Computational Analysis of 3’ Ends of ESTs Shows Four Classes of Alternative Polyadenylation in Human, Mouse, and Rat

Jun Yan¹ and Thomas G. Marr¹,²

¹ Institute of Arctic Biology
² Arctic Region Supercomputing Center

311 Irving 1
University of Alaska
Fairbanks, AK, USA
99775-7000

Corresponding authors: Jun Yan (E-mail: fsjy1@uaf.edu) and Tom Marr (E-mail: fftgm@uaf.edu). Tel: 907.474.5673. Fax: 907.450.8601.
Abstract

Alternative initiation, splicing, and polyadenylation are key mechanisms used by many organisms to generate diversity among mature mRNA transcripts originating from the same transcription unit. While previous computational analyses of alternative polyadenylation have focused on polyadenylation activities within or downstream of the normal 3' terminal exons, we present the results of the first genome-wide analysis of patterns of alternative polyadenylation in the human, mouse, and rat genomes occurring over the entire transcribed regions of mRNAs using 3' ESTs with polyA tails aligned to genomic sequences. Four distinct classes of patterns of alternative polyadenylation result from this analysis: tandem polyA sites, composite exons, hidden exons, and truncated exons. We estimate that at least 49% (human), 31% (mouse), and 28% (rat) of polyadenylated transcription units have alternative polyadenylation. A portion of these alternative polyadenylation events result in new protein isoforms.

Supplemental materials are available at website: http://physics.nyu.edu/~jy272/altA.
Introduction

It is well known that polyadenylation at different positions along pre-mRNAs leads to mature mRNAs with different 3' untranslated regions (UTRs). 3' UTRs have been shown to contain regulatory elements controlling mRNA stability (Touriol et al. 1999, Dreyfus and Regnier 2002), translational efficiency (Knirsch et al. 2000), and intracellular localization (Kislauskis et al. 1994). Reviews of the known molecular mechanisms involved in 3' end formation in eukaryotes can be found in Zhao et al. (1999). Alternative initiation, splicing, and polyadenylation are key mechanisms used by organisms to generate diversity among mature mRNA transcripts originating from the same transcription unit. Recent studies have shown that mRNA processing, including capping, splicing, and polyadenylation, occur co-transcriptionally and that these are highly coupled and perhaps co-regulated reactions involving both cis- and trans-acting elements (Dye and Proudfoot 2001, Proudfoot et al. 2002, Levine and Tjian 2003, Kornblihtt et al. 2004).

Edwalds-Gilbert et al. (1997) first surveyed the recurring patterns of alternative polyadenylation based on the experimental literature. In this survey, three distinct patterns of alternative polyadenylation were noted: tandem polyA sites, composite exons, and skipped exons. Exons are generally categorized as 5' terminal exon, internal exons, or 3' terminal exon. In tandem polyA sites, multiple polyA sites are found within the same 3' terminal exon. In composite exons, 5' splice sites can sometime be silent, causing them to behave as 3' terminal exons, or sometime be active, thereby causing them to behave as internal exons. In skipped exons, either the first alternative 3' terminal exon is used, or that exon is skipped entirely and the second 3' terminal exon is spliced into the transcript. Since this paper, there have been several computational studies on alternative polyadenylation, mostly using Expressed Sequence Tags (ESTs). Gautheret et al. (1998) found alternative polyadenylation in 189 out of 1,000 human EST clusters. Aligning ESTs to the 3' UTRs from UTRdb database (Pesole et al. 2000), Beaudoin et al. (2000) found that 29% of human genes were alternatively polyadenylated. Beaudoin and Gautheret (2001) detected the biases among the EST libraries in the alternative polyadenylation for 1,450 human and 200 mouse mRNAs, suggesting that alternative polyadenylation is tissue- or disease-specific in each case. Iseli et al. (2002) found that at least half of the human genes were alternatively polyadenylated. They also noted that a
significant portion of polyadenylation sites spread over distances in the kilobase range. However, all of the above computational analyses have focused on the polyadenylation activities within or downstream of the normal 3’ terminal exon. Therefore, they can only identify tandem polyA sites as described in (Edwalds-Gilbert et al. 1997). In this paper, we carry out the first genome-wide analysis on the alternative polyadenylation over the entire transcribed regions of mRNAs in the human, mouse, and rat genomes.

