DNA RESEARCH **18**, 189–200, (2011)
Advance Access Publication: 17 May 2011

doi:10.1093/dnares/dsr011

In silico Analysis of 3′-End-Processing Signals in *Aspergillus oryzae* Using Expressed Sequence Tags and Genomic Sequencing Data

MIZUKI Tanaka[1,2,†], YOSHIFUMI Sakai[3,†], OSAMU Yamada[4], TAKAHIRO Shintani[1], and KATSUYA Gomi[1,*]

Laboratory of Bioindustrial Genomics, Department of Bioindustrial Informatics and Genomics, Graduate School of Agricultural Science, Tohoku University, 1-1 Tsutsumidori-Amamiyamachi, Aoba-ku, Sendai 981-8555, Japan[1]; Japan Society for the Promotion of Science, 8 Ichibancho, Chiyoda-ku, Tokyo 102-8472, Japan[2]; Laboratory of Bioindustrial Informatics, Department of Bioindustrial Informatics and Genomics, Graduate School of Agricultural Science, Tohoku University, 1-1 Tsutsumidori-Amamiyamachi, Aoba-ku, Sendai 981-8555, Japan[3] and National Research Institute of Brewing, 3-7-1 Kagamiyama, Higashi-hiroshima, Hiroshima 739-0046, Japan[4]

*To whom correspondence should be addressed. Tel. +81 22-717-8901. Fax. +81 22-717-8902.
E-mail: gomi@biochem.tohoku.ac.jp

Edited by: Katsumi Isono
(Received 16 November 2010; accepted 17 April 2011)

Abstract

To investigate 3′-end-processing signals in *Aspergillus oryzae*, we created a nucleotide sequence data set of the 3′-untranslated region (3′ UTR) plus 100 nucleotides (nt) sequence downstream of the poly(A) site using *A. oryzae* expressed sequence tags and genomic sequencing data. This data set comprised 1065 sequences derived from 1042 unique genes. The average 3′ UTR length in *A. oryzae* was 241 nt, which is greater than that in yeast but similar to that in plants. The 3′ UTR and 100 nt sequence downstream of the poly(A) site is notably U-rich, while the region located 15–30 nt upstream of the poly(A) site is markedly A-rich. The most frequently found hexanucleotide in this A-rich region is AAUGAA, although this sequence accounts for only 6% of all transcripts. These data suggested that *A. oryzae* has no highly conserved sequence element equivalent to AAUAAA, a mammalian polyadenylation signal. We identified that putative 3′-end-processing signals in *A. oryzae*, while less well conserved than those in mammals, comprised four sequence elements: the furthest upstream U-rich element, A-rich sequence, cleavage site, and downstream U-rich element flanking the cleavage site. Although these putative 3′-end-processing signals are similar to those in yeast and plants, some notable differences exist between them.

Key words: *Aspergillus oryzae*; 3′-end processing signal; polyadenylation signal; 3′-untranslated region (3′ UTR)

1. Introduction

In eukaryotes, most mRNAs have a poly(A) tail at their 3′ end. 3′-end-processing of eukaryotic pre-mRNA involves endonucleolytic cleavage and polyadenylation.[1–3] The 3′-end cleavage and polyadenylation site are regulated by several sequence elements, which have been extensively studied in mammalian, yeast, and plant cells.[4–8] In mammals, three elements are known as primary sequence elements: the polyadenylation signal, cleavage site, and downstream U/GU-rich elements. In addition, two auxiliary sequence elements (upstream U-rich elements and downstream G-rich elements) have also been identified. Among these elements, the polyadenylation signal, which is the hexanucleotide AAUAAA or its variant AUUAAA, located 10–35 nucleotides (nt) upstream of the poly(A) site is the most highly conserved sequence. In yeast and plants, A-rich sequence elements also exist ∼10–30 nt upstream of the

[†] These authors contributed equally to this work.

cleavage site, but these elements are less well conserved compared with mammalian polyadenylation signals. Among many sequences identified as A-rich sequences, AAUAAA is the most well-conserved sequence in both yeast and plants. In addition to A-rich sequence elements, further upstream elements, designated efficiency elements in yeast or far upstream elements in plants, the cleavage site, and the downstream U-rich element flanking the cleavage site have been described.

In Japan, the filamentous fungus *Aspergillus oryzae* has long been used for the production of traditional fermented foods, such as sake, soy sauce, and miso (soybean paste), and its long history of use in the food industry is a testament to its safety.[9] In addition, *A. oryzae* has the ability to secrete large amounts of protein, and therefore, it has recently gained recognition as a favourable host organism for recombinant protein production.[10,11] However, secretion yields of heterologous proteins from *A. oryzae* are low compared with those of homologous proteins or proteins from closely related fungal species.[12] Recently, we revealed that the transcript of a heterologous gene containing the AT-biased codon was prematurely polyadenylated within the coding region of *A. oryzae*.[13] This premature polyadenylation was prevented by the alteration of its codon to better suit *Aspergillus* codon usage. This result suggested that cryptic 3′-end-processing signals are recognized by *A. oryzae* within the coding region of the heterologous gene and that these signals are eliminated by codon optimization. However, no experimental data exist on 3′-end-processing signals in filamentous fungi, including *A. oryzae*.

To elucidate 3′-end-processing signals in *A. oryzae*, we created a nucleotide sequence data set of the 3′-untranslated region (3′ UTR) and 100 nt downstream of the poly(A) site using *A. oryzae* expressed sequence tags (ESTs) and genomic sequencing data. Using this data set, we identified several putative 3′-end-processing signals in *A. oryzae*. To our knowledge, this is the first report of the identification of 3′-end-processing signals in filamentous fungi.

2. Materials and methods

2.1. Creation of the A. oryzae poly(A) data set

A total of 21 446 EST sequences in the *A. oryzae* EST database (http://nribf2.nrib.go.jp/EST2/index.html), created by sequencing from the 5′ end of the cDNA insert,[14] were searched for sequences that contained at least eight consecutive A residues, yielding 1647 EST sequence entries. Subsequently, EST sequences containing oligo(A) stretches inherently present in the genome were eliminated by comparison with genomic DNA sequences.[15] In addition, to eliminate mis-annotated genes, only EST sequences in which the poly(A) site was located within 1000 nt downstream of the stop codon were selected. EST sequences with poly(A) sites located within the coding region, probably caused by internal priming, were also eliminated in this manner. Nine pairs of redundant EST sequences with identical poly(A) sites were considered to be derived from a single cDNA and were removed from the data set. Finally, 1065 EST sequences were selected by these processes. Genomic DNA-based sequences within the 3′ UTR and 100 nt sequence downstream of the poly(A) site were extracted for the poly(A) site data set (T residues were converted to U residues). This data set contained 22 pairs of EST sequences derived from the same gene with different poly(A) sites. Therefore, this data set comprised only sequences derived from 1043 unique genes.

