Sequence determinants in human polyadenylation site selection.

Matthieu Legendre, Daniel Gautheret

BMC Genomics

BioMed Central

Research article

Sequence determinants in human polyadenylation site selection

Matthieu Legendre and Daniel Gautheret*

Address: INSERM ERM-206, Luminy Case 906, 13288 Marseille Cedex 09, France

Email: Matthieu Legendre - legendre@tagc.univ-mrs.fr; Daniel Gautheret* - gautheret@esil.univ-mrs.fr

* Corresponding author

Published: 25 February 2003

BMC Genomics 2003, **4**:7

Received: 10 December 2002

Accepted: 25 February 2003

This article is available from: http://www.biomedcentral.com/1471-2164/4/7

Abstract

Background: Differential polyadenylation is a widespread mechanism in higher eukaryotes producing mRNAs with different 3' ends in different contexts. This involves several alternative polyadenylation sites in the 3' UTR, each with its specific strength. Here, we analyze the vicinity of human polyadenylation signals in search of patterns that would help discriminate strong and weak polyadenylation sites, or true sites from randomly occurring signals.

Results: We used human genomic sequences to retrieve the region downstream of polyadenylation signals, usually absent from cDNA or mRNA databases. Analyzing 4956 EST-validated polyadenylation sites and their -300/+300 nt flanking regions, we clearly visualized the upstream (USE) and downstream (DSE) sequence elements, both characterized by U-rich (not GU-rich) segments. The presence of a USE and a DSE is the main feature distinguishing true polyadenylation sites from randomly occurring A(A/U)UAAA hexamers. While USEs are indifferently associated with strong and weak poly(A) sites, DSEs are more conspicuous near strong poly(A) sites. We then used the region encompassing the hexamer and DSE as a training set for poly(A) site identification by the ERPIN program and achieved a prediction specificity of 69 to 85% for a sensitivity of 56%.

Conclusion: The availability of complete genomes and large EST sequence databases now permit large-scale observation of polyadenylation sites. Both U-rich sequences flanking both sides of poly(A) signals contribute to the definition of "true" sites. However, the downstream U-rich sequences may also play an enhancing role. Based on this information, poly(A) site prediction accuracy was moderately but consistently improved compared to the best previously available algorithm.

Background

Alternative splicing, alternative transcription initiation and alternative polyadenylation are three mechanisms through which a variety of transcripts can be synthesized from a single eukaryotic gene. Beside the sheer number of genes, the combination of these mechanisms largely contributes to transcript diversity in complex genomes such as those of mammalians. Alternative polyadenylation occurs when two or more polyadenylation sites are present in the 3' untranslated region (UTR) of an mRNA, producing transcript isoforms of variable lengths. At least 22% of mRNAs [1] undergo alternative polyadenylation, often in a tissue- or time-specific manner [2]. Since 3' UTRs may host important regulatory elements – for instance affecting stability, localization or translation – alternative polyadenylation may strongly affect the fate of mRNA and therefore gene function. Although critical, the mechanisms of polyadenylation site selection are not yet fully

understood. In particular, it is not clear how the polyadenylation machinery is able to distinguish bona fide polyadenylation signals from the multiple look-alikes interspersed along the 3' UTRs.

Polyadenylation sites are primarily defined by a hexameric polyadenylation signal (PAS) of sequence AAUAAA or a one-base variant, located about 15 bases upstream of the cleavage site. This central motif can be flanked by optional upstream and downstream sequence elements (USE and DSE). The DSE is described as a U-rich [3], or GU-rich element located 20–40 bases 3' to the cleavage site ([4,5] for reviews). It is present in a large proportion of genes, and its deletion has been shown to suppress polyadenylation [6]. The presence of a USE has been described in fewer cases: in viruses [5] and in four human genes [7–9]. The position and sequence of the USE are poorly defined, although U-richness is also suspected. EST counts suggest that, when several polyadenylation sites are present in the same UTR, the distal site is generally the most efficient [1]. However, polyadenylation efficiency does not strictly depend on the hexameric signal: "strong" sites may contain variant AGUAAA signals while "weak" sites may contain canonical AAUAAA hexamers. It is therefore suspected that sequence determinants affecting polyadenylation efficiency may lie in the flanking USE and DSE.

We used bioinformatics analysis of EST and genomic sequences to characterize biases in the regions encompassing 600 nucleotides around the cleavage site. Correlations between poly(A) site efficiency and sequence biases in flanking regions were identified. In addition, we observed that sequences in a downstream region broader than the DSE discriminate actual poly(A) sites from randomly occurring AAUAAA hexamers. We exploited this information in a computer program for polyadenylation site detection that presents a better specificity/sensitivity ratio than previous algorithms.