Results

As described in detail in Methods section, mRNA sequences from RefSeq database (Pruitt and Meglott 2001) are aligned to the human, mouse, and rat genomes. The cluster of overlapping mRNAs on the genome is identified as transcription unit. One mRNA sequence is chosen as the reference sequence to represent each transcription unit. The genomic sequence surrounding the reference sequence is extracted as the “extended genomic sequence”. ESTs with putative polyA tails from dbEST database (Boguski et al. 1993) are aligned to these extended genomic sequences. Only the aligned ESTs satisfying stringent criteria are identified as polyA ESTs. PolyA sites are then identified from the 3’ ends of polyA ESTs.

The polyA ESTs can be categorized into four classes as shown in Figure 1. Class I (Figure 1A), the 3’ end of the EST falls within or downstream of the 3’ terminal exon of the reference sequence. Class II (Figure 1B), the 3’ end of the EST falls in the intron of the reference sequence and the 3’ terminal exon of the EST overlaps with the internal exon of the reference sequence. Class III (Figure 1C), the 3’ end of the EST falls in the intron of the reference sequence but the 3’ terminal exon of the EST does not overlap with any exon of the reference sequence. Class IV (Figure 1D), the 3’ end of the EST falls in the internal exon of the reference sequence. The abundance of four classes of polyA ESTs is shown in Table 1.

For almost all polyA sites that we identified, the polyA ESTs using the same polyA site all belong to the same class as defined above. So we can also categorize most polyA sites into four classes according to their polyA ESTs. The abundance of four classes of polyA sites is shown in Table 2.

When the transcription unit contains at least one polyA site, it is referred as
being polyadenylated. When the transcription unit contains more than one polyA sites, it is referred as being alternatively polyadenylated. When the transcription unit contains more than one class I polyA sites, it is referred as having tandem polyA sites as described in (Edwalds-Gilbert et al. 1997). When the transcription unit contains class II polyA sites, it is referred as having composite exons as described in (Edwalds-Gilbert et al. 1997) because the internal exon in the reference sequence behaves as the 3’ terminal exon in class II EST. When the transcription unit contains class III polyA sites, it is referred as having hidden exons because class III EST uses the alternative 3’ terminal exon hidden inside the intron of the reference sequence. This is essentially the skipped exons as described in (Edwalds-Gilbert et al. 1997) in that the reference sequence skips the 3’ terminal exon of class III EST. When the transcription unit contains class IV polyA sites, it is referred as having truncated exons because the internal exon of the reference sequence is truncated by polyadenylation in class IV EST. The abundance of transcription units with polyadenylation, tandem polyA sites, composite exons, hidden exons, truncated exons, and alternative polyadenylation is summarized in Table 3.

More detailed information including all the mRNA and EST alignments are available at web site http://physics.nyu.edu/~jy272/altA. One representative for each class of transcription unit is given below. As an explicit example of our alignment method, we show the mRNA and EST alignments of Lamin A/C (composite exons) in Figure 2.

Rhodopsin has tandem polyA sites. We found five (human), six (mouse), and four (rat) class I polyA sites in rhodopsin. All of them are downstream of the stop codon of the reference sequence. Transcripts using these polyA sites have different 3’ UTRs but encode the same protein. Three of the polyA sites are highly conserved across human, mouse, and rat.