The poly(A) site was designated as the last nucleotide in the genome sequence preceding the poly(A) tail. When an adenine residue was found at the poly(A) site in the genome sequence, this adenine was termed the poly(A) site nucleotide according to the recent reports of 3′-end-processing signals in plants and *Chlamydomonas* on the basis of EST sequencing data,[8,16] indicating that the first adenine of a poly(A) tail tended to be transcribed from the genomic DNA.[17–19]

2.2. DNA microarray analysis

The *A. oryzae* wild-type strain RIB40, which was used for genome sequencing analysis,[15] was grown in sterilized wheat bran media (3.0 g wheat bran with 1.8 ml distilled water) at 30°C for 33 h. Total RNA extraction, mRNA preparation, and DNA microarray analysis were performed according to the methods of Tamano *et al*.[20] Purified Cy3- or Cy5-labelled cDNA probes were hybridized using 12 K *A. oryzae* oligonucleotide microarrays (Fermlab, Tokyo, Japan). After global normalization, the relative fluorescence intensity of each gene was normalized to that of the *histone H4* gene, which was used as a reference. We selected 5384 genes whose intensities were found to be reproducible and reliable ($P < 0.1$) in dye-swap experiments.

2.3. Generation of sequence logos

Sequence logos of around the poly(A) site were generated using the enoLOGOS web tool.[21]

2.4. Analysis of oligonucleotide frequencies

A standard score (*Z*-score) was used to detect the most over-represented hexanucleotide sequences from -30 to -15 nt (region II), according to the zeroth- and first-order Markov chain models.[22] The *Z*-score of a hexanucleotide sequence ($w = x_1\ x_2\ x_3\ x_4\ x_5\ x_6$, where x is the nucleotide sequence) was calculated as follows:

$$\frac{f_{\mathrm{obs}}(w) - f_{\mathrm{exp}}(w)}{\left(f_{\mathrm{exp}}(w) \times \dfrac{1 - f_{\mathrm{exp}}(w)}{n}\right)^{1/2}}$$

In this definition, $f_{\mathrm{obs}}(w)$ denotes the observed frequency of w, i.e. the number of occurrences of w in s divided by the number of occurrences of sequences having the same length as w in s, where s ranges over all sequences of length 6 located in region II; $f_{\mathrm{exp}}(\mathrm{w})$ denotes the expected frequency of w, determined as the value $f_{\mathrm{obs}}(x_1) \times f_{\mathrm{obs}}(x_2) \times f_{\mathrm{obs}}(x_3) \times f_{\mathrm{obs}}(x_4) \times f_{\mathrm{obs}}(x_5) \times f_{\mathrm{obs}}(x_6)$ in the zeroth-order model or $f_{\mathrm{obs}}(x_1\ \ x_2) \times f_{\mathrm{obs}}(x_2\ \ x_3) \times f_{\mathrm{obs}}(x_3\ \ x_4) \times f_{\mathrm{obs}}(x_4\ \ x_5) \times f_{\mathrm{obs}}(x_5\ \ x_6)/(f_{\mathrm{obs}}(x_2) \times f_{\mathrm{obs}}(x_3) \times f_{\mathrm{obs}}(x_4) \times f_{\mathrm{obs}}(x_5)$ in the first-order model; and n denotes the number of sequences of length 6 located in region II.

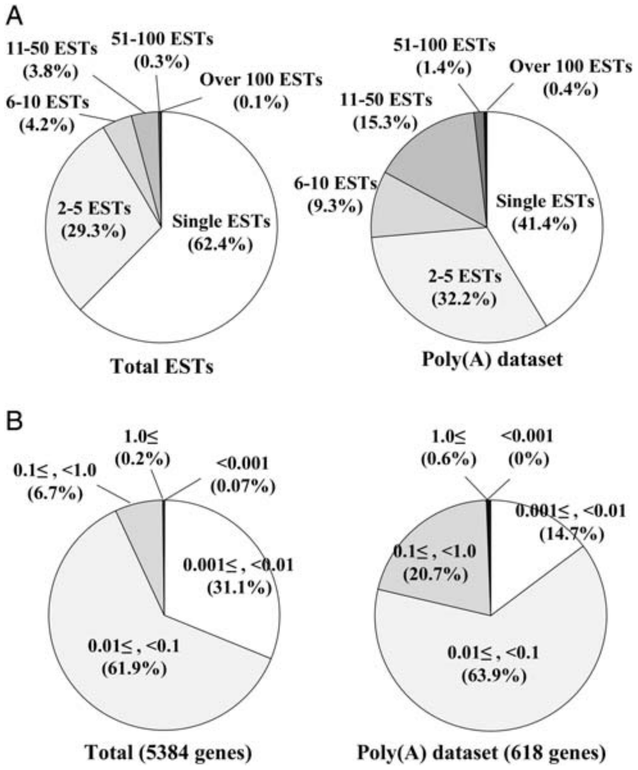

Figure 1. Profile of the *A. oryzae* poly(A) data set. (**A**) Frequency distribution of EST contigs based on the EST copy number. The EST copy number of each contig contained in the *A. oryzae* poly(A) data set was obtained from the *A. oryzae* EST database (http://nribf2.nrib.go.jp/EST2/index.html). Data on the total EST contigs were obtained from the study by Akao *et al.*[14] (**B**) Gene expression levels determined by DNA microarray analysis. The fluorescence intensity of each gene was normalized to that of the *histone H4* gene.

2.5. *Search for protein factors involved in pre-mRNA 3′-end-processing in A. oryzae*

Homologs of protein factors involved in eukaryotic pre-mRNA 3′-end-processing were retrieved by searching the *A. oryzae* genome database (http://www.bio.nite.go.jp/dogan/project/view/AO, http://nribf2.nrib.go.jp/) using the BlastP program.

3. Results and discussion

3.1. *Profile of the A. oryzae poly(A) data set*

We obtained 1065 sequences for the *A. oryzae* poly(A) data set from the EST database, as described in the Materials and methods section. First, 1043 unique genes contained in the *A. oryzae* poly(A) data set were classified into functional categories known as eukaryotic orthologous groups,[23] according to the gene list in the *A. oryzae* genome database. Compared with the classification of all genes found in the genome database, the number of genes classified into the Unannotated category was markedly lower in the poly(A) data set [43% in the genome database vs. 30% in the poly(A) data set]. In contrast, the number of genes classified into Information storage and processing and Cellular processes and signalling categories was higher in the poly(A) data set [7 and 12% in the genome database vs. 14 and 20% in the poly(A) data set, respectively]. The number of genes classified into Metabolism and Poorly characterized categories was similar between the genome database and poly(A) data set. These results indicated that the poly(A) data set covers a wide range of genes classified into diverse functional categories despite the poly(A) data set comprising only 1043 unique genes of the 12 074 genes predicted in the *A. oryzae* genome database.

