Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

Under-representation of PolyA/PolyT tailed ESTs in Human ESTdb: an obstacle to alternative polyadenylation inference

Roi Gilat, Sergey Goncharov, Nir Esterman and Dorit Shweiki*

Bioinformatics Program, School of Computer Science, The Academic College of Tel Aviv-Yaffo, Tel Aviv, Israel;
Dorit Shweiki* - Email: dorits@mta.ac.il; Phone: +972 3 5211853; Fax: +972 3 5211871;
* Corresponding author

received September 05, 2006; revised October 02,2006; accepted October 02, 2006; published online October 07, 2006

Abstract:

Alternative polyadenylation is a key regulatory process which affects the 3' end formation of variants of the same transcription unit, thus altering gene expression pattern, and transcripts' cellular behaviour and characteristics. The common methodology for computational analysis of alternative polyadenylation signal utilization is based on EST data, specifically on PolyA/PolyT tailed ESTs. Studying the human ESTs dataset we detected a significant under-representation of PolyA/PolyT tailed ESTs, constituting only 10% of most libraries. Consequently, more than 50% of false-negative events are revealed in the analysis of alternatively polyadenylated variants' expression. We therefore argue that the ratios of PolyA/PolyT tailed ESTs, as represented in the human EST database, do not reflect the true-picture of 3' end variants formation of a given physiological situation. Thus the EST database should not be considered a reliable source for alternative polyadenylation signal usage inference.

Keywords: polyadenylation; human EST; PolyA/PolyT; alternative polyadenylation inference

Background:

The 3’ UTR of eukaryotic mRNAs contains valuable regulatory information, vastly affecting mRNA translation levels, mRNA stability and mRNA cellular localization. **[1-4]** Thus, mRNA 3’-end determination through alternative polyadenylation may significantly alter transcripts' cellular behavior and characteristics (e.g., cellular scatter, levels of stability or levels of translation), and consequently gene expression pattern.

A significant number of pre-mRNAs are subjected to differential polyadenylation events; i.e. alternative polyadenylation signal selection and the formation of mature transcripts with 3’ UTRs of variable length. **[5, 6]** The scope of alternative polyadenylation phenomenon was assessed several times by various computational studies. ESTs based analyses estimated alternative polyadenylation regulation to be implemented to at least half of the human genes. **[5, 7-10]**

The common methodology for computational analysis of alternative polyadenylation signal utilization is based on EST data. PolyA/PolyT tailed ESTs are being used to assess the 3' end of mRNA variants, of a given tissue. Based on EST analysis, a growing number of genes have been reported to go through alternative polyadenylation signal utilization and to produce 3’ UTR mRNA variants, some in a tissue and a process-specific manner. **[6, 11]** Zhang et al reported on biased alternative polyadenylation in human tissues. Indicating the existence of tissue-specific strong and weak polyadenylation site usage. **[11]**

ESTs are an informative, significant tool, valuable for many studies, and have been used successfully in gene identification and gene~~s~~ structure analysis **[12]**, tissue-specific alternative splicing **[13]**, transcriptome maps and large-scale analysis of gene~~s~~ expression studies. **[14]** The EST database suffers from several inherent problems due to intrinsic contaminations and artifacts which may contribute to biased results in EST-based studies. It is a redundant database, contaminated with vector or linker sequences and sometimes with genomic sequences. **[15]** EST database was reported to contain a high rate of sequence errors. **[16]** Some ESTs are chimeric ones, the result of two or more expressed sequences from different transcribed units. **[17]** Pre-mature mRNAs may also be represented in the EST dataset, thus intronic sequences are represented as exonic ones. **[18]**

Contaminations are not the only obstacles in computational analysis which is based on the EST dataset. The question of an accurate representation of mRNA variants and of gene expression levels is equally significant. In the study of alternatively polyadenylated genes and their polyadenylation signal usage, one must first verify that the common tool being used, i.e. PolyA/PolyT tailed ESTs alignment is valid.