Results & Discussion

Out of 13680 UTR sequences analyzed, 4956 contained at least one EST-supported polyadenylation site (see Methods for definition of "EST-supported"). Considering that UTRs may contain several EST-supported polyadenylation sites, the total number of sites was 6563. For each site, flanking sequences of 300 nt in both 5' and 3' direction were retrieved in the human genome, possibly encompassing regions in the last intron or past the last poly(A) site, into the cleaved part of the primary transcript (Figure 1). This last point is worth noting, since previous analyses based on cDNA or EST sequences had limited access to sequences 3' of the poly(A) site. Not all UTR sequences had a match in the human genome that was both reliable and long enough to cover the 600 nt region. We finally obtained 5069 extended sequences, which we will henceforth call "terminal sequences" (Figure 1).

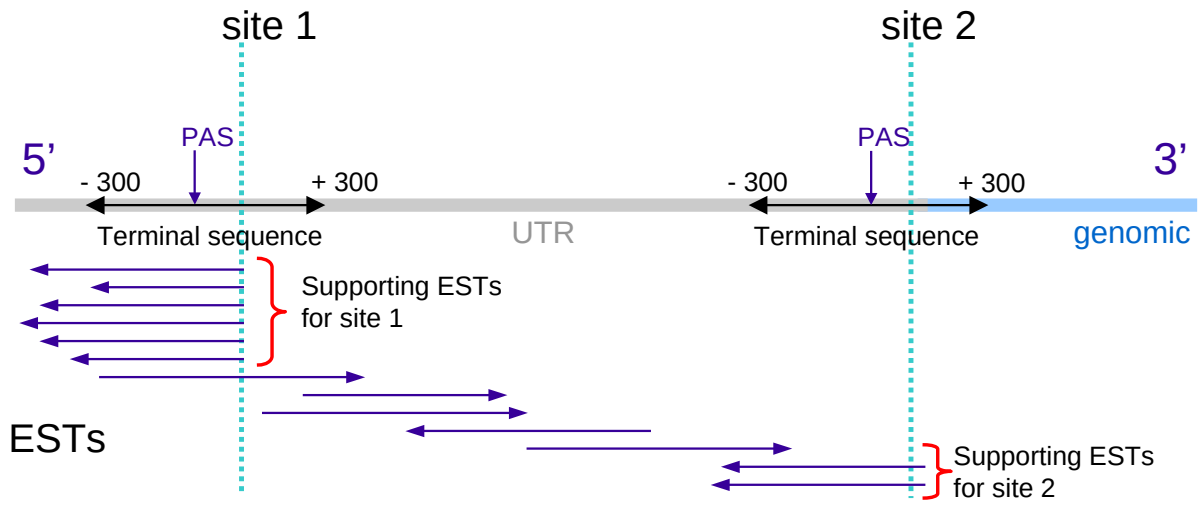

Figure 1
Schematic view of EST-based polyadenylation site identification. Each UTR is aligned onto the complete EST database. A poly(A) site is validated when at least two ESTs match this site while respecting specific length, position and quality criteria (see Methods). The -300/+300 nt fragment surrounding each site (here called "terminal sequence") is then extracted for further analysis. For sites located near the 3' end of the UTR, we use the corresponding genomic sequence to complete the terminal sequence.

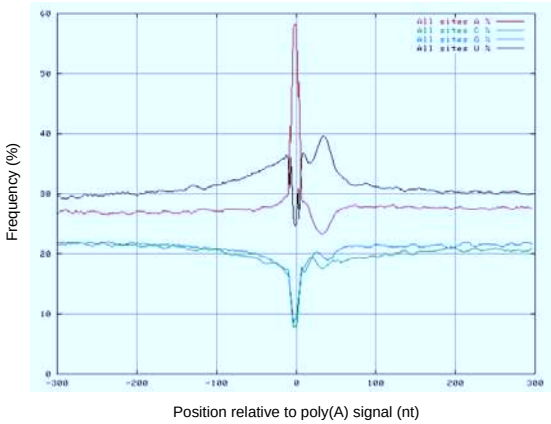

Figure 2
Nucleotide composition in terminal sequences. Position 0 corresponds to the 3' base of the polyadenylation signal. Nucleotide positions were averaged in a sliding 11 nt window.