In contrast, because EST sequences were accumulated by single-pass sequencing of the 5′ end of the cDNA insert,[14] the poly(A) data set could cover <10% of the total genes, and thus, the poly(A) data set might show some bias towards highly expressed genes. To assess this possibility, we compared the frequency distributions of EST contigs that corresponded to each of the 1043 genes in the poly(A) data set with those of the total 7589 contigs in the EST database[14] (Fig. 1A). Whereas contigs with frequencies of >6 accounted for ~10% of the total EST contigs, contigs with corresponding frequencies accounted for ~25% of the poly(A) data set. However, singletons accounted for 40% of the poly(A) data set. In addition, we examined the expression levels of genes in the poly(A) data set by DNA microarray analysis (Fig. 1B). Of the total of 5384 genes selected by microarray analysis, the number of relatively highly expressed genes (expression ratio > 0.1) accounted for approximately 7%, whereas it accounted for 20% of the 618 genes of the poly(A) data set. The remaining genes (80%) were expressed at low levels. These results suggested that the poly(A) data set was somewhat biased towards highly expressed genes, but this fact enabled the identification of 3′-end-processing signals.

In eukaryotes, many genes including >50% of human and rice genes have multiple polyadenylation sites.[8,24] This alternative polyadenylation has been recognized as an important mechanism for gene expression regulation.[25] However, no study has investigated on alternative polyadenylation on the basis of bioinformatics analyses in filamentous fungi. Although only 22 pairs of duplicated EST sequences with alternative poly(A) sites were included in the poly(A) data set, 14 pairs of these sequences had distant poly(A) sites located at least 30 nt apart (Supplementary Table S1). This result suggested that alternative polyadenylation also generally occurs in filamentous fungi.

3.2. *Analysis of 3′ UTR length and sequence elements of 3′ UTR-binding proteins*

In eukaryotes, 3′ UTR regulates mRNA stability and translational efficiency through sequence elements for 3′ UTR-binding proteins and microRNAs or through its length.[26–31] In *A. nidulans*, stability of transcripts involved in nitrogen metabolism was dependent on their 3′ UTRs.[32,33] Therefore, the 3′ UTRs may play an important role in gene expression regulation in filamentous fungi. However, no comprehensive information exists about 3′ UTRs in filamentous fungi. Hence, we analysed the distribution of 3′ UTR lengths in *A. oryzae* and determined their average and median lengths to compare with those in yeast and plants, which were also determined by analysis of EST sequencing data. In *A. oryzae*, 3′ UTR lengths were predominantly distributed in the range of 51 to 350 nt (Fig. 2). The average 3′ UTR length in *A. oryzae* was 241 nt, while the median 3′ UTR length was 203 nt. The average 3′ UTR length in *Saccharomyces cerevisiae* is 144 nt (median 3′ UTR length is 121 nt)[5] and that in plants is 289 nt (*Oryza sativa*) and 223 nt (*Arabidopsis thaliana*).[8] These results suggested that 3′ UTR length in *A. oryzae* is greater than that in yeast but similar to that in plants.

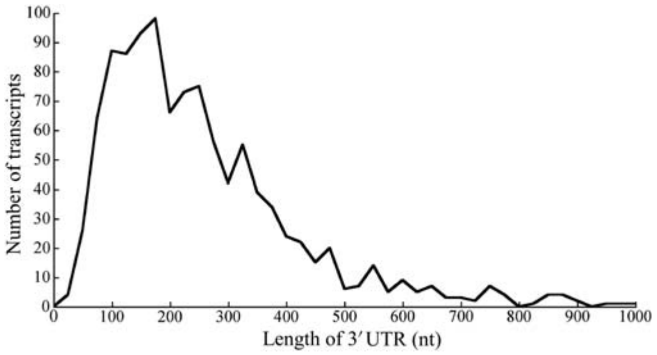

Figure 2. Distribution of 3′ UTR lengths determined for 1065 unique EST sequences. The average length is 241 nt.

The most well-known sequence elements for 3′ UTR-binding proteins in eukaryotes are AU-rich elements (AREs) and the PUF consensus motif.[34,35] We searched for transcripts containing the yeast putative AREs (UAUUUAUU and UUAUUUAU) and PUF consensus motif (UGUANAUA) within the *A. oryzae* 3′ UTR.[36–38] In the poly(A) data set, 12 and 23 genes possessed AREs and the PUF consensus motif within the 3′ UTR, respectively (Supplementary Table S2). One gene (AO090011000041) particularly exhibited overlapping AUUUA sequences (AUUUAUUUA), a typical ARE motif. In addition, we found orthologs of the yeast ARE-binding protein (Pub1) and four of six yeast Puf family proteins (Puf1, Puf3, Puf4, and Puf6) in the *A. oryzae* genome (Supplementary Table S3). These results suggested the existence of a regulation system for gene expression that utilizes 3′ UTR-binding proteins in filamentous fungi.

3.3. *Nucleotide profile of the* A. oryzae *3′ UTR*

To determine 3′ end processing elements in *A. oryzae*, we first measured the single nucleotide frequencies for all positions within the 3′ UTR and 100 nt sequence downstream of the poly(A) site (set at position 0). As shown in Fig. 3A, this region was notably U-rich, while AU accounted for 62% of nucleotides in this region (U = 34%; A = 28%). Meanwhile, AU content of the coding region in *A. oryzae* was 48% (http://www.kazusa.or.jp/codon/), suggesting that a high AU content is characteristic of this region. The 3′ UTR was markedly U-rich, but a A-rich region was observed upstream of the poly(A) site—particularly, the −29 to −14 nt region had a high A content with >30%. In addition, a high U content was also observed in the +1 to +20 nt region immediately downstream of the poly(A) site, but A and U content in the downstream +20 to +100 nt region was almost equal. This AU-rich element (ARE) located in the region immediately downstream of the U-rich region flanking the poly(A) site was also found in yeast and plants, but it has not been defined as the 3′-end-processing element in those organisms.[4,7,8] Moreover, the poly(A) site (position 0) had an extremely high A content (78%), and as described in the Materials

