

Bioinformatic identification of candidate *cis*-regulatory elements involved in human mRNA polyadenylation

JUN HU,¹ CAROL S. LUTZ,¹ JEFFREY WILUSZ,² and BIN TIAN¹

¹Department of Biochemistry and Molecular Biology, New Jersey Medical School, University of Medicine and Dentistry of New Jersey, Newark, New Jersey 07101, USA

²Department of Microbiology, Immunology and Pathology, Colorado State University, Fort Collins, Colorado 80523, USA

ABSTRACT

Polyadenylation is an essential step for the maturation of almost all cellular mRNAs in eukaryotes. In human cells, most poly(A) sites are flanked by the upstream AAUAAA hexamer or a close variant, and downstream U/GU-rich elements. In yeast and plants, additional *cis* elements have been found to be located upstream of the poly(A) site, including UGUA, UAU, and U-rich elements. In this study, we have developed a computer program named PROBE (Polyadenylation-Related Oligonucleotide Bidimensional Enrichment) to identify *cis* elements that may play regulatory roles in mRNA polyadenylation. By comparing human genomic sequences surrounding frequently used poly(A) sites with those surrounding less frequently used ones, we found that *cis* elements occurring in yeast and plants also exist in human poly(A) regions, including the upstream U-rich elements, and UAU and UGUA elements. In addition, several novel elements were found to be associated with human poly(A) sites, including several G-rich elements. Thus, we suggest that many *cis* elements are evolutionarily conserved among eukaryotes, and human poly(A) sites have an additional set of *cis* elements that may be involved in the regulation of mRNA polyadenylation.

Keywords: polyadenylation; *cis* elements; regulation

INTRODUCTION

mRNA polyadenylation is the cellular process that adds poly(A) tails to maturing mRNAs. A poly(A) tail is found on nearly all mRNAs in eukaryotes (with the exception of most histone mRNAs), and it is involved in virtually every aspect of mRNA metabolism, including mRNA stability, translation, and mRNA transport (Jacobson and Peltz 1996; Sachs et al. 1997; Wickens et al. 1997). The process of polyadenylation is composed of two tightly coupled steps (Colgan and Manley 1997; Edmonds 2002): In the first step, an endonucleolytic cleavage takes place at a site determined by surrounding RNA sequences (*cis* elements) and their binding proteins (*trans* factors); the second step involves polymerization of an adenosine tail. In vertebrates, the average length of the poly(A) tail is ~200 nucleotides (nt).

It is generally accepted that signals required for recognition of sites for polyadenylation reside near the cleavage site (poly[A] site). The strength of the *cis* elements for binding *trans* factors can determine the efficiency of mRNA production and influence the amount of mature, exported mRNA (Edwards-Gilbert et al. 1993). For simplicity, the genomic sequence surrounding a poly(A) site is referred to as the poly(A) region. Sequences upstream and downstream of a human poly(A) site are generally U-rich (Legendre and Gautheret 2003; Tian et al. 2005), which is an important feature that can be used to predict poly(A) sites (Tabaska and Zhang 1999; Legendre and Gautheret 2003). A hexamer AAUAAA or a close variant is located between 10 and 35 nt upstream of many mammalian poly(A) sites and is usually referred to as the polyadenylation signal (PAS). Sequences 10–40 nt downstream of the cleavage site are also known to be involved in directing polyadenylation. These elements do not have a clear consensus sequence but can be characterized as U-rich or GU-rich (Zarudnaya et al. 2003). PAS and U/GU-rich elements are usually referred to as core elements for polyadenylation. Furthermore, a number of auxiliary upstream elements (USEs) or downstream elements (DSEs) have been identified in viral and cellular systems

Reprint requests to: Bin Tian, Department of Biochemistry and Molecular Biology, New Jersey Medical School, University of Medicine and Dentistry of New Jersey, Newark, NJ 07101, USA; e-mail: btian@umdnj.edu; fax: (973) 972-5594.

Article published online ahead of print. Article and publication date are at <http://www.majournal.org/cgi/doi/10.1261/rna.2107305>.

that play regulatory roles in polyadenylation, including Simian Virus 40 (SV40), Human Immunodeficiency Virus type 1 (HIV-1), human C2 complement, collagen, cyclooxygenase-2, etc. (Carswell and Alwine 1989; Brown et al. 1991; Valsamakis et al. 1991; Moreira et al. 1995; Arhin et al. 2002; Natalizio et al. 2002; Hall-Pogar et al. 2005). The existence of these auxiliary elements indicates that polyadenylation can be regulated by *cis* elements other than PAS and U/GU-rich elements in human cells. In fact, <60% of human and mouse poly(A) sites are preceded by the AAUAAA PAS (Tian et al. 2005), raising the possibility that there may exist other enhancing elements for polyadenylation.

Yeast and plant genes utilize a different set of *cis* elements for polyadenylation (Graber et al. 1999; Zhao et al. 1999). While AAUAAA is a prominent hexamer located upstream of poly(A) sites in these species, it occurs to a much lesser extent. Other A-rich elements seem to be equivalent to the AAUAAA, and U-rich elements have been found both upstream and downstream of the poly(A) site. In addition, upstream elements have been found for both yeast and plant poly(A) sites. For example, UAUUA and UAUGUA elements are the efficiency elements (EEs) located 30–70 nt upstream of yeast poly(A) sites (Graber 2003). It is not known, however, to what extent these elements occur in human poly(A) regions.

We recently mapped a large number of poly(A) sites on human and mouse genes (Tian et al. 2005) and wished to address whether additional *cis* elements are associated with human poly(A) sites besides the well-characterized PAS and U/GU-rich elements. To this end, we have developed a computer program named PROBE (Polyadenylation-Related Oligonucleotide Bidimensional Enrichment) to identify *cis* elements that may play regulatory roles in mRNA polyadenylation. By comparing poly(A) regions of frequently used poly(A) sites, termed strong poly(A) sites, and those of less frequently used ones, termed weak poly(A) sites, we found that most *cis* elements found in yeast and plants also exist in human poly(A) regions, and their presence was biased to strong poly(A) sites. In addition, several novel elements were found to be associated with human poly(A) sites, including several G-rich elements. Thus, we suggest that many *cis* elements surrounding the poly(A) site appear to be evolutionarily conserved among all eukaryotes, and human poly(A) sites have an additional set of *cis* elements involved in the regulation of mRNA polyadenylation.