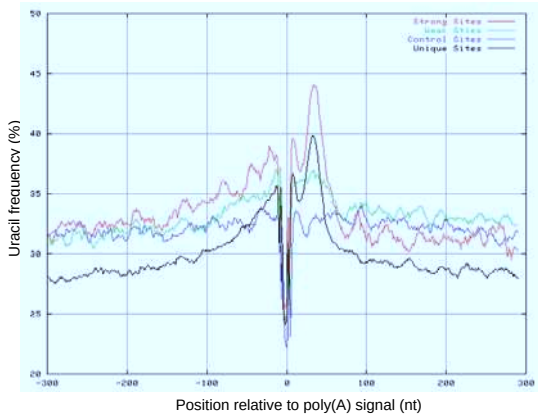

Figure 3
Uracil frequencies in a 11 nt window, in the vicinity of "strong" poly(A) sites (645 sequences), "weak" sites (1200 sequences), "unique" sites (3776 sequences) and controls (1249 sequences).

For genes with multiple polyadenylation sites, each site was classified into one of two categories according to its putative polyadenylation efficiency, as measured by relative EST counts (see Methods). "Strong" sites were those supported by more than 70% of the ESTs for this gene (645 sites), while "weak" sites were those supported by less than 30% of the ESTs (1200 sites). Genes with less than 10 supporting ESTs were ignored for this classification. Let us emphasize here that EST counts do not accurately reflect polyadenylation efficiency for any specific gene, since the observed poly(A) variants may come from different EST libraries with different sizes and amplification protocols. Here we used only non-normalized libraries and, since we made observations over a large number of genes, the biases observed for specific genes should neutralize each other and allow for general tendencies to emerge.

Two further groups of poly(A) sites were built: "unique" sites from UTR with a single poly(A) site (3776 sites), and "control" sites from AAUAAA signals for which no EST was ever observed (1249 sites).

Figure 2 shows base composition in the terminal sequence for all site types, averaged over a window of 11 nt. The Adenine peak at position 0 corresponds to the A-rich polyadenylation signal. There is a visible rise in Uracil frequency 5' of the signal, which could correspond to an Upstream Sequence Element. For simplification purposes, we will call USE this U-rich region although some authors have used the term differently. Increase in Uracil frequency is yet more significant in the 3' region, with a peak at +15/+30 nt from the cleavage site (or +30/+45 from the poly(A) signal). This corresponds to the DSE, as described by Chou et al [10], at +15/+25 from the cleavage site. The rise in %U at the DSE is closely mirrored by a decline in %A, while the G and C curves remain roughly flat.

A first consequence of Figure 2 is that the DSE is generally not a GU-rich region as often stated, but rather a U-rich region. A second concerns the importance of the USE element: although less conspicuous than the 3' element, upstream U-rich elements are certainly present in a large number of mRNAs, even if few USEs have been documented to date.

We next questioned whether the USE and DSE were specifically associated to strong *vs*. weak sites, or active *vs*. inactive sites. Figure 3 shows fluctuations of Uracil composition (U%) in a 11-nt window around different types of polyadenylation sites: "strong", "weak", defined as above for sequences having two or more sites, as well as "unique" (single-site UTR) and "control" (silent AAUAAA hexamers) sites. USE and DSE elements flank all types of true poly(A) sites, while they are essentially absent in the vicinity of control sites. This suggests that the presence of a USE and/or DSE distinguishes *bona fide* polyadenylation sites from randomly occurring AAUAAA hexamers.

In the DSE, U-richness is significantly higher for "strong" or "unique" sites, than for "weak" sites. Student's test P value for U frequencies in the +20/+60 region is $2.8\ 10^{-10}$

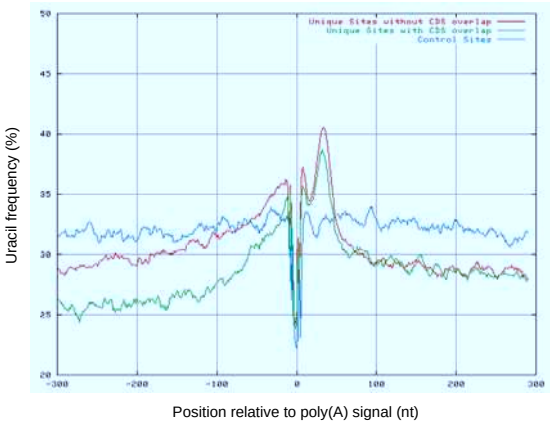

Figure 4
Uracil frequencies in a 11 nt window, in the vicinity of "control" sites and two types of "unique" poly(A) sites: those located less than 300 nt from the Stop codon (CDS overlap: 1328 sequences), and those located more than 300 nt from the Stop codon (no CDS overlap: 2448 sequences).

for "strong" *vs*. "weak" and 2.2 10^{-16} for "weak" *vs*. "control". This suggests that the U-rich element in the +20/+60 region may act as an enhancer and help distinguish the most efficient sites in case of alternative polyadenylation. There is no such obvious correlation in the USE, which does not vary as strongly between "strong" and "weak" sites. Therefore, the USE is not as strongly linked to processing efficiency as the DSE.