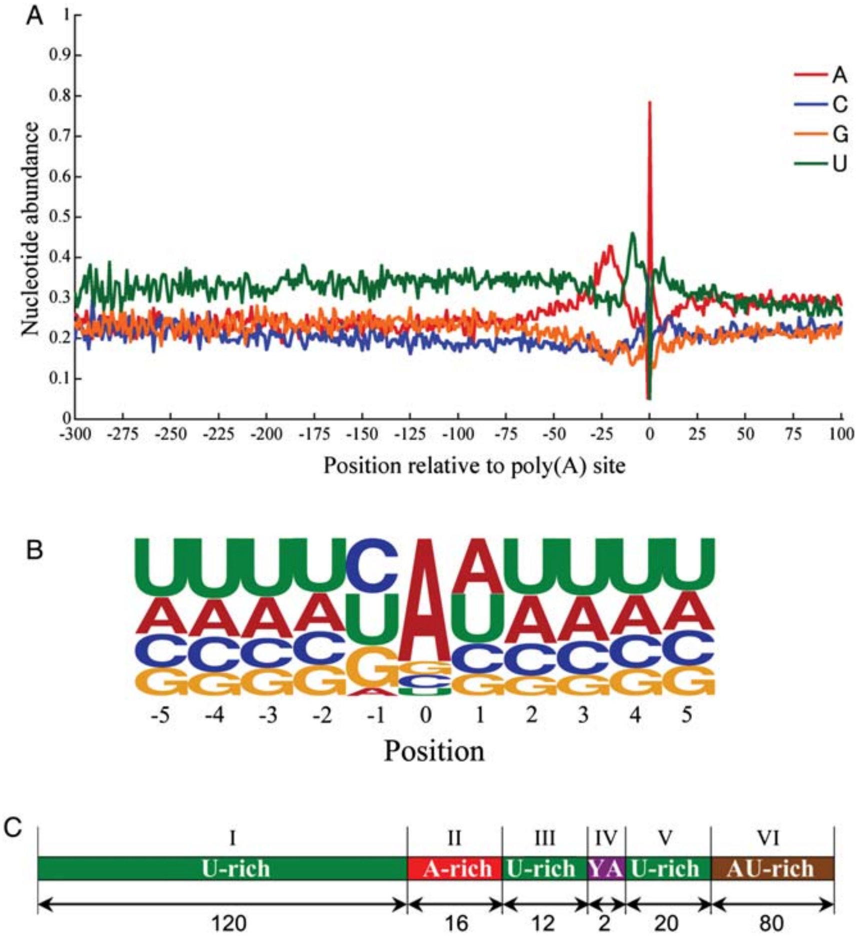

Figure 3. Single nucleotide frequencies in the 3′ UTR and 100 nt sequence downstream of the poly(A) site. (**A**) Single nucleotide profile in the 3′ UTR and 100 nt sequence downstream of the poly(A) site. The poly(A) site is at position 0. The upstream sequence of the poly(A) site is designated minus and the downstream sequence is designated plus. (**B**) Sequence logo generated from the actual frequency of occurrence of each of the four nucleotides around the cleavage site. (**C**) Six regions of the 3′ UTR and 100 nt sequence downstream of the poly(A) site formed according to the single nucleotide profile. The cleavage and polyadenylation site is located between regions IV and V.

and methods section, the first adenine of the poly(A) tail was designated as the poly(A) site nucleotide. High C nucleotide usage was observed at position −1 immediately before the poly(A) site compared with other positions (position −3, 20%; position −2, 21%; position −1, 37%; and position 0, 7%; Fig. 3B). The content of pyrimidine nucleotides (C and U) at position −1 was 68%, suggesting that CA or UA dinucleotides form the optimal cleavage site in *A. oryzae*, similar to that observed in plants.

Importantly, the nucleotide distribution profile of the 3′ UTR in *A. oryzae* was similar to that in plants,[7,8,39] yeast,[4] and mammals,[24] although the U-rich region was expanded towards the coding region of *A. oryzae*. On the basis of the nucleotide profile observed, the 3′ UTR plus 100 nt sequence downstream of the poly(A) site was divided into six signal element regions, designated regions I–VI, to identify the sequence elements for 3′-end-processing (Fig. 3C).

3.4. *Search for nucleotide sequence elements for 3′-end-processing*

To identify 3′-end-processing elements, we searched for tetramer–heptamer nucleotide sequences that appeared most frequently in each signal element region (Table 1, the top 50 list is available in Supplementary Table S4). In region II, equivalent to the region containing the polyadenylation signal in mammals, no significantly conserved hexanucleotide sequence was observed, similar to that observed in yeast and plants. The top-ranked hexanucleotide was AAUGAA in region II. The top two pentanucleotides (AAUGA and AUGAA) were partial sequences of AAUGAA, and all of the top three heptanucleotides contained the AAUGAA sequence (Table 1). In addition, according to the zeroth- and first-order Markov chain models, calculation of a standard score (*Z*-score) to measure the standard deviation of the hexanucleotide sequences from its

Table 1. The top five sequences (4−7 nt) that most frequently appear in 3′ ends

Region I (from −149 to −30 nt)							
4 nt	Number[a]	5 nt	Number	6 nt	Number	7 nt	Number
UUUG	629	UGUUU	343	UUCUUU	172	UUUCUUU	99
UGUU	628	UGUAU	341	UUUCUU	162	UUUUCUU	84
UUGU	624	UUUCU	316	UGUUUU	152	UUCUUUU	82
GUUU	619	UCUUU	310	UCUUUU	149	UGUAUAU	61
AUUU	617	UUGUU	301	UUUUCU	144	UUUGUUU	60
				UUGUUU	144		
Region II (from −29 to −14 nt)							
4 nt	Number	5 nt	Number	6 nt	Number	7 nt	Number
AAUA	286	AAUGA	119	AAUGAA	64	AAAUGAA	23
AAUG	257	AUGAA	110	AUGAAU	48	AAUGAAA	22
AAAU	233	AAUAU	99	AAUAAA	44	AAUGAAU	20
AUGA	216	AAUAA	93	AAUAUA	39	AAAUAAA	18
UAAU	215	AUAAU	92	AAAUGA	37	AAUAUGA	17
		AAAUA	92			AAUAAAU	17
Region III (from −13 to −2 nt)							
4 nt	Number	5 nt	Number	6 nt	Number	7 nt	Number
UUUU	170	CUUUU	64	UCUUUU	24	UUUUGUU	11
AUUU	158	UUUUC	58	UUCUUU	23	UUUCUUU	11
UUUC	150	AUUUU	56	UUUUCU	22	UUUUCUU	10
CUUU	136	UUUCU	55	UUUCUU	22	UUCUUUU	10
UUAU	129	UUUAU	51	UUUUGU	19	UGUUUAU	10
		UCUUU	51				
Region IV (from −1 to 0 nt)							
2 nt	Number						
CA	328						
UA	269						
GA	235						
UG	36						
UC	32						
Region V (from +1 to +20 nt)							
4 nt	Number	5 nt	Number	6 nt	Number	7 nt	Number
UUCU	186	UUUUC	76	UUUUCU	36	UUUUUCU	17
UCUU	184	UUUCU	68	UUUUUC	32	UUUUCUU	16
UUUC	169	CUUUU	68	UCUUUU	31	UUUUCUC	14
UUUU	157	UUCUU	67	UUCUUU	27	UUUCUUU	14
CUUU	149	UUUUU	61	CUUUUU	27	UUUCUCU	14
						CUUUUUU	14
Region VI (from +21 to +100 nt)							
4 nt	Number	5 nt	Number	6 nt	Number	7 nt	Number
AUAU	465	AUGUA	180	UAUGUA	74	AUAUGUA	32
UGUA	453	UUGUA	177	AGAAAA	74	AUAUAUA	30
AAUA	426	UAUAU	177	AUAUAU	66	AAAGAAA	30
UAUA	423	GUAGA	177	AUUGUA	64	UAUAUAU	28
UAGA	412	UGUAG	166	AAAGAA	64	AAGAAAA	28

[a] The number of transcripts with at least one occurrence.