Lamin A/C has a composite exon. Two polyA sites are present in human, mouse, and rat (Figures 2). We searched LocusLink (Pruitt and Maglott 2001) for Lamin A/C in three species. We found all full-length mRNA transcripts using these two polyA sites. When the transcript uses the promoter-distal polyA site in exon 12, it gives rise to Lamin A mRNA. When the transcript uses the promoter-proximal polyA site hidden in the intron of Lamin A mRNA between exon 10 and 11, it gives rise to Lamin C mRNA. When translated into proteins, Lamin A has a different C-terminal domain from Lamin C (Alsheimer et al. 2001). Lamin A has a CaaX box (C,
cysteine; a, aliphatic; X, any amino acid) in the C-terminal domain. Isoprenylation of the cysteine residue of the CaaX box is essential for lamin attachment to the inner nuclear membrane. Lamin C lacks the CaaX box and requires the presence of other lamins for nuclear envelope attachment. In mouse and rat, there is a third isoform Lamin C2 (Alsheimer et al. 2001). Like Lamin C, it uses the promoter-proximal polyA site but has alternative 5’ terminal exon (alternative initiation). In human, there is also a third isoform Lamin Adel10. Like Lamin A, it uses the promoter-distal polyA site but skips exon 10 (alternative splicing).

Cell Division Cycle 42 (CDC42) has a hidden exon. We found six (human), five (mouse), and five (rat) polyA sites in CDC42. With the exception that one polyA site appears to be human-specific, all of the other five polyA sites are conserved among human, mouse, and rat. Polyadenylation at three (four in human) promoter-distal polyA sites leads to the transcripts encoding protein isoform 1. Polyadenylation at the other two promoter-proximal polyA sites leads to the transcripts encoding protein isoform 2. These two protein isoforms have the exactly same lengths but differ in their last 10 C-terminal amino acids. Whether this leads to different functions of the two protein isoforms is worth further investigation.

WW domain binding protein 2 (Wbp2) has a truncated exon in rat but not in human or mouse. 67 (human), 31 (mouse), 14 (rat) polyA ESTs use a conserved class I polyA site. Only one rat polyA EST (UI-R-CA0-bgv-h-11-0-U1.s1, GenBank gi: 11378731) uses a class IV polyA site within the last internal exon of the reference sequence. This polyA EST truncates the coding sequence before it reaches the stop codon. No canonical polyA signals (AATAAA and ATTAAA) or their variants (Beaudoing et al. 2000) are found near the polyA site. This raises the suspicion that this polyA EST may be an experimental artifact.

Discussion

To study how alternative polyadenylation contributes to protein diversity, we must first identify the protein coding regions on the alternative transcripts. This is a difficult task because ESTs are only partial sequences of the whole transcripts. So far, all the known examples (Edwalds-Gilbert et al. 1997 and Zhao et al. 1999) suggest that alternative polyadenylation does not change the open reading frame. So we will assume that all four classes of ESTs follow the same open reading frame as the reference sequence, even
though open reading frame shift due to alternative initiation or splicing can occur in principle. Most of class I polyA sites (99% in human) fall downstream of the stop codon of the reference sequence. They share the stop codon with the reference sequence and encode the same protein. Most of class II polyA sites (93% in human) and class III polyA sites (90% in human) fall between the start and stop codons of the reference sequence on the genome. If exist, the new stop codons of these class II and III ESTs can only lie in the region where they overlap with the introns of the reference sequence. We search for the new stop codons in these regions following the open reading frame of the reference sequence. The result is summarized in Table 4. The new stop codons exist for ~80% class II and ~90% class III polyA sites. The transcripts using these polyA sites are translated into new protein sequences with different C-terminal domains. Interestingly, we note that the percentage of the stop codon presence is about 10% higher in class III than in class II polyA sites in all three species. Most of class IV polyA sites (87% in human) fall between the start and stop codons of the reference sequence on the genome. These class IV ESTs truncate the coding exon of the reference sequence and lack the stop codon. If not experimental artifacts, most of class IV ESTs and a small portion of class II and III ESTs can only be non-coding transcripts. Non-coding transcripts have been shown to exist extensively in human (Cawley et al. 2004). They can simply be the mistakes made by the transcription machinery and undergo the degradation soon after the transcription, or they can have important regulatory functions (Martens et al. 2004 and Cawley at al. 2004).