Unique sites also display a relatively low U% background in both the 5' and 3' region, compared to alternatively polyadenylated sites or control sites (Figure 3). We suspected that the composition in the 5' region could be affected by the proximity of the coding sequence, unique sites being, in average, closer to the coding sequence (35% of unique sites are within 300 bp of the stop codon, while only 21% of non-unique sites are in this range). We therefore distinguished those sites occurring less than 300 nt from the Stop codon from those occurring at a larger distance (shown for unique sites, Figure 4). Sites with a 300 nt upstream region overlapping the coding sequence clearly have a lower U content in 5', in agreement with the generally higher G+C content of the coding region. Therefore, the overall lower U% upstream of unique poly(A) sites in the -300 to -100 region can be attributed to the proximity of coding sequences. As for the low U% in the downstream region, between position +80 and +300 (Figure 3), it is an interesting specificity of unique sites that is worth noting.

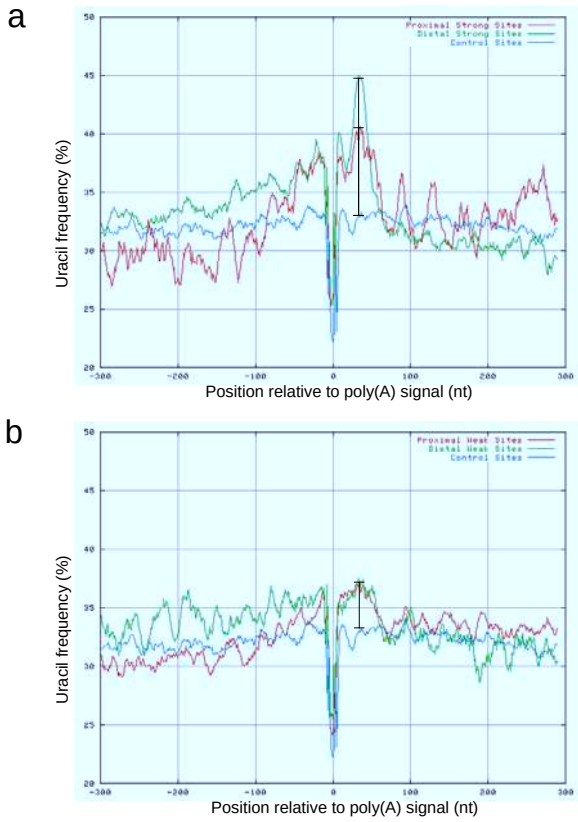

Figure 5
Uracil frequencies in a 11 nt window in the vicinity of alternative poly(A) sites, distinguishing proximal sites from distal sites. (a) : "strong" poly(A) sites (129 proximal, 499 distal); (b) : "weak" poly(A) sites (655 proximal, 210 distal).

We showed previously that, in UTRs with multiple sites, the strongest poly(A) sites were often the most distal ones [1]. We thus questioned whether the apparent "strong site" characteristics in Figure 3 could be associated instead to distal sites, independently of their strength. Figure 5 shows U% variations in distal *vs*. proximal polyadenylation sites, for strong (a) and weak sites (b). Due to the small number of sites considered in some cases (especially proximal/strong and distal/weak), the corresponding average curves are somewhat jaggy but, in any case, strong proximal sites do not differ significantly from strong distal sites in the DSE region (Figure 5.a: T-test P value = 0.09). However, a higher %U peak in the DSE region is definitely characteristic to strong sites, independently of their position in the UTR. Although both strong and weak sites display a significant uracil rise in the DSE (T-test P values <

10^{-15} for either strong or weak sites *vs*. control sites, Fig. 5.a and 5.b), the difference between strong and weak sites is also highly significant (T-test P value = 8.0 10^{-10} for proximal strong sites vs. proximal weak sites). As for the upstream region, although %U level in the USE is consistently higher than background in strong poly(A) sites, it is also higher in weak sites, suggesting again that the 5' bias occurs in both classes of sites.