Table 2. Top 10 hexanucleotide sequences mostly over-represented in region II

Rank	Markov order = 0			Markov order = 1		
	Word	*Z*-score	Number of occurences[a]	Word	*Z*-score	Number of occurences[a]
1	AAUGAA	16.603	67	AAUGAA	9.856	67
2	AUGAAU	13.146	48	AUGAAU	6.086	48
3	GAAUGA	11.594	31	GAAUGA	6.067	31
4	UGAAUG	10.594	25	GUCAAU	6.002	16
5	CAAUGC	10.083	17	GUCGCG	5.727	3
6	AAUGCA	9.026	25	CAAUGC	5.711	17
7	UCAAUG	8.684	21	UCGCGU	5.59	4
8	AUGCAA	8.576	24	AAUACA	5.196	29
9	AAAUGA	7.962	38	GGCAGU	5.027	5
10	GGAAUG	7.865	14	UCAAUU	4.994	23
70	AAUAAA	4.116	46			

Z-scores of the most over-represented hexanucleotide sequences in region II, according to the zeroth- and first-order Markov chain models.

[a]The number of hexanucleotide sequences found in region II.

expected occurrence revealed that AAUGAA was the most over-represented hexanucleotide sequence in region II (Table 2). These results suggested that AAUGAA is the most predominant hexanucleotide sequence in region II, although it accounted for only 6% of all transcripts (64 of1043). In contrast, according to the order of *Z*-scores, the AAUAAA sequence was not the major hexanucleotide sequence in region II, although it ranked third in the list of hexanucleotides. This was also demonstrated by plotting the distribution of hexanucleotide sequences, including AAUGAA and AAUAAA, in the region ranging from -40 to -1 nt (Fig. 4). The AAUGAA sequence was a single nucleotide variant of AAUAAA, but no study has reported that AAUGAA is the most effective A-rich sequence for the 3′-end-processing element in any eukaryote. Interestingly, point mutation of AAUAAA to AAUGAA results in a significant reduction of polyadenylation efficiency by *in vitro* 3′-end-processing reactions, using nuclear extracts from *Xenopus* and mammalian cells.[18,40] Thus, the 3′-end-processing machinery in *A. oryzae* may be somewhat different from that in higher eukaryotes.

Predominant sequence motifs in the upstream of the A-rich region (region I), called efficiency elements in yeast and far upstream elements in plants, have been identified. The best sequence for yeast efficiency elements is UAUAUA and its single nucleotide variants (UAUGUA and UACAUA).[4,41,42] In contrast, the best sequence of plant far upstream elements is UGUA.[8,38] In addition, this region in mammalian cells is defined as the auxiliary upstream element, and the UGUAN sequence element may function as a recognition element for 3′-end-processing proteins in case of A(A/U)UAAA-lacking 3′ UTRs.[43] However, these sequences were not predominant in region I of the *A. oryzae* poly(A) data set (Table 1). Moreover, no other sequence motif was highly conserved in this region, although the top nucleotide sequences were notably U-rich sequences. Similarly, no conserved sequence motif was observed in two other U-rich regions (regions III and V), suggesting that these sequences can be defined only as U-rich elements.

In region IV, equivalent of the cleavage site, the CA sequence ranked top and this motif existed in 31% of the sequences (Table 1). This suggested that the CA sequence is the most optimal cleavage site in *A. oryzae*. However, the GA sequence ranked third, and this motif existed in 22% of the sequences, suggesting that CA or UA dinucleotide sequences are not strictly conserved as the cleavage site. In region VI, no commonality was observed in the high-ranked

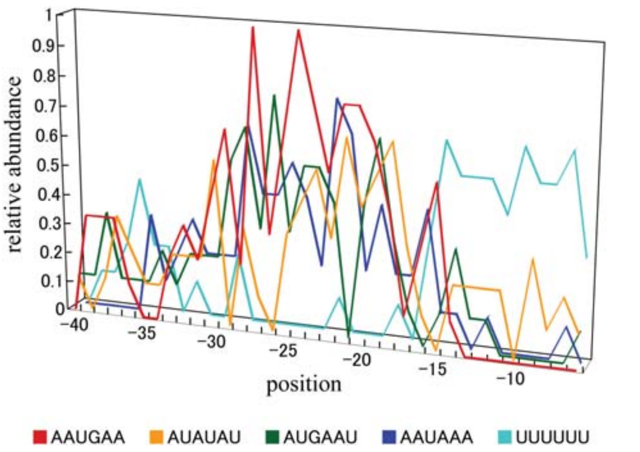

Figure 4. Representative hexanucleotide signals in the poly(A) signal region (from -40 to -1 nt).

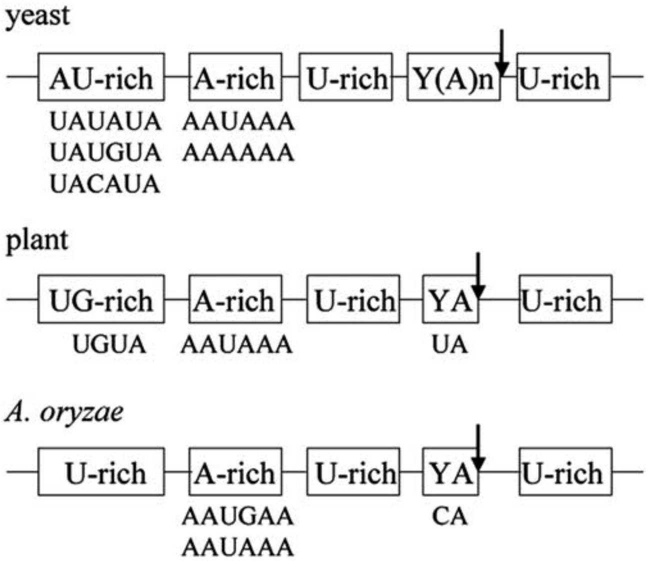

Figure 5. A schematic representation of the alignment of 3′-end-processing signals in *A. oryzae*, yeast, and plants. The arrow indicates the cleavage and polyadenylation site.

tetramer–heptamer sequences (Table 1), suggesting that this region cannot be defined as a 3′-end-processing element, similar to that in yeast and plants.