We screened human EST libraries for PolyA/PolyT tailed ESTs fraction in order to assess whether these ESTs can be utilized to evaluate polyadenylation signal usage in alternatively polyadenylated genes. 546 alternatively polyadenylated genes were used as our training-set on which polyadenylation signal usage analysis was carried out. Our results clearly demonstrate that PolyA/PolyT tailed ESTs are vastly under-represented in human EST libraries, thus they may not be a valid tool to evaluate alternative polyadenylation signal usage.

Methodology:

Differential polyadenylation training-set construction: In our designing and construction of a differentially polyadenylated gene training-set we were guided primarily by criteria of experimental validation. Therefore although it is estimated that approximately half of human genes undergo alternative polyadenylation, only mRNAs with verified functional polyadenylation signals were retrieved and further analyzed. 27565 Human RefSeq mRNA sequences (28 August 2005 release) were screened for poly(A) signal feature key at their annotation (**[19]**, personal communication). 1688 mRNA sequences contained more than one polyadenylation signal. Intra-segmental sequence is defined as the sequence between the end of the first polyadenylation signal to the beginning of the last one. Intra-segmental sequence size ranges between 0 (consecutive signals) to 8272 nucleotides. The default criterion for further analysis was an intra-segmental sequence of at least 50 nucleotides. 755 non redundant, of the 1235 sequences identified, contained more than one polyadenylation signal with an intra-segmental sequence at the size of 50 nucleotides or longer (the variants with the longest UTR were choose for the non redundant training set). The number of signals per mRNA molecule~~s~~ varies: out of 755 non-redundant mRNAs, 546 sequences contained two verified polyadenylation signals (72.3 %), while others contained up to 8 consecutive polyadenylation signals. Polyadenylation signal usage in differential polyadenylated genes was calculated only for the 546 mRNAs containing two-signals.

ESTs based analysis of genes expression levels: Gene expression level, of the different tissues was evaluated utilizing Human ESTs extracted from dbEST (September 2004 release). **[20]** ESTs shorter than 25 bp, from normalized or subtracted libraries, cancerous, embryonic and pooled libraries were discarded. EST tissue attribution was parsed either from the library description field or the tissue_type qualifier field in the source feature key. A synonym name table for the different tissues was used (data not shown). Cancerous ESTs were classified as such when any of the following terms appeared in the description field of the library or its source feature key field: cancer, adenoma, carcinoma, tumor, blastoma, glioma, leukemia, lymphoma, invasive, sarcome, melanome, myeloma, metastatic, metastasis, anaplastia, anaplast, anaplasis (as a word or as a suffix of a word). Normal ESTs were classified as such by the process of elimination if no cancer related word appeared in their annotation. Embryonic ESTs were classified as such when any of the following terms appeared in the description field of the library or its source feature key field: embryo, fetal, fetus, trophoblast. Adult ESTs were classified as such when any of the following terms appeared in the description field of the library or its source feature key field: adult, year and a number indication, post natal or newborn. 2318258 ESTs remained representing normal tissue libraries of a human adult. ESTs were masked using RepeatMasker (**[21]**) and RepBase database (release 10.06). **[22]** Masked ESTs were aligned to Human RefSeq mRNAs using standalone BLAST 2.2.12. **[23]** Variants of the same gene were identified using the GeneID qualifier field in the gene feature key. ESTs aligned to more than one GeneID with a score >80 were discarded. Numbers of aligned ESTs per gene per tissue stage were calculated.

ESTs based analysis of polyadenylation signal usage: 3' PolyA tailed and 5' PolyT tailed ESTs were identified from the non-masked Human ESTs file. ESTs had to fulfill the following criteria: EST length >=50, PolyA and PolyT tails had to be a consecutive sequence of As or Ts of at list 10 bp, within 10 bp from EST end, and followed by at list 25 bp of non-PolyA or non-PolyT sequence. PolyT ESTs were reverse complemented. A total of 275358 PolyA/PolyT ESTs were retrieved, representing 11.9% of the total adult normal ESTs.