We tried to further characterize the sequence bias in the USE and DSE regions using word count and Gibbs sampling algorithms. Although these methods are able to identify any k-letter words that is significantly enriched in a specific sequence set, they failed to identify any recurrent motif in either the USE or DSE region, whatever subset of sequence was considered: strong, weak, distal, proximal, etc. (data not shown). Therefore, there is no specific sequence motif associated with the general USE or DSE element. For the DSE, an absence of determined sequence motif is consistent with previous experiments showing that point mutations in this region had no effect on polyadenylation efficiency [11]. However, our analysis does not exclude the presence of gene-specific motifs in certain USEs or DSEs.

Polyadenylation site detection

One of the main difficulties of eukaryotic gene annotation is the delineation of the first and last exons, mainly because they lack the splicing and codon usage constraints of internal exons. Recently, there has been some improvement in the detection of 5' exons, thanks to an improved recognition of CpG islands and promoter sequences [12]. However, gene annotation programs still use crude approaches to detect the 3' boundary of the last exon. In Genscan [13], 3' end detection is done by a simple search for A(A/U)UAAA (AWUAAA) hexamers. This captures about 85% of the true poly(A) sites, but also finds about one false positive per kilobase in a database of randomized UTR sequences (Table 2). This is a high rate if one considers that 3' UTRs often span several kilobases. To reduce levels of false positives, Tabaska & Zhang (1999) [14] have developed a quadratic discriminant function that analyzes both the poly(A) signal and sequence biases in the DSE region. In our test conditions, their program, POLYADQ, detects only about 55% of the true poly(A) sites, but finds 4 to 5 times fewer false positives than a simple hexamer count, which is a considerable gain in specificity.

Table 1: Measure of True Positives (TP), False Negatives (FN) and Sensitivity (SN) in the prediction of polyadenylation signals by the POLYADQ and ERPIN programs, based on a dataset of 982 annotated UTR sequences from the EMBL database. See Methods for information on database construction. ERPIN parameters were adjusted to match the sensitivity of POLYADQ.

Program	TP	FN	Sensitivity SN
Erpin	549	433	55.9 %
Polyadq	547	435	55.7 %

Table 2: Negative predictions and accuracy of the ERPIN and POLYADQ program, evaluated for different control sequences not containing polyadenylation sites: coding sequences (CDS), introns, and two types of randomized UTR sequences: simple shuffling or first order Markov simulation.

Negative set	AWUAAA per 100 kb	Program	TN	FP	FP per 100 kb	Specificity SP	Accuracy CC
CDS	31.2	Erpin	880	102	3.7	84.33 %	0.483
		Polyadq	862	120	3.8	82.01 %	0.459
Introns	156.4	Erpin	741	241	38.9	69.49 %	0.320
		Polyadq	718	264	42.0	67.45 %	0.293
UTR shuffled	109.6	Erpin	888	94	11.0	85.38 %	0.494
		Polyadq	826	156	17.4	77.81 %	0.415
UTR Markov 1st order	94.49	Erpin	772	210	21.9	72.33 %	0.354
		Polyadq	733	249	23.9	68.72 %	0.309

See Methods for information on database construction. Each row shows the number of potential A(A/U)UAAA signals per 100 kb in the dataset, True Negatives (TN), False Positives (FP), False Positives per 100 kb, Specificity (SP) and Accuracy (CC). Calculation of CC uses TP and TN from Table 1.

Table 3: Compared accuracy of the ERPIN and POLYADQ programs for the prediction of EMBL annotated poly(A) sites, and of EST-derived weak poly(A) sites identified in this study. The negative set for SP and CC calculation is "UTR shuffled" from Table 2.

Program	Prediction of:	Sensitivity SN	Specificity SP	Accuracy CC
Erpin	EMBL annotated sites	55.91 %	85.38 %	0.494
Polyadq		55.70 %	77.81 %	0.415
Erpin	Weak sites	31.28 %	80.21 %	0.262
Polyadq		30.54 %	70.45 %	0.171