3.5. Putative 3′-end-processing signals in A. oryzae

Based on the information presented in this study, we proposed putative 3′-end-processing signals in *A. oryzae* (Fig. 5). The putative 3′-end-processing signals in *A. oryzae* were similar to those in yeast and plants but some differences were observed between them. First, A-rich sequences upstream of the poly(A) site were less well conserved in all three species than in mammals, and the predominant hexanucleotide in this region of *A. oryzae* differed from that of yeast and plants. The canonical hexanucleotide AAUAAA signal in mammals is also the most frequently occurring signal in this 3′ UTR of yeast and plants, whereas it is found only in ~13% and 7–10% of yeast and plant genes, respectively. In contrast, the most over-represented hexanucleotide in *A. oryzae* was AAUGAA, although this sequence accounted for only 6% of all transcripts, similar to yeast and plant AAUAAA sequences. Second, in the upstream of the A-rich region, while most dominant sequence motifs are well defined in yeast and plants (UAUAUA in yeast and UGUA in plants), no conserved sequence motif was observed in *A. oryzae*, except for the U-rich elements described earlier.

In a previous study, we showed that a cDNA of the mite *Dermatophagoides farinae*, known as Der f7, contains the AT-biased codon and therefore is prematurely polyadenylated within the coding region of *A. oryzae*. We also showed that codon optimization circumvents this premature polyadenylation.[13] The GC content of the native Der f7 open reading frame (ORF) was 37.8%, while that of the codon-optimized Der f7 ORF was 52.8%. Thus, A- and U-rich sequences within the coding region of native Der f7 cDNA were eliminated by codon optimization. The putative 3′-end-processing signals in *A. oryzae* deduced from this study supported that the A- and U-rich sequences present within the coding region of native Der f7 pre-mRNA were involved in incorrect 3′-end-processing. Although two AAUGAA sequences were present in the coding region of native Der f7 pre-mRNA, neither were located within the region located 10–30 nt upstream of the premature poly(A) sites.[13] This suggested that the AAUGAA sequence within the coding region of the AT-rich heterologous gene could not function by itself as an efficient 3′-end-processing signal in *A. oryzae*. The A- and U-rich sequences located upstream of the cleavage site might work co-operatively in 3′-end-processing. In future, whether the elimination of the top-ranked A-rich sequences within the coding region of heterologous genes results in the prevention of aberrant, premature transcription termination must be examined empirically.

3.6. Protein factors involved in the pre-mRNA 3′-end-processing machinery of A. oryzae

The recognition mechanism of 3′-end-processing signals has been well studied in yeast and mammals, and a large number of protein factors, e.g. ~14 proteins in mammals and ~20 proteins in yeast, are required for 3′-end-processing.[2,3] To examine whether these factors involved in 3′-end-processing are conserved in *A. oryzae*, we searched for homologous proteins of 20 yeast polyadenylation factors in the *A. oryzae* genome (Table 3). Most homologs of yeast polyadenylation factors, except for 3 factors (Ref2, Syc1, and Pti1), were found in the *A. oryzae* genome. These three factors are components of the cleavage and polyadenylation factor in yeast,[44] but no homologous proteins of these three factors are observed in plant and mammal genomes, suggesting that they are specific to yeast. In comparison with polyadenylation factors in human genomes, although no homologs of CFIm68 and CFIm59 were found in yeast, *A. oryzae*, and plant genomes, the homologue of CFIm25 was present in *A. oryzae* and plant genomes but not in the yeast genome. In contrast, the homologous protein (AO090001000725) of yeast Hrp1, reported to bind to RNA with specificity for the AU-rich efficiency element in yeast,[45,46] was found in the *A. oryzae* genome, whereas no Hrp1 homologue with higher similarity was found in plant

Table 3. Comparison of protein factors involved in pre-mRNA 3′-end-processing between *Aspergillus oryzae*, yeast, plants, and human

Aspergillus oryzae	*Saccharomyces cerevisiae*	*Arabidopsis thaliana*	*Homo sapiens*	BlastP score to yeast homologue	BlastP score to plant homologue	BlastP score to human homologue
	CFIB					
AO090001000725	Hrp1	None	None	3e−52	—	—
	CFIA	**AtCstF**	**CstF**			
AO090003000655	Rna14	AT1G17760 (AtCstF77)	CstF77	2e−69	6e−40	2e−46
AO090011000789	Rna15	AT1G71800 (AtCstF64)	CstF64	1e−12	1e−19	2e−35
None	None	AT5G60940 (AtCstF50)	CstF50	—	——	—
			CFIIm			
AO090026000698	Clp1	AT3G04680 (AtCLPS3)	hClp1	9e−34	4e−45	9e−47
AO090012001002	Pcf11	AT4G04885 (AtPCFS4)	hPcf11	3e−22	2e−15	4e−18
	CPF	**AtCPSF**	**CPSF**			
AO090103000017	Yhh1	AT5G51660 (AtCPSF160)	CPSF160	3e−69	3e−83	e−108
AO090005001277	Ydh1	AT5G23880 (AtCPSF100)	CPSF100	5e−26	6e−24	2e−25
AO090005001001	Ysh1	AT1G61010 (AtCPSF73-I)	CPSF73	e−168	e−140	7e−155
AO090005000813	Yth1	AT 1G30460 (AtCPSF30)	CPSF30	2e−40	5e−14	5e−28
AO080531000089[a]	Fip1	AT5G58040 (AtFIPS5)	hFip1	4e−10	4e−06	2e−11
AO090011000862	Pfs2	AT5G13480 (AtFY)	hPfs2 (WDR33)	1e−82	3e−89	1e−90
AO090103000067	Pta1	AT1G27595 (AtSYM2)	Symplekin	3e−21	0.085	7e−05
		AT5G01400 (AtSYM5)			—	
None	None	AT2G01730 (AtCPSF73-II)	CPSF73L	—	—	—
AO090005001504	Ssu72	AT1G73820 (Ssu72-like)	hSsu72	4e−48	1e−38	3e−41
AO090701000351	Glc7	AT2G39840 (AtPP1)	PP1α	e−157	e−141	1e−154
			PP1β			1e−151
None	Ref2	None	None	—	—	—
AO090001000739	Mpe1	AT5G47430	RBBP6	1e−70	2e−34	6e−34
None	Syc1	None	None	—	—	—
AO090120000355	Swd2	AT5G14530	WDR82	2e−52	5e−40	3e−54
None	Pti1	None	None	—	—	—
AO090005001182	Pap1	AT1G17980 (AtPAPS1)	PAP	e−151	e−103	e−114
		AT2G25850 (AtPAPS2)			e−102	
		AT4G32850 (AtPAPS4)			e−101	
			CFIm			
None	None	None	CFIm68	—	—	—

Continued

Table 3. Continued

Aspergillus oryzae	*Saccharomyces cerevisiae*	*Arabidopsis thaliana*	*Homo sapiens*	BlastP score to yeast homologue	BlastP score to plant homologue	BlastP score to human homologue
None	None	None	CFIm59	—	—	—
AO090003001316	None	AT4G25550 (AtCFIS2)	CFIm25	—	1e−60	5e−60