We observe that different polyA sites in the same transcription unit can have different polyadenylation efficiencies. Some polyA sites are used more often and associated with a large number of polyA ESTs. The others are used more rarely and associated with only one or a few polyA ESTs. We define the polyadenylation efficiency of the polyA site as the number of polyA ESTs using the polyA site. Within each class of polyA site, the average polyadenylation efficiency is simply the total number of polyA ESTs (shown in Table 1) divided by the total number of polyA sites (shown in Table 2). In human, the average polyadenylation efficiencies are: 14.53 (class I), 2.65 (class II), 3.09 (class III), and 2.00 (class IV). In all three species, we found that the average polyadenylation efficiencies follow the same order: class I > class III > class II = class IV. As pointed out in (Edwalds-Gilbert et al. 1997 and Zhao et al. 1999), there is competition between splicing and polyadenylation during the mRNA post-transcriptional processing. Recently, Qiu et al. (2004) showed an interesting distance effect on the
polyadenylation efficiency. They found that the alternative polyadenylation of AAV5 RNA within an intron is inhibited by U1 snRNP binding to the 5’ splice site immediately upstream to the polyA site. When the distance between the 5’ splice site and the polyA site was increased by inserting a heterologous DNA sequence, the inhibition was reduced. Similar inhibition and distance effect was observed in HIV-1 where the alternative polyadenylation at the promoter-proximal polyA site was inhibited by U1 snRNP binding to the 5’ splice site immediately downstream of the polyA site. In our study, class II and class III polyA sites are immediately downstream of the 5’ splice site whereas class IV polyA sites are immediately upstream of the 5’ splice site. Generally, the distance between the polyA site and the 5’ splice site is longer for class III (9,708 bps (base pairs) in rat CDC42), but shorter for class II and IV (110 bps in rat Lamin A/C and 40 bps in rat Wbp2). This may explain why class III sites are more efficient than class II and IV sites. Qiu et al. (2004) also pointed out that the polyadenylation efficiency is proportional to the distance between promoter and polyA site. This is consistent with the observation that the most efficient polyA site is often the most distal one from the promoter. Class I polyA sites are most distal to the promoter among four classes of polyA sites. This may explain why class I polyA sites are most efficient at polyadenylation.

We compared the transcription units containing class II, III, and IV polyA sites across human, mouse, and rat. We found that alternative polyadenylation in Lamin A/C (class II), CDC42 (class III), Nucleophosmin (class III), Olfactomedin 1 (class III) are conserved in all three species. They all result in new protein isoforms with different C-terminal domains. Since there are only a relatively small number of transcription units in common among three species in our study and human, mouse, and rat have different redundancies of ESTs, the above list is by no means complete.

Four classes of alternative polyadenylation are distinct in their splicing-polyadenylation patterns, protein-coding capacities, and polyadenylation efficiencies. Strong evidences including full-length mRNA sequences and protein sequences exist for class I, II, and III transcripts. But the role of class IV transcripts remains elusive. To tell whether they are experimental artifacts or non-coding transcripts, future experimental verifications will be very necessary.
Methods

We extracted 20,860 (human), 16,743 (mouse), and 4,843 (rat) mRNA sequences from the NCBI RefSeq database release 4 (Pruitt and Maglott 2001). Only the mRNAs with accession numbers starting with NM (mRNA sequences that are experimentally verified) were selected. The accession numbers, start codon positions, and stop codon positions were also extracted. These sequences were masked by RepeatMasker (Smit and Green 2004, setting -w) program using WU-Blast program (Gish 2003). The repeat databases were obtained from RepBase Update (http://www.girinst.org).

Human (build 34), mouse (build 32), and rat (build 2) genomes were downloaded from NCBI. We aligned the masked sequences onto their genomes using the megablast program (Altschul 1990, settings -p 98 -D 3). Only the alignments with percent identities greater than 98% were kept. The minimum and maximum positions of each mRNA sequence alignment on each genomic contig were calculated and the genomic sequence starting from 40 kbps upstream of the minimum position to 40 kbps (kilo base pairs) downstream of the maximum position were extracted from the genomic contig. The unmasked mRNA sequences were realigned to their corresponding extracted genomic sequences one by one using sim4 program (Florea 1998, settings A=0 P=1). Only the sim4 alignments satisfying the following high quality controls were kept: 1. All splice sites were in the same orientations. 2. The entire sequence except the 50bps at both ends was aligned. 3. The average percent identity of alignments was greater than 95%. The average percent identity given by sim4 was defined as the alignment score. Overall, we aligned 20,095 (96%) human mRNAs, 14,861 (89%) mouse mRNAs, and 3,803 (79%) rat mRNAs to their genomes. The disparity in the percentages of mRNAs aligned was largely due to the differences in the completeness of three genomes.