Having a larger collection of EST-validated poly(A) sites and a better knowledge of upstream and downstream regions discriminating true poly(A) sites from random AWUAAA hexamers, we asked whether this could be exploited to further improve poly(A) site detection. We used the 2327 "strong" and "unique" terminal sequences as a training set for the ERPIN program. ERPIN [15] is an extension of the classical weight matrix representation of sequence alignments, adapted to RNA sequences containing base-paired regions. From an RNA alignment, the program creates weight matrices corresponding to different parts of the molecule, as defined by users, and finds all sequences matching these matrices above a certain score threshold. Matrices can be defined for each helix, gapped single strand or ungapped single strand. Here we used only ungapped single strands, which are modelled by ERPIN using dinucleotide frequencies (see Methods). Dinucleotides are usually better than mononucleotides at capturing sequence biases in non-coding DNA or RNA sequences. We then tested different search strategies, varying the size and position of the regions used to define the weight matrices. In spite of the significant sequence bias in the USE region, no upstream region was beneficial to search accuracy. The highest accuracy was obtained using both the hexameric signal and the 0 to +46 nt downstream region. Table 1 and 2 show results obtained using this region, compared to POLYADQ results. To facilitate comparisons, ERPIN was calibrated to achieve the same sensitivity as POLYADQ on a test set of 982 annotated UTRs (Table 1). Then, the programs were compared for their ability to filter out false positives (Table 2).

Our control sets were sequences known not to contain poly(A) signals: coding sequences, introns and randomized UTRs. Each AWUAAA hexamer in these sequences was considered as a negative site. This AWUAAA background varies significantly between CDS (31 sites / 100 kb), randomized UTRs (about 100 sites / 100 kb) and introns (156 sites / 100 kb). We thus expected more false positives per kilobase in introns or UTRs than in CDS, and this is indeed what we observed, for both POLYADQ and ERPIN. However, both programs were more efficient in filtering out false positives in CDS or shuffled UTRs, than in 1st order Markov model UTRs or introns. For instance, there were only 3.7 false positives per 100 kb in CDS, instead of about 40 / 100 kb in introns (Table 2). The good performance on shuffled UTRs relative to 1st order Markov models may be due to dinucleotide biases in the downstream region that are lost after shuffling, although such biases do not appear very significant in dinucleotide counts (data not shown). Poly(A) site prediction is worst in intronic and 1st order Markov model UTRs, where about one out of four AWUAAA sites is predicted as true (specificity around 70%). However, we note a moderate but consistent gain in overall accuracy from POLYADQ to ERPIN. The smaller gain is for CDS sequences (0.483 instead of 0.459) and the largest in shuffled UTR sequences (0.494 instead of 0.415). This gain is surprising if one considers the simplicity of the ERPIN algorithm for ungapped single strands (based on a simple dinucleotide weight matrix) compared to POLYADQ.

A useful application of improved predictions would be in the detection of weak or cryptic poly(A) sites. Table 3 shows the prediction quality obtained for annotated poly(A) sites in the EMBL databank (numbers taken from Table 2) and for weak sites, as defined above based on EST counts. For both programs, false positives and true negatives were computed in the "UTR shuffled" dataset (Table 2). For both programs, prediction sensitivity for weak sites is only about about 30%, which is explained by the absence of a strong DSE. However, it is noteworthy that ERPIN provides a 53% accuracy improvement over POLYADQ, through a better filtering of false positives. Using a specific "weak sites" training set in lieu of our "strong sites" set did not further improve ERPIN predictions, indicating that there are no common features of weak sites that can be exploited in a weight matrix approach.

Conclusions

A large fraction of human polyadenylation sites are flanked by U-rich elements, both upstream (USE) and downstream (DSE) of the cleavage site, located around positions 0 to -50 and +20 to +60, relative to the poly(A) signal. USE and DSE clearly distinguish true polyadenylation sites from randomly occurring AAUAAA hexamers.

While the USE is not specifically associated with "efficient" poly(A) sites, a "U-rich" DSE may act as an enhancer specifying the most efficient sites in case of alternative polyadenylation. For the purpose of predicting poly(A) signals in genomic sequences, the most discriminating region includes the poly(A) signal and the 46 nt downstream sequence. Using the ERPIN program and a simple dinucleotide weight matrix description, we were able to improve the accuracy of poly(A) site predictions by 5% to 19% relative to POLYADQ, depending on sequence context. Such a gain can be helpful in annotation projects aiming at a better characterization of 3' non coding sequences.

Methods

Terminal sequence extraction

Polyadenylation sites were identified using the EST-Parser program (Beaudoing et al. 2000 [16]), based on the mapping of ESTs to UTR sequences. When at least two ESTs finished at the same position and less than 36 bases from a potential PAS (AAUAAA or one of 11 variants), a hypothetic cleavage site was considered as valid. Moreover, in the case of multiple polyadenylation sites, the numbers of ESTs observed at each site were taken as a measure of relative polyadenylation efficiency. For this measure, only non-normalized and non-subtracted EST libraries were considered. The EST database was dbEST (October 2001 release) and the UTR database was UTRdb release 13 [17]. The procedure identified 6563 distinct Polyadenylation sites from 4956 3' UTRs.