RESULTS AND DISCUSSION

Identification of *cis* elements by PROBE

We are interested in elucidating *cis* elements that may play enhancing roles in polyadenylation. We reasoned that putative *cis* elements would be over-represented in poly(A) regions of frequently used poly(A) sites (strong sites), and

under-represented in those of less frequently used poly(A) sites (weak sites). Our computer program PROBE is similar in spirit to that used by Fairbrother et al. (2002) for identifying exonic splicing enhancers, but with several major modifications (see below and Materials and Methods). We first classified poly(A) sites as strong or weak sites using the number of supporting cDNA/ESTs as a guide. A strong site was classified here as a site in a gene that was utilized more than 75% of the time. Other poly(A) sites in that same gene would be classified as weak sites. Thus, in our study only poly(A) sites belonging to genes having multiple poly(A) sites were considered. Among 29,283 human poly(A) sites in our recently created database for mammalian polyadenylation, polyA_DB (Zhang et al. 2005), 22,865 poly(A) sites belong to 7524 genes that have multiple poly(A) sites. Among all poly(A) sites considered, 3711 sites were classified as strong sites and 5663 sites as weak sites, corresponding to 12.7% and 19.3% of all human poly(A) sites, respectively. Our classification of strong and weak sites is similar to what Legendre and Gautheret (2003) previously reported. To enrich our data set, we used cDNA/ESTs from both normalized and non-normalized cDNA libraries, and used 75% as the cutoff. These two measures have opposing effects on selection of strong sites: Inclusion of normalized libraries would make selection less stringent as normalization narrows the difference of usage between strong and weak sites, whereas using the cutoff of 75% requires at least threefold difference in poly(A) site usage. In addition, we used cDNA/ESTs derived from a large number of cDNA libraries, corresponding to a wide spectrum of tissue sources. Thus, “strong” and “weak” should represent overall poly(A) site utilization in most tissues.

We then obtained sequences 100 nt upstream (–100) and 100 nt downstream (+100) of the poly(A) site (which was set at position 0). We divided the sequences into four subregions—–100/–41, –40/–1, +1/+40, and +41/+100—to assist in the identification of *cis* elements (Fig. 1). This step is based on two considerations: (1) Most known *cis* elements for polyadenylation reside within 100 nt from the poly(A) site and have spatial preference in polyadenylation. For example, PAS is located 10–35 nt upstream of the poly(A) site, and the U/GU-rich element is located within ~40 nt downstream of the poly(A) site (Chen et al. 1995). (2) The nucleotide compositions of –100/–41, –40/–1, +1/+40, and +41/+100 regions appear to be distinct from one another, and from regions >100 nt upstream of and >100 nt downstream of the poly(A) site (Tian et al. 2005). By considering these regions independently, we can use region-specific nucleotide background models to identify *cis* elements (see below), which will enhance both specificity and sensitivity. For convenience, we have named putative *cis* elements in the following way: Elements in the –100/–41 region are called auxiliary upstream elements (AUEs); elements in the –40/–1 region, core upstream elements (CUEs); elements in the +1/+40 region, core downstream elements (CDEs); elements in the +41/+100 region, auxiliary downstream elements (ADEs).