Protein factors involved in pre-mRNA 3′-end-processing in yeast, plants, and humans are based on the data described in the studies by Mandel *et al*.,[2] Millevoi and Vagner,[3] and Hunt *et al*.[48]

[a]Homologue of *A. oryzae* Fip1 was retrieved by searching the *A. oryzae* genome database deposited by the National Research Institute of Brewing, Japan (http://nribf2.nrib.go.jp/genome/blastscope.html).

and mammalian genomes. Furthermore, although homologs of human CstF-50 and CPSF73-II were present in the plant genome, these were not found in yeast and *A. oryzae* genomes. These observations suggested that the protein factors involved in the 3′-end-processing machinery of filamentous fungi resemble, in part, those of yeast and those of plants. This could be indicative of the evolutionary relationship between filamentous fungi, plants, and yeast. Some protein factors homologous to their counterparts in other organisms show differences in their RNA-binding specificity, positioning, and function. For example, while mammalian CPSF160 binds directly to the hexanucleotide AAUAAA signal, the homologue of yeast Yhh1 binds near the A-rich cleavage site but not the A-rich polyadenylation signal.[47] In this regard, because no sequence motif equivalent to the yeast AU-rich efficiency element or the mammalian UGUAN sequence was observed among the putative 3′-end-processing signals in *A. oryzae*, the RNA-binding specificity and positioning of the homologs Hrp1 and CFIm25 involved in 3′-end-processing in *A. oryzae* must be investigated.

3.7. Conclusions

In this study, we identified putative 3′-end-processing signals in *A. oryzae* using EST and genomic sequencing data. The putative 3′-end-processing signals in *A. oryzae* identified in this study comprised four elements: the furthest upstream U-rich element; A-rich sequence element (the most dominant sequence being AAUGAA); cleavage site (the most dominant sequence being CA); and U-rich element flanking the cleavage site. Although these putative 3′-end-processing signals in *A. oryzae* were similar to those found in yeast and plants, obvious differences were observed in the furthest upstream element and A-rich sequence element. To our knowledge, this is the first study of 3′-end-processing signals in filamentous fungi, and we believe that the data presented in this paper will provide knowledge critically important to the understanding of pre-mRNA 3′-end-processing in eukaryotes. In addition, this study also provides useful information on codon optimization of heterologous genes to prevent aberrant, premature polyadenylation within the coding region of filamentous fungi.

Acknowledgements: We are grateful to Dr Sachiko Shiro-Hasegawa for DNA microarray analysis.

Supplementary data: Supplementary data are available at www.dnaresearch.oxfordjournals.org.

Funding

This work was supported, in part, by a Grant-in-Aid for Scientific Research on Priority Areas, Applied Genomics (no. 17019001), from the Ministry of Education, Culture, Sports, Science, and Technology of Japan. M. T. is a JSPS Research Fellow.

References

1. Zhao, J., Hyman, L. and 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**, 405–45.
2. Mandel, C.R., Bai, Y. and Tong, L. 2008, Protein factors in pre-mRNA 3′-end processing, *Cell. Mol. Life Sci.*, **65**, 1099–122.
3. Millevoi, S. and Vagner, S. 2010, Molecular mechanisms of eukaryotic pre-mRNA 3′ end processing regulation, *Nucleic Acids Res.*, **38**, 2757–74.
4. Graber, J.H., Cantor, C.R., Mohr, S.C. and Smith, T.F. 1999, *In slico* detection of control signals: mRNA 3′-end-processing sequences in diverse species, *Proc. Natl Acad. Sci. USA*, **96**, 14055–60.
5. Graber, J.H., Cantor, C.R., Mohr, S.C. and Smith, T.F. 1999, Genomic detection of new yeast pre-mRNA 3′-end-processing signals, *Nucleic Acids Res.*, **27**, 888–94.
6. Hu, J., Lutz, C.S., Wilusz, J. and Tian, B. 2005, Bioinformatic identification of candidate *cis*-regulatory elements involved in human mRNA polyadenylation, *RNA*, **11**, 1485–93.