The next task was to obtain the +/-300 nt region around each PAS. This region usually goes past the boundaries of the UTR, either because the PAS is too close to the 3' extremity of the UTR (most frequent case) or because it is too close from the Stop codon. In order to retrieve the 5' or 3' genomic regions, UTRs were aligned on the human genome working draft (NCBI Oct-2001 version) with the BLAST program. We retained alignments meeting the following criteria: length > 60 nt; E-value < 0.001; identity > 98%; dangling ends of less than 10 nt in both directions. The complete UTR and 300 nt flanking regions was then extracted from the genomic sequence, producing 5110 "extended UTR" sequences. Then, for each polyadenylation site in a UTR, the +/-300 nt region around the PAS was extracted. This region is referred to as a "terminal sequence" hereafter.

Polyadenylation site classification

For UTRs with alternative polyadenylation, sites were classified as follows (any site may belong to more than one category):

• *Strong sites* (645 terminal sequences): sites in UTR with at least 10 matching ESTs, more than 70% of which are associated to this site.

• *Weak sites* (1200 terminal sequences): sites in UTR with at least 10 matching ESTs, less than 30% of which are associated to this site.

Strong and weak sites were further subdivided into "distal" or "proximal", according to their position in the UTR. In each UTR, the 3'-most site was said "distal" and the 5'-most site was said "proximal". Numbers of terminal sequences were as follows. Strong proximal sites: 129; Strong distal sites: 499; Weak proximal sites: 655; Weak distal sites: 210.

We then defined as "Unique" sites (3776 terminal sequences) those sites from UTRs with a single EST-supported poly(A) site. Unique sites were further distinguished according to their proximity to the Stop codon: sites within 300 nt of Stop codon (1328 terminal sequences) and sites at distance 300 nt or more from Stop codon (2448 terminal sequences).

Finally, we defined "Control" sites (1249 terminal sequences) as regions containing an AATAAA hexamer and no matching EST within 60 nt around the hexamer. These controls did not include any AAUAAA hexamer located within 50 nt of the 3' end of the UTR, to avoid real sites not covered by ESTs.

Nucleotide composition analysis

Position-dependent compositions in the terminal sequences were measured in an 11 nucleotide sliding window, advancing by one nucleotide steps over the 600 nt region.

Erpin runs

We used Erpin version 3.1 http://tagc.univ-mrs.fr/pub/erpin/. The training set was made of 2,327,600 nt terminal sequences (1632 "unique" sites + 695 "strong" sites) and is available at the same location. Optimal search parameters were determined empirically. The best results were obtained when searching first for the hexameric PAS with a score cutoff of 70%, and then searching for the 46 nt region immediately downstream of the PAS with a score cutoff of 74%. Score cutoffs are expressed in percentage of training set sequences retained. A 70% cutoff for the hexamer region amounts to retaining either AAUAAA or AUUAAA and rejecting any other variant. Searched regions were defined as ungapped single strands. Therefore, their weight matrices were computed by ERPIN using a lod-score for each pair of consecutive bases at positions i and i+1 [15]:

$$S_{i,i+1} = \log\left(\frac{O_{i,i+1}}{E_i \times E_{i+1}} \right)$$

Where Oi,i+1 is the observed frequency for the pair of consecutive bases at position i and i+1, and E_i, E_{i+1} the expected frequencies of individual bases. All searches were conducted using the "-unifstat" option, which sets all expected base frequencies to 0.25, thus reducing the number of false positives in GC-rich regions.

Measure of site detection accuracy

True Positives (TP) and False Negative (FN)

Computational detection of "real" poly(A) sites was evaluated based on 982 EMBL sequence fragments containing 2000 nt upstream and 200 nt downstream of an annotated PAS (total: 2.16 Mb). For genes with multiple annotated PAS, only the most distal one was retained. We ensured that none of the training set sequences above was present in this database. TP was then measured as the number of annotated signals detected by the program and FN as the number of annotated signals undetected (Table 1). ERPIN parameters were adjusted so that TP and FN were nearly the same as with POLYADQ, which explains the nearly identical sensitivities in Table 1.