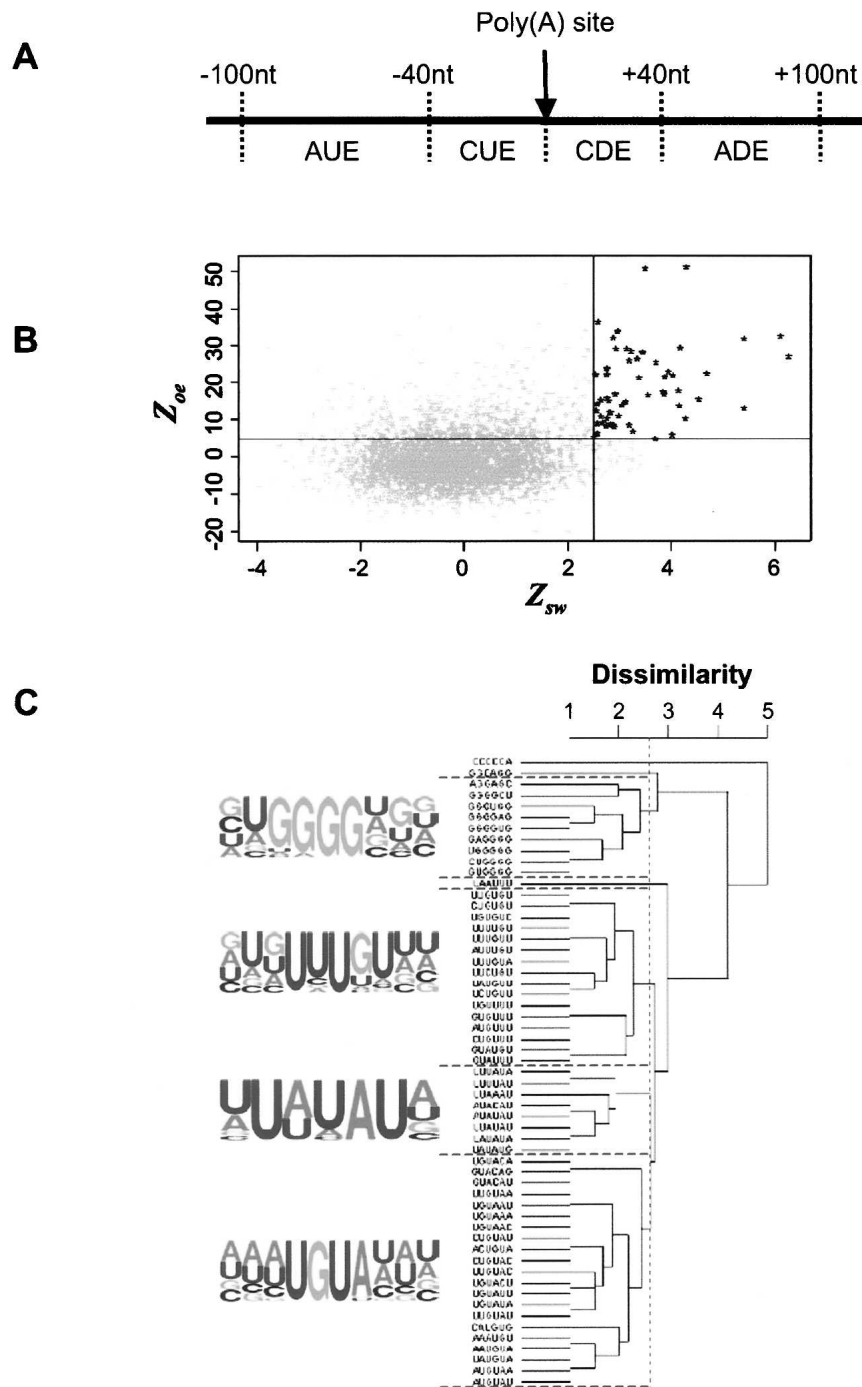


FIGURE 1. Schematic of a poly(A) region and identification of *cis* elements in the $-100/-41$ region. (A) A poly(A) region is a genomic sequence containing a poly(A) site. The poly(A) site is considered at position 0. Four subregions were investigated in this study, namely $-100/-41$, $-40/-1$, $+1/+40$, and $+41/+100$. Elements identified for these regions are named AUE, CUE, CDE, and ADE, respectively. (B) Scatter plot of all 4096 hexamers. (X-axis) z_{sw} , corresponding to the difference between strong and weak poly(A) sites in a specific region (the $-100/-41$ region in this graph); (Y-axis) z_{oe} , corresponding to the difference between observed and expected values in the specific poly(A) region. (Black asterisks) Hexamers whose z_{sw} and z_{oe} are above the respective cutoffs; (gray asterisks) the rest of the hexamers. (C) Clustering of hexamers. Hexamers were clustered according to their mutual dissimilarity by agglomerative hierarchical clustering, which is shown at *right*. Dissimilarity value 2.6 was used to group hexamers (see Materials and Methods). Grouped hexamers were aligned by a multiple sequence alignment method described in Materials and Methods. Each hexamer group gave rise to a sequence logo, shown at *left*.

It should be pointed out that the naming method here is arbitrary and does not necessarily reflect functions.

We then calculated the frequency of occurrence for all 4096 hexamers in the four regions surrounding the poly(A) site. The reason we used hexamers was based on two considerations: (1) Motif identification by hexamer enumeration was previously shown to be highly sensitive for RNA elements (Fairbrother et al. 2002); and (2) many known RNA functional elements have a size ~ 6 nt, such as AAUAAA for the PAS and AUUUA for the AU-rich elements (Chen and Shyu 1995; Bakheet et al. 2001). Two values were used to select hexamers that were overrepresented in specific regions of strong poly(A) sites. First, we used the score z_{sw} for measuring the difference of frequency of occurrence for a hexamer in a specific region of strong poly(A) sites versus weak poly(A) sites (see Materials and Methods for its calculation). We used the cutoff 2.5, which corresponds to p -value ~ 0.01 in the normal distribution. Second, we used the score z_{oe} to measure how significant a hexamer is in a specific poly(A) region in general (see Materials and Methods for its calculation). z_{oe} is the difference between observed frequency of occurrence and expected frequency of occurrence. We calculated expected values by using first-order Markov Chain (MC) models, i.e., dinucleotide frequencies, that are specific for regions. For each region, a cutoff was chosen, which was based on the 99th-percentile value (corresponding to p -value 0.01) of an extreme value distribution (EVD) of z_{oe} , derived from random sequences generated by the first-order MC model for the region. Hexamers with both z_{sw} and z_{oe} scores above the respective cutoffs were selected (Fig. 1B; also Supplemental Fig. 1 at http://exon.umdj.edu/suppl_data/PROBE/). This two-dimensional selection should result in less than one falsely identified hexamer ($4096 \times 0.01 \times 0.01 = 0.4$) in this study.

Selected hexamers were clustered according to their reciprocal sequence similarity (see Materials and Methods). Similar hexamers were grouped together and aligned by a multiple se-

quence alignment method similar to ClustalW with several modifications designed to (1) favor the most frequently occurring hexamer and (2) prevent identified *cis* elements from becoming too long after merge of hexamers. Aligned hexamers were used to generate sequence logos, representing *cis* elements. An example of this process is given in Figure 1C for the $-100/-41$ region. Graphs for other regions are provided as supplementary data (see Supplemental Fig. 1 at http://exon.umdj.edu/suppl_data/PROBE/). PROBE can conceivably be used for identification of *cis* elements in other systems and is available upon request.

Using PROBE, we identified 15 *cis* elements in four regions, including four AUEs, two CUEs, four CDEs, and five ADEs. Their sequence logos (Schneider and Stephens 1990), the number of supporting hexamers, the top hexamers with respect to z_{oe} , and their occurrences in specific regions of all poly(A) sites are listed in Table 1. In addition, z_{sw} and z_{oe} scores for all hexamers in different regions are provided as supplemental data (see Supplemental Table 1 at http://exon.umdj.edu/suppl_data/PROBE/). For each *cis* element, we derived a position-specific scoring matrix
















(PSSM) which we used to search poly(A) regions for the presence of the element. A positive score indicates that the likelihood of the presence of an element is higher than expected by random chance. A sequence having a positive score is called a hit. We searched all 29,283 poly(A) sites in the polyA_DB in the $-125/+125$ region and obtained scores for all elements at every location. We then calculated “fraction of hits” and “average score of hits” to characterize the elements. These values represent two characteristics of a *cis* element: The fraction of hits indicates the occurrence of all positive hits for a *cis* element, whereas the average score of hits indicates how close a hit’s sequence is to the consensus sequence of the *cis* element. The data obtained are summarized below.

AAUAAA PAS

The CUE.2 element (Table 1) identified in our survey is likely to be the AAUAAA PAS. The fact that AAUAAA is biased to strong poly(A) sites is in line with the notion that AAUAAA is an efficient signal for polyadenylation. Since we used poly(A) sites regardless of their association with

PAS, this result also validates our approach to identify *cis* elements that enhance polyadenylation. Using its scoring matrix to search all poly(A) sites in the $-125/+125$ region, we found the element was prominently present in the $-40/-1$ region (Fig. 2A). There is also a conspicuous difference between strong and weak poly(A) sites, with respect to both average score of hits and fraction of hits (Fig. 2A). To further validate this result, we did an exact word search in the $-40/-1$ region for all known PAS elements reported before (Beaudoing et al. 2000; Tian et al. 2005), including AAUAAA and 11 single nucleotide variants. We divided poly(A) sites into four groups: strong poly(A) sites; weak poly(A) sites; median poly(A) sites; which are the poly(A) sites from genes that have multiple poly(A) sites but no strong or weak poly(A) sites; and constitutive poly(A) sites, which are the sites from genes with only one poly(A) site. As expected, we found that the frequency of AAUAAA being used by strong poly(A) sites is similar to that used by constitutive sites, and higher than that used by weak or median sites (Fig. 2B). Interestingly, the close variant AUUAAA has the same frequency of occurrence in all types of poly(A) sites. Thus this result confirms that AAUAAA