7. Loke, J.C., Stahlberg, E.A., Stenski, D.G., Haas, B.J., Wood, P.C. and Li, Q.Q. 2005, Complication of mRNA polyadenylation signals in *Arabidopsis* revealed a new signal element and potential secondary structures, *Plant Physiol.*, **138**, 1457–68.
8. Shen, Y., Ji, G., Hass, B.J., et al. 2008, Genome level analysis of rice mRNA 3′-end processing signals and alternative polyadenylation, *Nucleic Acids Res.*, **36**, 3150–61.
9. Machida, M., Yamada, O. and Gomi, K. 2008, Genomics of *Aspergillus oryzae*: learning from the history of koji mold and exploration of its future, *DNA Res.*, **15**, 173–83.
10. Christensen, T., Woeldike, H., Boel, F., et al. 1988, High level expression of recombinant genes in *Aspergillus oryzae*, *Bio/Technology*, **6**, 1419–22.
11. Lubertozzi, D. and Keasling, J.D. 2009, Developing *Aspergillus* as a host for heterologous expression, *Biotechnol. Adv.*, **27**, 53–75.
12. Gouka, R.J., Punt, P.J. and van den Hondel, C.A.M.J.J. 1997, Efficient production of secreted proteins by *Aspergillus*: progress, limitations and prospects, *Appl. Microbiol. Biotechnol.*, **47**, 1–11.
13. Tokuoka, M., Tanaka, M., Ono, K., Takagi, S., Shintani, T. and Gomi, K. 2008, Codon optimization increases steady-state mRNA levels in *Aspergillus oryzae* heterologous gene expression, *Appl. Environ. Microbiol.*, **74**, 6538–46.
14. Akao, T., Sano, M., Yamada, O., et al. 2007, Analysis of expressed sequence tags from the fungus *Aspergillus oryzae* cultured under different conditions, *DNA Res.*, **14**, 47–57.
15. Machida, M., Asai, K., Sano, M., et al. 2005, Genome sequencing and analysis of *Aspergillus oryzae*, *Nature*, **438**, 1157–61.
16. Shen, Y., Liu, Y., Liu, L., Ling, C. and Li, Q.Q. 2008, Unique features of nuclear mRNA poly(A) signals and alternative polyadenylation in *Chlamydomonas reinhardtii*, *Genetics*, **179**, 167–76.
17. Moore, C.L., Skolnikdavid, H. and Sharp, P.A. 1986, Analysis of RNA cleavage at the Adenovirus-2 L3 polyadenylation site, *EMBO J.*, **5**, 1929–38.
18. Sheets, M.D., Ogg, S.C. and Wickens, M.P. 1990, Point mutations in AAUAAA and the poly(A) addition site: effects on the accuracy and efficiency of cleavage and polyadenylation *in vitro*, *Nucleic Acids Res.*, **18**, 5799–805.
19. Chen, F., Macdonald, C.C. and Wilusz, J. 1995, Cleavage site determinants in the mammalian polyadenylation signal, *Nucleic Acids Res.*, **23**, 2614–20.
20. Tamano, K., Sano, M., Yamane, N., et al. 2008, Transcriptional regulation of genes on the non-syntenic blocks of *Aspergillus oryzae* and its functional relationship to solid-state cultivation, *Fungal Genet. Biol.*, **45**, 139–51.
21. Workman, C.T., Yin, Y., Corcoran, D.L., Ideker, T., Stormo, G.D. and Benos, P.V. 2005, enoLOGOS: a versatile web tool for energy normalized sequence logos, *Nucleic Acids Res.*, **33**, W389–92.
22. Van Helden, J., Del Olmo, M. and Perez-Ortin, J.E. 2000, Statistical analysis of yeast genomic downstream sequences reveals putative polyadenylation signals, *Nucleic Acids Res.*, **28**, 1000–10.
23. Tatusov, R.L., Federova, N.D., Jackson, J.D., et al. 2003, The COG database: an updated version includes eukaryotes, *BMC Bioinformatics*, **4**, 41–54.
24. Tian, B., Hu, J., Zhang, H. and Lutz, C.S. 2005, A large-scale analysis of mRNA polyadenylation of human and mouse genes, *Nucleic Acids Res.*, **33**, 201–12.
25. Edwalds-Gilbert, G., Veraldi, K.L. and Milcarek, C. 1997, Alternative poly(A) site selection in complex transcription units: means to an end?, *Nucleic Acids Res.*, **25**, 2547–61.
26. Pesole, G., Mignone, F., Gissi, C., Grillo, G., Licciulli, F. and Liuni, S. 2001, Structural and functional features of eukaryotic mRNA untranslated regions, *Gene*, **276**, 73–81.
27. Kuersten, S. and Goodwin, E.B. 2003, The power of the 3′ UTR: translational control and development, *Nat. Rev. Genet.*, **4**, 626–37.
28. Mazumder, B., Seshadri, V. and Fox, P.L. 2003, Translational control by the 3′-UTR: the ends specify the means, *Trends Biochem. Sci.*, **28**, 91–8.
29. Kozak, M. 2004, How strong is the case for regulation of the initiation step of translation by elements at the 3′ end of eukaryotic mRNAs?, *Gene*, **343**, 41–54.
30. Bartel, D.P. 2009, MicroRNAs: target recognition and regulatory functions, *Cell*, **136**, 215–233.
31. Kebaara, B.W. and Atkin, A.L. 2009, Long 3′-UTRs target wild-type mRNAs for nonsense-mediated mRNA decay in *Saccharomyces cerevisiae*, *Nucleic Acids Res.*, **37**, 2771–8.
32. Morozov, I.Y., Martinez, M.G., Jones, M.G. and Caddick, M.X. 2000, A defined sequence within the 3′ UTR of the *areA* transcript is sufficient to mediate nitrogen metabolite signaling via accelerated deadenylation, *Mol. Microbiol.*, **37**, 1248–57.
33. Caddick, M.X., Jones, M.G., van Tonder, J.M., et al. 2006, Opposing signals differentially regulate transcript stability in *Aspergillus nidulans*, *Mol. Microbiol.*, **62**, 509–19.
34. Wickens, M., Bernstein, D.S., Kimble, J. and Parker, R. 2002, A PUF family portrait: 3′ UTR regulation as a way of life, *Trends Genet.*, **18**, 150–7.
35. Barreau, C., Paillard, L. and Osborne, B. 2006, AU-rich elements and associated factors: are there unifying principles?, *Nucleic Acids Res.*, **33**, 7138–50.
36. Vasudevan, S. and Peltz, S.W. 2001, Regulated ARE-mediated mRNA decay in *Saccharomyces cerevisiae*, *Mol. Cell*, **7**, 1191–200.
37. Puig, S., Askeland, E. and Thiele, D. 2005, Coordinated remodeling of cellular metabolism during iron deficiency through targeted mRNA degradation, *Cell*, **120**, 99–110.
38. Jackson, J.S., Houshmandi, S.S., Leban, F.L. and Olivas, W.M. 2004, Recruitment of the Puf3 protein to its mRNA target for regulation of mRNA decay in yeast, *RNA*, **10**, 1625–36.
39. Dong, H., Deng, Y., Chen, J., et al. 2007, An exploration of 3′-end processing signals and their tissue distribution in *Oryza sativa*, *Gene*, **389**, 107–13.
40. Wickens, M. and Stephenson, P. 1984, Role of the conserved AAUAAA sequence: for AAUAAA point mutations prevent messenger RNA 3′ end formation, *Science*, **226**, 1045–51.

41. Guo, Z., Russo, P., Yun, D.F., Butler, J.S. and Sherman, F. 1995, Redundant 3′ end-forming signals for the yeast CYC1 mRNA, *Proc. Natl. Acad. Sci. USA*, **92**, 4211–14.
42. Graber, J.H., McAllister, G.D. and Smith, T.F. 2002, Probabilistic prediction of *Saccharomyces cerevisiae* mRNA 3’-processing sites, *Nucleic Acids Res.*, **30**, 1851–58.
43. Venkataraman, K., Brown, K.M. and Gilmartin, G.M. 2005, Analysis of a noncanonical poly(A) site reveals a tripartite mechanism for vertebrate poly(A) site recognition, *Genes Dev.*, **19**, 1315–27.
44. Nedea, E., He, X., Kim, M., et al. 2003, Organization and function of APT, a subcomplex of the yeast cleavage and polyadenylation factor involved in the formation of mRNA and small nucleolar RNA 3’-ends, *J. Biol. Chem.*, **278**, 33000–10.
45. Kessler, M.M., Henry, M.F., Shen, E., et al. 1997, Hrp1, a sequence-specific RNA-binding protein that shuttles between the nucleus and the cytoplasm, is required for mRNA 3′-end formation in yeast, *Genes Dev.*, **11**, 2545–56.
46. Pérez-Cañadillas, J.M. 2006, Grabbing the message: structural basis of mRNA 3′ UTR recognition by Hrp1, *EMBO J.*, **25**, 3167–78.
47. Dichtl, B., Blank, D., Sadowski, M., Hubner, W., Weiser, S. and Keller, W. 2002, Yhh1p/Cft1p directly links poly(A) site recognition and RNA polymerase II transcription termination, *EMBO J.*, **21**, 4125–35.
48. Hunt, A.G., Xu, R., Addepalli, B., et al. 2008, Arabidopsis mRNA polyadenylation machinery: comprehensive analysis of protein-protein interactions and gene expression profiling, *BMC Genomics*, **9**, 220.