNA 3′-end-processing sequences in diverse species. Proc. Natl. Acad. Sci. U. S. A. 96, 14055–14060.

Kislauskis, E.H., Zhu, X., Singer, R.H., 1994. Sequences responsible for intracellular localization of beta-actin messenger RNA also affect cell phenotype. J. Cell Biol. 127, 441–451.

Knirsch, L., Clerch, L.B., 2000. A region in the 3′ UTR of MnSOD RNA enhances translation of a heterologous RNA. Biochem. Biophys. Res. Commun. 272, 164–168.

Loke, J.C., Stahlberg, E.A., Strenski, D.G., Haas, B.J., Wood, P.C., Li, Q.Q., 2005. Compilation of mRNA polyadenylation signals in *Arabidopsis* revealed a new signal element and potential secondary structures. Plant Physiol. 138, 1457–1468.

Mazumder, B., Seshadri, V., Fox, P.L., 2003. Translational control by the 3′-UTR: the ends specify the means. Trends Biochem. Sci. 28, 91–98.

Meyers, B.C., et al., 2004. Analysis of the transcriptional complexity of *Arabidopsis thaliana* by massively parallel signature sequencing. Nat. Biotechnol. 22, 1006–1011.

Reinert, G., Schbath, S., Waterman, M.S., 2000. Probabilistic and statistical properties of words: an overview. J. Comput. Biol. 7, 1–46.

Rothnie, H.M., 1996. Plant mRNA 3′-end formation. Plant Mol. Biol. 32, 43–61.

Schbath, S., 2000. An overview on the distribution of word counts in Markov chains. J. Comput. Biol. 7, 193–201.

Shannon, C.E., 1948. A mathematics theory of communication. Bell Syst. Tech. J. 27, 379-423-623-656.

van Helden, J., del Olmo, M., Perez-Ortin, J.E., 2000. Statistical analysis of yeast genomic downstream sequences reveals putative polyadenylation signals. Nucleic Acids Res. 28, 1000–1010.

Yan, J., Marr, T.G., 2005. Computational analysis of 3′-ends of ESTs shows four classes of alternative polyadenylation in human, mouse, and rat. Genome Res. 15, 369–375.

Yu, J., et al., 2002. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science 296, 79–92.

Zhang, Z., Schwartz, S., Wagner, L., Miller, W., 2000. A greedy algorithm for aligning DNA sequences. J. Comput. Biol. 7, 203–214.

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.