TABLE 1. *Cis* elements identified in four regions surrounding the poly(A) site

Region	<i>Cis</i> element	Name ^a	Number of Hex ^b	Top 3 hexamers ^c	Percent of Hits ^d
$-100/-41$		AUE.1	9	GGGGAG, GUGGGG, GGGUGG	48%
		AUE.2	16	UUUGUA, GUAUUU, CUGUGU	93%
		AUE.3	8	UAUAUA, AUAUAU, UUAUA	51%
		AUE.4	21	UGUAUA, AUGUAU, UGUAAU	82%
$-40/-1$		CUE.1	17	UAUUUU, UGUUUU, UUUUUU	82%
		CUE.2	23	AAUAAA, AUAAG, AAAUAA	90%
$+1/+40$		CDE.1	25	GUGUCU, CUGCCU, UGUCUC	87%
		CDE.2	14	UUAUUU, UUUCUU, UGUUUU	90%
		CDE.3	21	UGUGUG, GUGUGU, CUGUGU	66%
		CDE.4	23	CUGGGG, UGUCUG, GUCUGU	68%
$+41/+100$		ADE.1	5	CCUCCC, CUCCCC, CACCCC	21%
		ADE.2	6	CCCGCC, CCCC GC, CCCGCG	29%
		ADE.3	15	GGUGGG, GGCUGG, GGGUGG	74%
		ADE.4	6	GGGCAG, GGCCAG, GGGGCC	27%
		ADE.5	18	GGGAGG, GGAGGG, GGGGAG	85%

^aAUE, auxiliary upstream elements; CUE, core upstream elements; CDE, core downstream elements; ADE, auxiliary downstream elements.

^bThe number of hexamers selected for a *cis* element.

^cTop three hexamers among hexamers selected for a *cis* element with respect to z_{oe} .

^dPercent of hits is the percentage of poly(A) sites that have hits (sequences with a positive score while comparing to an element) in a specific region.

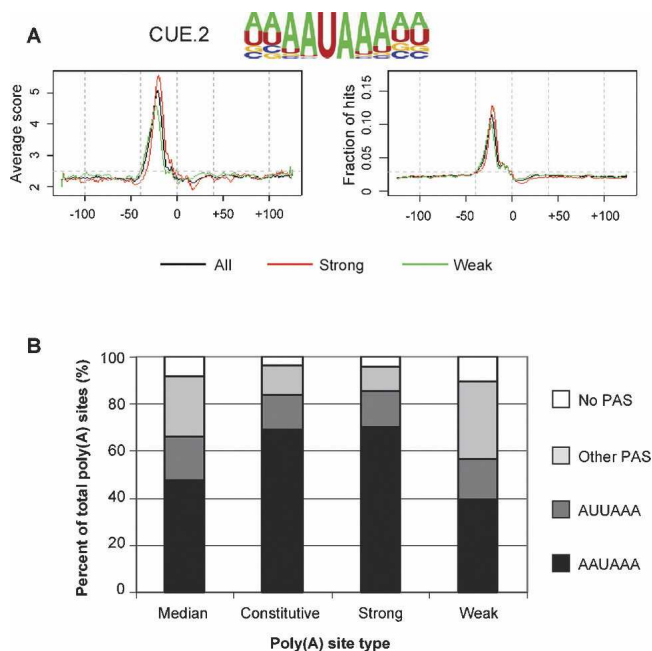


FIGURE 2. AAUAAA element. (A) Average score (left) and fraction of hits of CUE.2 in regions from all poly(A) sites (black lines), strong poly(A) sites (red lines), and weak poly(A) sites (green lines). (Dotted vertical lines) -100 -nt, -40 -nt, 0 -nt, $+40$ -nt, and $+100$ -nt positions; (dotted horizontal lines) the average of values in the -100 -nt to $+100$ -nt region from all poly(A) sites. (B) Association of different PAS hexamers with poly(A) sites of various types. Constitutive poly(A) sites are the sites from genes with only one poly(A) site. For genes with multiple poly(A) sites, the site that is utilized more than 75% of the time is classified as a strong poly(A) site. If there exists a strong site, other poly(A) sites in the same gene are classified as weak sites. If there is no strong site, all poly(A) sites are classified as median sites.

is an element that can distinguish strong and weak poly(A) sites, whereas AUUAAA is not.

U-rich elements

Three U-rich elements were found in different regions, namely, AUE.2, CUE.1, and CDE.2 (Table 1). A sequence search of these three elements showed almost identical profiles with respect to average score and fraction of hits (Fig. 3). Three peaks can be discerned in these graphs, corresponding to these three U-rich elements. The CDE.2 in the $+1/+40$ region is likely the binding site for CstF-64 (MacDonald et al. 1994). It ranges from $+1$ to $+40$, with a peak at $\sim+25$ (Fig. 3A), indicating that the U-rich element for binding CstF-64 is located within 40 nt downstream of the cleavage site, with a preferable position at $+25$ nt, which is in line with previous experimental data (Chou et al. 1994). The peaks corresponding to CUE.1 in both average score and fraction of hits plots are at ~-15 nt (Fig. 3B), which is just 3' to the average location of PAS at ~-20 . Strong sites have higher values at the peak in both cases

(Fig. 3B). Thus, it appears that a AAUAAA PAS element is usually followed (at its 3' end) by a U-rich element for strong poly(A) sites. In fact, this was observed for many model poly(A) sites, such as the poly(A) site of β -globin gene (Levitt et al. 1989) and SV-40 early poly(A) site (Ryner and Manley 1987; Wilusz and Shenk 1988). Interestingly, yeast and plant poly(A) sites all have U-rich elements located between the PAS and the poly(A) site (Graber et al. 1999). Our finding therefore suggests that the U-rich element located between the PAS and the poly(A) site is evolutionarily conserved. In addition, the presence of the upstream U-rich elements is in line with the finding that human Fip1, which binds U-rich RNAs and stimulates poly(A) polymerase, is associated with CPSF (Kaufmann et al. 2004).

GU-rich elements

GU-rich elements, such as UGUGUG, have been found in the downstream region of many poly(A) sites and, like the U-rich element, may serve as the binding site for CstF-64 (Perez Canadillas and Varani 2003). CDE.3 (Table 1) appears to correspond to a GU-rich element. In line with its activity for binding CstF-64, its occurrence is similar to the U-rich element in the downstream region, with a slight difference in that its peaks ($\sim+10$ nt, Fig. 4A) are located closer to the poly(A) site than the U-rich element ($\sim+25$ nt, Fig. 3A). Interestingly, CDE.1 (UGYCU; Y being U or C) and CDE.4 (UCUG) showed similar profiles to those of GU-rich elements (Fig. 4B,C). These sequences have not been known to play a role in polyadenylation. Thus, it will be interesting to assess fully their activity in the future.