We clustered all splicing variants based on their common exon boundaries. Two mRNA sequences were clustered together if they shared at least one 5’ or 3’ splice site. The mRNA sequence with the highest alignment score in the cluster was chosen as the reference sequence to represent that cluster. If the same reference sequence was aligned to multiple positions on the genome, the position with the highest alignment score was chosen. We identified the resulting clusters as transcription units. We obtained 16,018 (human), 14,406 (mouse), and 3,771 (rat) transcription units. We defined the “extended genomic sequence” as the genomic sequence spanning between 10 kbps upstream and 10 kbps downstream of the reference sequence.
alignments. The extended genomic sequences were extracted and converted into the same orientation as the reference sequences (5’ to 3’). The new alignments of the reference sequences on their extended genomic sequences were calculated.

We downloaded and parsed the NCBI dbEST database (08 April 2004 release, Boguski et al. 1993) for ESTs of human, mouse, and rat. The ESTs shorter than 100 bps were discarded. The EST sequences were searched for polyA tracks in the forward and polyT tracks in the reverse orientation. The polyA or polyT tracks must have more than 10 contiguous A’s or T’s. One end of the polyA or polyT track must be within 10bps of the EST end. The other end of the polyA or polyT track was identified as the putative start of polyA tail. These ESTs were identified as the putative polyA ESTs and were converted into the forward orientation such that they all have polyA tracks at the 3’ ends. Overall, we obtained 691,293 ESTs with putative polyA tails for human, 227,576 for mouse, and 223,367 for rat. They were then masked by RepeatMasker program (setting -w). Masked EST sequences were aligned to the extended genomic sequences by megablast program only in the forward orientation (settings -p 95 -D 3 -S 1). Only the alignments with total coverage higher than 50% of the EST sequence lengths and percent identities higher than 95% were kept. The minimum and maximum positions of each EST sequence alignment on each extended genomic sequence were calculated and the sequence spanning between 4 kbps upstream of the minimum position and 4 kbps downstream of the maximum position was extracted from the extended genomic sequences. The unmasked EST sequences were realigned to their corresponding extracted sequences one by one using sim4 program (settings A=0 P=1). Only the sim4 alignments satisfying the following quality controls were kept: 1. All splice sites were in the same orientation. 2. The entire sequence was aligned except the 50 bps at 5’ end and the 3’ end of alignment must occur within 5 bps of the putative start of polyA tail. 3. The average percent identities of alignment must be higher than 90%.

The 3’ end of alignment was identified as putative polyA cleavage site. 50 bps downstream sequences of the putative polyA cleavage sites were extracted and searched for the polyA tracks. The polyA track was identified if there were more than 5 contiguous A’s or more than 8 A’s out of any 10 bps sliding window. If the polyA track was found in the downstream sequence, the polyA cleavage site was identified as the putative internal priming site. The aligned ESTs without the putative internal priming sites
were identified as polyA ESTs. To further filter out the polyA ESTs that may belong to the intronic genes or overlapping genes, the polyA ESTs must satisfy the following criteria: if the EST was spliced, it must share at least one 3’ or 5’ splice site with the reference sequence; if the EST was not spliced, it must overlap with the 3’ terminal exon of the reference sequence. The 3’ ends of the polyA ESTs on the same extended genomic sequence were clustered if they were less than 50 bps apart. This was because the 3’ ends of the polyadenylated transcripts were known to fluctuate within a small distance after the polyA signals. The clusters of the 3’ ends of polyA ESTs were identified as the polyA sites.