True Negative (TN) and False Positives (FP)

Mispredictions were evaluated on four distinct databases not expected to contain polyadenylation sites (Table 2), the size of which was adjusted so that the number of false sites (TN+FP) was the same as the number of true sites (TP+FN). The following databases were used: CDS sequences (4448 seq, 3.15 Mb) extracted from Genbank release 127; intronic sequences (154 seq, 628 kb) of the first intron from Genbank release 127; randomized UTR sequences (551 seq, 896 kb) of same mononucleotide composition as human 3' UTRs; and randomized UTR sequence (699 seq, 1.04 Mb) of same 1st order Markov model as human 3' UTRs. FP is the number of signals detected in those sequences and TN the number of AAUAAA or AUUAAA signals not detected in these sequences.

Predictive accuracy was then measured as follows:

$$\text{Sensitivity}: \ SN = \frac{TP}{TP + FN}$$

$$\text{Specificity}: \ SP = \frac{TP}{TP + FP}$$

$$\text{Accuracy}: \ CC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

For Table 3, we compared polyadenylation site prediction in the 982 annotated EMBL sequences above to that obtained in 848 complete UTR sequences (1.8 Mb) containing at least one "weak" poly(A) site, defined as above based on EST counts.

Author's contributions

ML developed the computer scripts and conducted the statistical analyzes. DG directed the study and drafted the manuscript.

Acknowledgements

We thank Dr. Philippe Benech and Dr. Pascal Hingamp for their detailed reading of the manuscript and useful suggestions.

References

1. Beaudoing E, Freier S, Wyatt JR, Claverie JM and Gautheret D **Patterns of variant polyadenylation signal usage in human genes** *Genome Res* 2000, **10:**1001-1010
2. Edwalds-Gilbert G, Veraldi KL and Milcarek C **Alternative poly(A) site selection in complex transcription units: means to an end?** *Nucleic Acids Res* 1997, **25:**2547-2561
3. Chen F, MacDonald CC and Wilusz J **Cleavage site determinants in the mammalian polyadenylation signal** *Nucleic Acids Res* 1995, **23:**2614-2620
4. Proudfoot N **Poly(A) signals** *Cell* 1991, **64:**671-674
5. Colgan DF and Manley JL **Mechanism and regulation of mRNA polyadenylation** *Genes Dev* 1997, **11:**2755-2766
6. Zhao J, Hyman L and Moore C **Formation of mRNA 3' ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis** *Microbiol Mol Biol Rev* 1999, **63:**405-445
7. Moreira A, Takagaki Y, Brackenridge S, Wollerton M, Manley JL and Proudfoot NJ **The upstream sequence element of the C2 complement poly(A) signal activates mRNA 3' end formation by two distinct mechanisms** *Genes Dev* 1998, **12:**2522-2534
8. Brackenridge S and Proudfoot NJ **Recruitment of a basal polyadenylation factor by the upstream sequence element of the human lamin B2 polyadenylation signal** *Mol Cell Biol* 2000, **20:**2660-2669
9. Aissouni Y, Perez C, Calmels B and Benech PD **The cleavage/polyadenylation activity triggered by a U-rich motif sequence is differently required depending on the poly(A) site location at either the first or last 3'-terminal exon of the 2'-5' oligo(A) synthetase gene** *J Biol Chem* 2002, **277:**35808-35814
10. Chou ZF, Chen F and Wilusz J **Sequence and position requirements for uridylate-rich downstream elements of polyadenylation signals** *Nucleic Acids Res* 1994, **22:**2525-2531
11. Zarkower D and Wickens M **A functionally redundant downstream sequence in SV40 late pre-mRNA is required for mRNA 3'-end formation and for assembly of a precleavage complex in vitro** *J Biol Chem* 1988, **263:**5780-5788
12. Davuluri RV, Grosse I and Zhang MQ **Computational identification of promoters and first exons in the human genome** *Nat Genet* 2001, **29:**412-417
13. Burge C and Karlin S **Prediction of complete gene structures in human genomic DNA** *J Mol Biol* 1997, **268:**78-94
14. Tabaska JE and Zhang MQ **Detection of polyadenylation signals in human DNA sequences** *Gene* 1999, **231:**77-86
15. Gautheret D and Lambert A **Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles** *J Mol Biol* 2001, **313:**1003-1011

16. Beaudoing E and Gautheret D **Identification of alternate polyadenylation sites and analysis of their tissue distribution using EST data** *Genome Res* 2001, **11:**1520-1526
17. Pesole G, Liuni S, Grillo G, Licciulli F, Larizza A, Makalowski W and Saccone C **UTRdb and UTRsite: specialized databases of sequences and functional elements of 5' and 3' untranslated regions of eukaryotic mRNAs** *Nucleic Acids Res* 2000, **28:**193-196