UAUA and UGUA elements

AUE.3 (Table 1) contains a UAUA sequence, and AUE.4 (Table 1) contains a UGUA sequence. UAUA and UGUA elements are similar in sequence to the EE elements of yeast and plant poly(A) sites. Thus, they appear to represent a group of evolutionarily conserved elements for polyadenylation. Interestingly, AUE.3 and AUE.4 are the only elements whose average score profile is different than the fraction of hits profile (Fig. 5A,B), indicating that other elements that are similar to or overlap with AUE.3 and AUE.4 may exist in various regions. It remains to be tested whether these elements, while biased to strong poly(A) sites, are also functional in human cells. It is worth noting that this bioinformatic finding is consistent with a recent biochemical study that implicated the UGUA element in human poly(A) site recognition (Venkataraman et al. 2005). Taken together with U-rich elements described above, our data suggest that some *cis* elements in yeast and plant poly(A) regions are conserved in human ones.

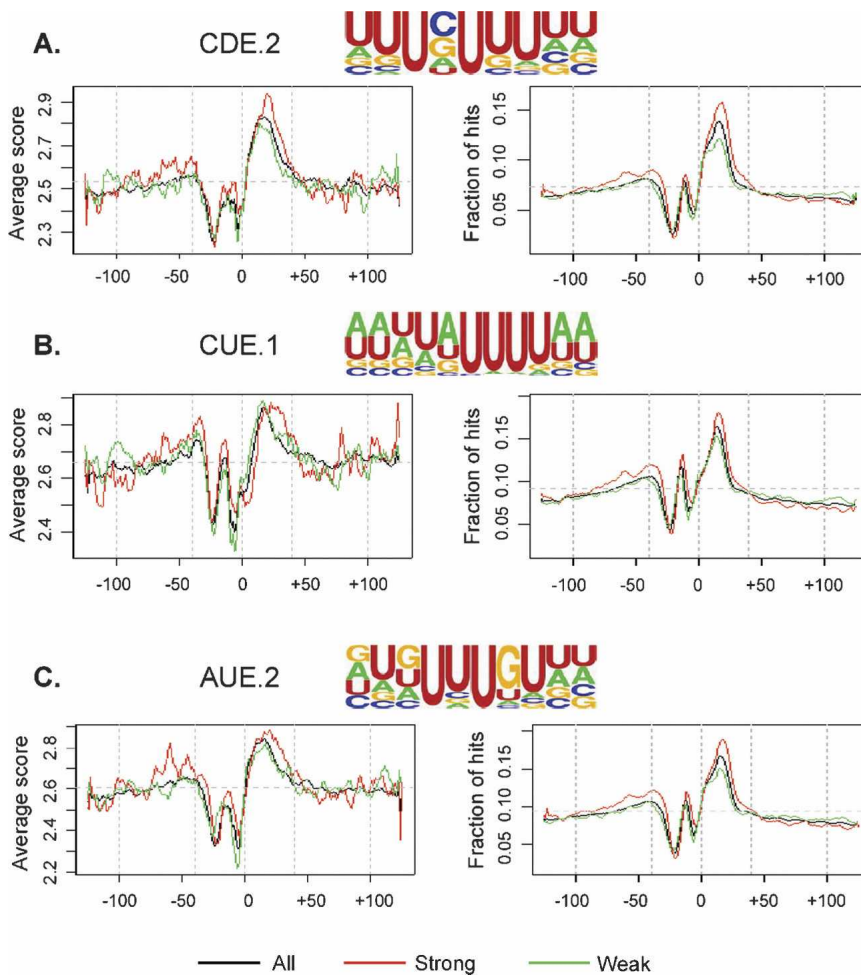


FIGURE 3. U-rich elements. (A) CDE.2; (B) CUE.1; (C) AUE.2. See the Figure 2A legend for detailed description of the graphs.

G-rich and C-rich elements

G-rich elements have been found in the $-100/-41$ and $+41/+100$ regions, including AUE.1, ADE.3, ADE.4, and ADE.5 (Table 1; Supplemental Fig. 2 at http://exon.umdj.edu/suppl_data/PROBE/). G-rich elements downstream to the poly(A) site have been implicated in enhancing polyadenylation in several previous studies (Bagga et al. 1995; Yonaha and Proudfoot 2000). The identification of a G-rich element in the $-100/-41$ upstream region (AUE.1) is novel. Interestingly, an early study of upstream elements of the polyadenylation signal in HIV-1 implicated a region containing a G-rich element in the regulation of polyadenylation (Valsamakis et al. 1992). The fact that G-rich elements are found in both upstream and downstream regions indicate that they could be general enhancers for polyadenylation independent of their position. One mechanism could be to recruit hnRNPH/H' proteins (Arhin et al. 2002).

ADE.1 and ADE.2 are C-rich elements (Table 1; Supplemental Fig. 2 at http://exon.umdj.edu/suppl_data/PROBE/). There have not been any reports for their role in polyadenyl-

ation. Thus, their functions remain to be validated in future experiments. One direction that needs to be investigated is the possibility of the base pairing of G-rich and C-rich elements, which can lead to secondary structures. On a similar note, G-rich sequences have been suggested to form G-quadruplex structures (Zarudnaya et al. 2003), and RNA secondary structures have been shown to play a role in mRNA polyadenylation (Graveley et al. 1996; Wu and Alwine 2004).

In summary, by comparing human genomic sequences surrounding strong poly(A) sites and those surrounding weak poly(A) sites, we have identified a number of *cis* elements that may play regulatory roles in polyadenylation. A schematic representation of these elements is shown in Figure 6. Our approach is validated by the presence of several known *cis* elements in our bioinformatic results, including the upstream AAUAAA element and the downstream U-rich and GU-rich elements. However, biochemical assays remain to be conducted to examine the functions of these *cis* elements. Equally important are the conservation of these *cis* elements across species and the possibility of using these *cis* elements to predict poly(A) sites on the genome. These will be pursued in future studies.

Several *cis* elements occurring in yeast and plants were also found to exist in human poly(A) regions, including the upstream U-rich elements and UAUA and UGUA elements. In addition, several novel elements were found to be associated with human poly(A) sites, including upstream and downstream G-rich elements, as well as C-rich elements. Thus, we suggest that many *cis* elements are evolutionarily conserved among all eukaryotes, and human poly(A) sites have an additional set of *cis* elements involved in the regulation of mRNA polyadenylation.