Acknowledgements

This project was supported by NIH grant RR-16466-01, NSF grant EPS-0092040, the Alaska BRIN program, and the University of Alaska Foundation. Leone Thierman in the Arctic Region Supercomputing Center kindly produced the figures. We thank anonymous referees for helpful comments.
Figure Legends

Figure 1. Four classes of patterns of alternative polyadenylation are shown together with the reference sequence. (A) tandem polyA sites. (B) composite exons. (C) hidden exons. (D) truncated exons.

Figure 2. Comparison between the full-length mRNA alignments and the polyA EST alignments of Lamin A/C in (A) human, (B) mouse, and (C) rat. In all three species, polyA ESTs are clustered into two clusters, each corresponding to a polyA site. Lamin A and C isoforms are present in all three species. In human, a third isoform Lamin Adel10 skips the exon 10 of Lamin A. In mouse and rat, a third isoform Lamin C2 has alternative initiation. GenBank gi numbers are shown for all mRNAs and ESTs.
## Tables

<table>
<thead>
<tr>
<th>PolyA ESTs</th>
<th>Class I</th>
<th>Class II</th>
<th>Class III</th>
<th>Class IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>329,966</td>
<td>3,230</td>
<td>2,257</td>
<td>1,816</td>
</tr>
<tr>
<td>Mouse</td>
<td>86,598</td>
<td>394</td>
<td>359</td>
<td>656</td>
</tr>
<tr>
<td>Rat</td>
<td>27,999</td>
<td>148</td>
<td>138</td>
<td>109</td>
</tr>
</tbody>
</table>

Table 1. The abundance of four classes of polyA ESTs in human, mouse, and rat

<table>
<thead>
<tr>
<th>PolyA sites</th>
<th>Class I</th>
<th>Class II</th>
<th>Class III</th>
<th>Class IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>22,712</td>
<td>1,221</td>
<td>730</td>
<td>907</td>
</tr>
<tr>
<td>Mouse</td>
<td>12,681</td>
<td>244</td>
<td>144</td>
<td>384</td>
</tr>
<tr>
<td>Rat</td>
<td>3,379</td>
<td>88</td>
<td>59</td>
<td>62</td>
</tr>
</tbody>
</table>

Table 2. The abundance of four classes of polyA sites in human, mouse, and rat

<table>
<thead>
<tr>
<th>Transcription units</th>
<th>Polyadenylated Tandem polyA sites</th>
<th>Composite exons</th>
<th>Hidden exons</th>
<th>Truncated exons</th>
<th>Alternatively polyadenylated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>13,268</td>
<td>5,521</td>
<td>1,084</td>
<td>650</td>
<td>794</td>
</tr>
<tr>
<td>Mouse</td>
<td>9,227</td>
<td>2,532</td>
<td>233</td>
<td>138</td>
<td>355</td>
</tr>
<tr>
<td>Rat</td>
<td>2,611</td>
<td>621</td>
<td>87</td>
<td>57</td>
<td>59</td>
</tr>
</tbody>
</table>

Table 3. The abundance of transcription units in human, mouse, and rat. The percentages of alternative polyadenylation are calculated relative to the polyadenylated transcription units.

<table>
<thead>
<tr>
<th>Presence of stop codons</th>
<th>Class II polyA sites</th>
<th>Class III polyA sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>81%</td>
<td>88%</td>
</tr>
<tr>
<td>Mouse</td>
<td>73%</td>
<td>84%</td>
</tr>
<tr>
<td>Rat</td>
<td>79%</td>
<td>89%</td>
</tr>
</tbody>
</table>

Table 4. The percentages of the presence of stop codons in class II and III polyA sites.
References


Qiu, J., and Pintel, D.J., 2004. Alternative polyadenylation of adeno-associated virus type 5 RNA within an internal intron is governed by the distance between the promoter and the intron and is inhibited by U1 small nuclear RNP binding to the intervening donor. *J. Biol. Chem.* **279**: 14889-14898.


Figure 2

B. Mouse Lamin A/C

C. Rat Lamin A/C