MATERIALS AND METHODS

Selection of hexamers based on z_{sw} and z_{oe}

Poly(A) regions surrounding human poly(A) sites were obtained as previously described (Tian et al. 2005). For genes with multiple poly(A) sites, we calculated the percentage of usage of each site using supporting cDNA/ESTs. A site supported by 75% of the total number of cDNA/ESTs was considered a strong site. If there exists a strong site in a

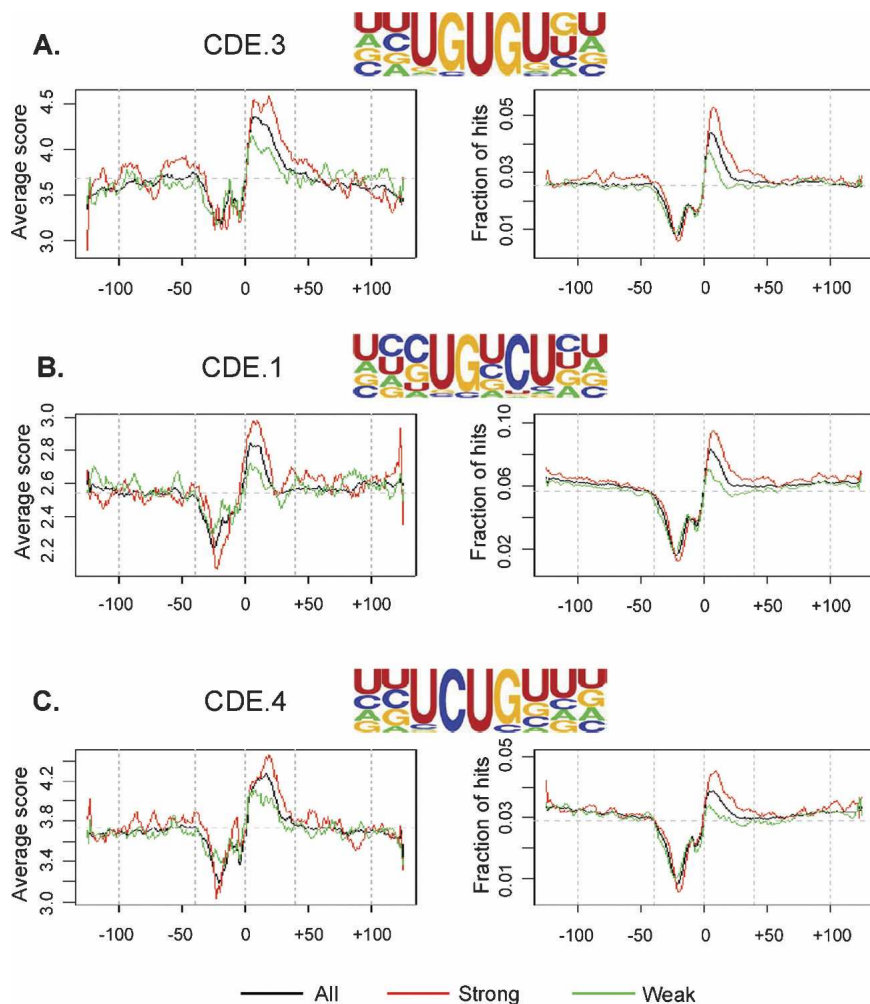


FIGURE 4. GU-rich elements. (A) CDE.3; (B) CDE.1; (C) CDE.4. See the Figure 2A legend for detailed description of the graphs.

particular gene, other sites of the same gene were considered weak sites.

For each hexamer (H), the difference of the frequency of occurrence in poly(A) regions of strong and weak poly(A) sites is represented by z_{sw} . It was calculated as follows:

$$z_{sw}(H) = \frac{f_s(H) - f_w(H)}{\sqrt{(1/N_s + 1/N_w)p(1-p)}}, \quad (1)$$

where

$$p = \frac{f_s(H) * N_s + f_w(H) * N_w}{(N_s + N_w)},$$

N_s and N_w are the total number of hexamers in a specific region of strong poly(A) sites and weak poly(A) sites, respectively. $f_s(H)$ and $f_w(H)$ are the frequency of occurrence of hexamer H in strong and weak poly(A) regions, respectively. The difference between observed and expected occurrences (z_{oe}) was calculated as follows:

$$z_{oe}(H) = \frac{N_o(H) - N_e(H)}{\sqrt{v_{oe}(H)}}, \quad (2)$$

where $N_o(H)$ is the occurrence of hexamer H in a specific region of all poly(A) sites (29,283 in total). Thus it is an observed value. $N_e(H)$ is expected occurrence and was calculated based on the first-order Markov Chain (MC) model of a specific poly(A) region of all poly(A) sites (Schbath 1997). $v_{oe}(H)$ is the estimated variance of $N_o(H) - N_e(H)$, calculated by a method described in (Schbath 1997). Since there is a wide difference in nucleotide composition in various regions surrounding the poly(A) site (Tian et al. 2005), we used MC models specific for particular regions. For example, the $-100/-41$ region used its own MC model derived from the $-100/-41$ region.

To get significant hexamers, we used a cutoff value of 2.5 for z_{sw} . We used an empirically determined cutoff value for z_{oe} which was derived as follows: We randomized sequences 500 times, maintaining dinucleotide frequency (first-order MC) and overall number of nucleotides. We used Equation 2 to calculate z_{oe} for all hexamers in each random set. The highest z_{oe} values from 500 random sets were used to derive an extreme value distribution (EVD). The 99th-percentile value (rank 495 in ascending order) was used as the cutoff. Hexamers with both z_{sw} and z_{oe} above their respective cutoffs were selected for further analysis.

Clustering of hexamers

Selected hexamers were grouped based on their mutual distance. The distance between two hexamers is their dissimilarity score (d) calculated as follows: $d = 6 - s$, where s is a similarity score. s was calculated by a dynamic programming method for sequence alignment that did not allow gaps. The match and mismatch scores were one and zero, respectively. Thus for perfectly matched hexamers, $s = 6$, and $d = 0$. Hexamers were clustered based on their mutual d scores using hierarchical clustering in program R with the “average” agglomeration method. A cutoff of 2.6 was used to group hexamers, which gave more robust *cis* elements than other cutoff values for different regions (data not shown). Groups containing more than three hexamers after clustering were selected for further analysis.

Making consensus sequence of *cis* elements

Hexamers belonging to the same group were aligned by the following multiple sequence alignment method: For each hexamer group, the hexamer with highest z_{oe} value was used as a seed for multiple alignment, instead of the neighbor joining method normally used in multiple alignment tools like Clus-

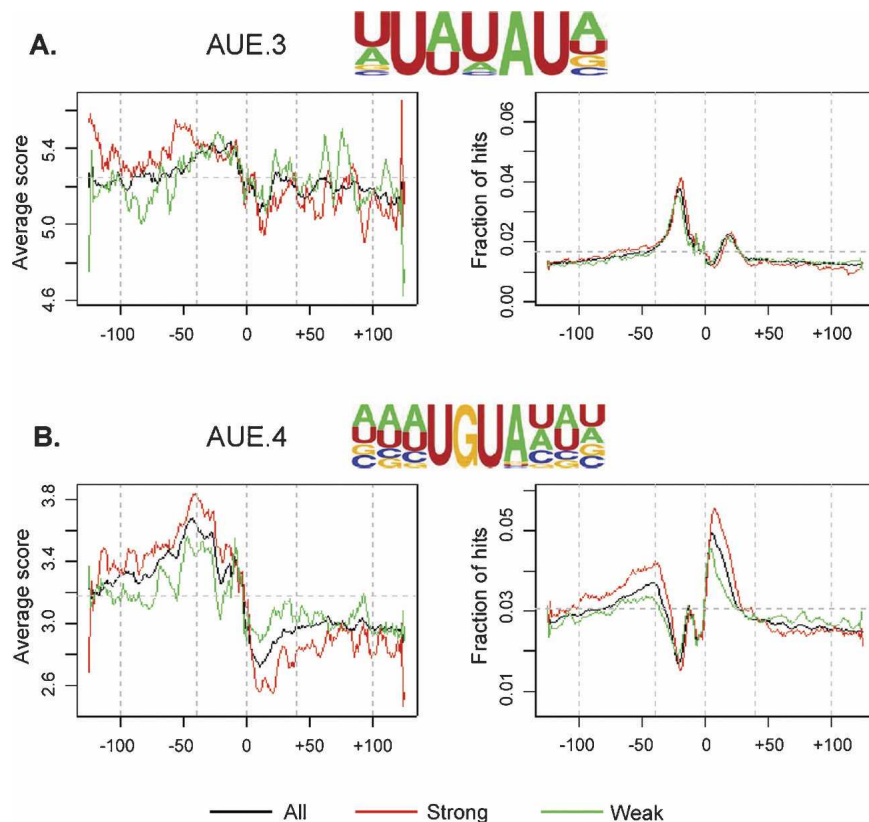


FIGURE 5. UAUA and UGUA elements. (A) AUE.3; (B) AUE.4. See the Figure 2A legend for detailed description of the graphs.

talW. We required all other hexamers to be aligned to the seed without gaps, which in effect limited expansion of identified *cis* elements. We then expanded the number of hexamers to 1000 by duplicating hexamers. The number of repetition of a hexamer is proportional to its occurrence in a specific poly(A) region, e.g., $-100/-41$. Gaps at the ends after the alignment are filled by randomly selecting nucleotides according to their frequency in the specific poly(A) region under study. For example, for *cis* elements identified in $-100/-41$, we used the nucleotide frequency of that region to fill gaps. Finally, the aligned hexamers were used to generate sequence logos by the Web Logo tool (Crooks et al. 2004).

Scoring matrices of *cis* elements and sequence search using the matrices

Aligned hexamers were used to generate position-specific scoring matrices (PSSM). For each position, the score was calcu-

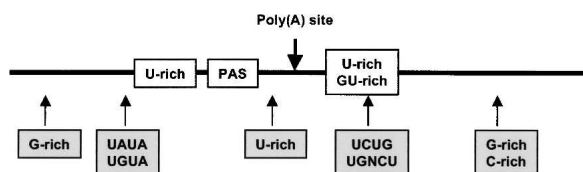


FIGURE 6. Schematic of *cis* elements for human poly(A) sites. (Gray boxes) Novel candidate *cis* elements identified in this study.

lated as follows: $S(n,p) = \log_2 (f(n,p) / f(n))$, (where $S(n,p)$ is the score for nucleotide n at position p , $f(n,p)$ is the frequency of occurrence of nucleotide n at the position p of the *cis* element, and $f(n)$ is the frequency of occurrence of nucleotide n in a specific poly(A) region, e.g., $-100/-41$). The matrices were used to search the $-125/+125$ region of all poly(A) sites using PERL. For a given region of a sequence with the length of a *cis* element, its score was the sum of individual scores at all nucleotide positions. Hits are sequences with positive scores. Results were plotted using program R. Lines in the graphs were smoothed by a moving window scheme, where the value of each position is the average of all values in a 7-n window centered at the position.

ACKNOWLEDGMENTS

We thank Michael Zhang and members of B.T. lab for helpful discussions of the methods. C.S.L. was supported by NSF award MCB-0426195 and NIH grant 1R03AR052038-01. J.W. was supported by NIH grant GM072481. B.T. was supported by The Foundation of UMDNJ.

Received May 11, 2005; accepted June 27, 2005.

REFERENCES

- Arhin, G.K., Boots, M., Bagga, P.S., Milcarek, C., and Wilusz, J. 2002. Downstream sequence elements with different affinities for the hnRNP H/H' protein influence the processing efficiency of mammalian polyadenylation signals. *Nucleic Acids Res.* **30**: 1842–1850.
- Bagga, P.S., Ford, L.P., Chen, F., and Wilusz, J. 1995. The G-rich auxiliary downstream element has distinct sequence and position requirements and mediates efficient 3' end pre-mRNA processing through a *trans*-acting factor. *Nucleic Acids Res.* **23**: 1625–1631.
- Bakheet, T., Frevel, M., Williams, B.R., Greer, W., and Khabar, K.S. 2001. ARED: Human AU-rich element-containing mRNA database reveals an unexpectedly diverse functional repertoire of encoded proteins. *Nucleic Acids Res.* **29**: 246–254.
- Beaudoing, E., Freier, S., Wyatt, J.R., Claverie, J.M., and Gautheret, D. 2000. Patterns of variant polyadenylation signal usage in human genes. *Genome Res.* **10**: 1001–1010.
- Brown, P.H., Tiley, L.S., and Cullen, B.R. 1991. Efficient polyadenylation within the human immunodeficiency virus type 1 long terminal repeat requires flanking U3-specific sequences. *J. Virol.* **65**: 3340–3343.
- Carswell, S. and Alwine, J.C. 1989. Efficiency of utilization of the simian virus 40 late polyadenylation site: Effects of upstream sequences. *Mol. Cell. Biol.* **9**: 4248–4258.
- Chen, C.Y. and Shyu, A.B. 1995. AU-rich elements: Characterization and importance in mRNA degradation. *Trends Biochem. Sci.* **20**: 465–470.

- Chen, F., MacDonald, C.C., and Wilusz, J. 1995. Cleavage site determinants in the mammalian polyadenylation signal. *Nucleic Acids Res.* **23**: 2614–2620.
- Chou, Z.F., Chen, F., and Wilusz, J. 1994. Sequence and position requirements for uridylate-rich downstream elements of polyadenylation signals. *Nucleic Acids Res.* **22**: 2525–2531.
- Colgan, D.F. and Manley, J.L. 1997. Mechanism and regulation of mRNA polyadenylation. *Genes & Dev.* **11**: 2755–2766.
- Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E. 2004. WebLogo: A sequence logo generator. *Genome Res.* **14**: 1188–1190.
- Edmonds, M. 2002. A history of poly A sequences: From formation to factors to function. *Prog. Nucleic Acid Res. Mol. Biol.* **71**: 285–389.
- Edwards-Gilbert, G., Prescott, J., and Falck-Pedersen, E. 1993. 3' RNA processing efficiency plays a primary role in generating termination-competent RNA polymerase II elongation complexes. *Mol. Cell. Biol.* **13**: 3472–3480.
- Fairbrother, W.G., Yeh, R.F., Sharp, P.A., and Burge, C.B. 2002. Predictive identification of exonic splicing enhancers in human genes. *Science* **297**: 1007–1013.
- 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., and Smith, T.F. 1999. In silico detection of control signals: mRNA 3'-end-processing sequences in diverse species. *Proc. Natl. Acad. Sci.* **96**: 14055–14060.
- Graveley, B.R., Fleming, E.S., and Gilmartin, G.M. 1996. RNA structure is a critical determinant of poly(A) site recognition by cleavage and polyadenylation specificity factor. *Mol. Cell. Biol.* **16**: 4942–4951.
- Hall-Pogar, T., Zhang, H., Tian, B., and Lutz, C.S. 2005. Alternative polyadenylation of cyclooxygenase-2. *Nucleic Acids Res.* **33**: 2565–2579.
- Jacobson, A. and Peltz, S.W. 1996. Interrelationships of the pathways of mRNA decay and translation in eukaryotic cells. *Annu. Rev. Biochem.* **65**: 693–739.
- Kaufmann, I., Martin, G., Friedlein, A., Langen, H., and Keller, W. 2004. Human Fip1 is a subunit of CPSF that binds to U-rich RNA elements and stimulates poly(A) polymerase. *EMBO J.* **23**: 616–626.
- Legendre, M. and Gautheret, D. 2003. Sequence determinants in human polyadenylation site selection. *BMC Genomics* **4**: 7.
- Levitt, N., Briggs, D., Gil, A., and Proudfoot, N.J. 1989. Definition of an efficient synthetic poly(A) site. *Genes & Dev.* **3**: 1019–1025.
- MacDonald, C.C., Wilusz, J., and Shenk, T. 1994. The 64-kilodalton subunit of the CstF polyadenylation factor binds to pre-mRNAs downstream of the cleavage site and influences cleavage site location. *Mol. Cell. Biol.* **14**: 6647–6654.
- Moreira, A., Wollerton, M., Monks, J., and Proudfoot, N.J. 1995. Upstream sequence elements enhance poly(A) site efficiency of the C2 complement gene and are phylogenetically conserved. *EMBO J.* **14**: 3809–3819.
- Natalizio, B.J., Muniz, L.C., Arhin, G.K., Wilusz, J., and Lutz, C.S. 2002. Upstream elements present in the 3'-untranslated region of collagen genes influence the processing efficiency of overlapping polyadenylation signals. *J. Biol. Chem.* **277**: 42733–42740.
- Perez Canadillas, J.M. and Varani, G. 2003. Recognition of GU-rich polyadenylation regulatory elements by human CstF-64 protein. *EMBO J.* **22**: 2821–2830.
- Ryner, L.C. and Manley, J.L. 1987. Requirements for accurate and efficient mRNA 3' end cleavage and polyadenylation of a simian virus 40 early pre-RNA in vitro. *Mol. Cell. Biol.* **7**: 495–503.
- Sachs, A.B., Sarnow, P., and Hentze, M.W. 1997. Starting at the beginning, middle, and end: Translation initiation in eukaryotes. *Cell* **89**: 831–838.
- Schbath, S. 1997. An efficient statistic to detect over- and under-represented words in DNA sequences. *J. Comput. Biol.* **4**: 189–192.
- Schneider, T.D. and Stephens, R.M. 1990. Sequence logos: A new way to display consensus sequences. *Nucleic Acids Res.* **18**: 6097–6100.
- Tabaska, J.E. and Zhang, M.Q. 1999. Detection of polyadenylation signals in human DNA sequences. *Gene* **231**: 77–86.
- 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**: 201–212.
- Valsamakias, A., Zeichner, S., Carswell, S., and Alwine, J.C. 1991. The human immunodeficiency virus type 1 polyadenylation signal: A 3' long terminal repeat element upstream of the AAUAAA necessary for efficient polyadenylation. *Proc. Natl. Acad. Sci.* **88**: 2108–2112.
- Valsamakias, A., Schek, N., and Alwine, J.C. 1992. Elements upstream of the AAUAAA within the human immunodeficiency virus polyadenylation signal are required for efficient polyadenylation in vitro. *Mol. Cell. Biol.* **12**: 3699–3705.
- 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**: 1315–1327.
- Wickens, M., Anderson, P., and Jackson, R.J. 1997. Life and death in the cytoplasm: Messages from the 3' end. *Curr. Opin. Genet. Dev.* **7**: 220–232.
- Wilusz, J. and Shenk, T. 1988. A 64 kd nuclear protein binds to RNA segments that include the AAUAAA polyadenylation motif. *Cell* **52**: 221–228.
- Wu, C. and Alwine, J.C. 2004. Secondary structure as a functional feature in the downstream region of mammalian polyadenylation signals. *Mol. Cell. Biol.* **24**: 2789–2796.
- Yonaha, M. and Proudfoot, N.J. 2000. Transcriptional termination and coupled polyadenylation in vitro. *EMBO J.* **19**: 3770–3777.
- Zarudnaya, M.I., Kolomiets, I.M., Potyahaylo, A.L., and Hovorun, D.M. 2003. Downstream elements of mammalian pre-mRNA polyadenylation signals: Primary, secondary and higher-order structures. *Nucleic Acids Res.* **31**: 1375–1386.
- Zhang, H., Hu, J., Recce, M., and Tian, B. 2005. PolyA_DB: A database for mammalian mRNA polyadenylation. *Nucleic Acids Res.* **33** (Database Issue): D116–D120.
- 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**: 405–445.



RNA

A PUBLICATION OF THE RNA SOCIETY

Bioinformatic identification of candidate *cis*-regulatory elements involved in human mRNA polyadenylation

JUN HU, CAROL S. LUTZ, JEFFREY WILUSZ, et al.

RNA 2005 11: 1485-1493

References

This article cites 42 articles, 20 of which can be accessed free at:
<http://rnajournal.cshlp.org/content/11/10/1485.full.html#ref-list-1>

License

Email Alerting Service

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#).

Dharmacon[™] Reagents
Custom synthesis, RNAi, and CRISPR solutions

Infinite Reliability

More

horizon
a PerkinElmer company

To subscribe to *RNA* go to:
<http://rnajournal.cshlp.org/subscriptions>