Masked ESTs, fulfilling the above mentioned criteria were then aligned to Human RefSeq mRNAs in order to discard ESTs aligning to more than one GeneID. ESTs at the size of 50 to 150 bp, score >65, identity > 75% and alignment length > 25bp; or ESTs larger than 150 bp, score > 80, were regarded as showing high sequence homology, and thus saved. We then parsed Blast result searching for alignments which specifically indicate a given polyadenylation signal usage. ESTs were considered to single-out a specific polyadenylation signal usage when the 3' end of the molecule was aligned in the range of 50 nucleotides from the beginning of a given polyadenylation signal. Accordingly, polyadenylation signal usage was calculated per gene, per tissue.

Results:

Low percentages of PolyA/PolyT tailed ESTs in ESTdb libraries

Normal human adult EST libraries of different tissues were screened to evaluate the percentage of PolyA/PolyT tailed ESTs from the total EST sequences in the library. The majority of human libraries are characterized by a very low percentage of PolyA/PolyT tailed ESTs. Figure 1 summarizes the distribution of PolyA/PolyT tailed ESTs in the different libraries. In the vast majority of human libraries, PolyA/PolyT tailed ESTs constitute less than 10% of library sequences. Only in a small fraction of the libraries, are ESTs 100% PolyA/PolyT tailed. Thus, indicating an extreme under-representation of PolyA/PolyT tailed ESTs in normal human adults' EST dataset, especially when taking into consideration that the vast majority of EST libraries are being prepared using oligo-dT primers.

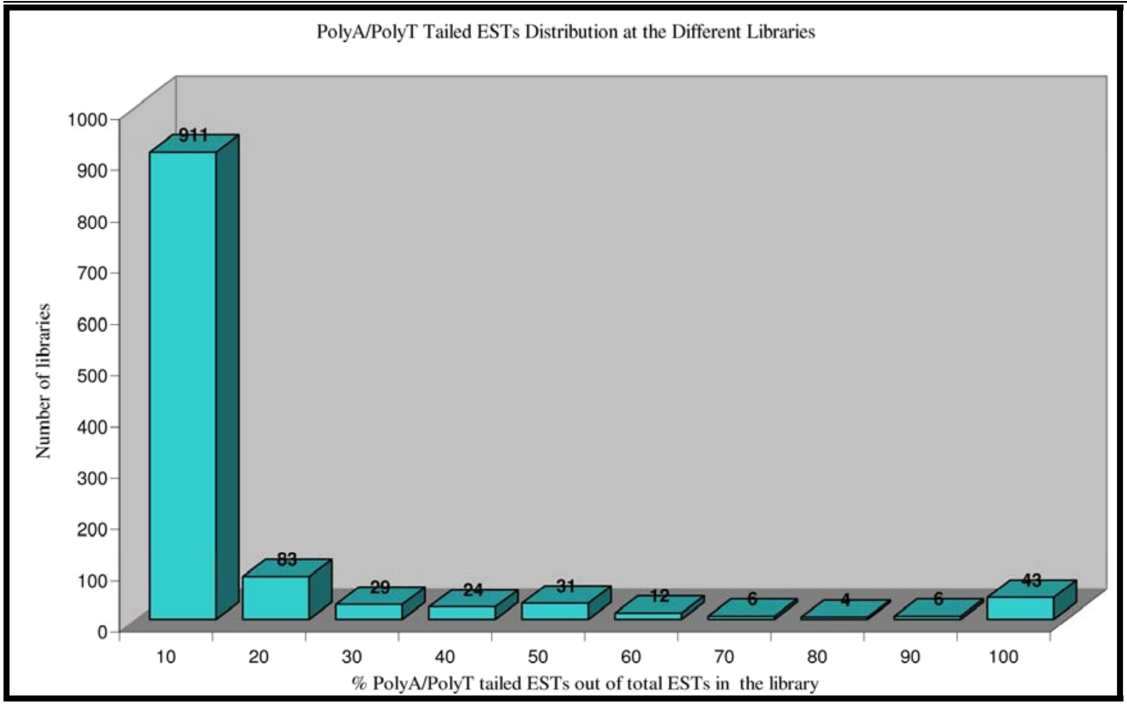

Figure 1: The figure summarizes the distribution of PolyA/PolyT tailed ESTs in the different libraries. For each library we calculated the percentage of PolyA/PolyT tailed ESTs out of the total number of ESTs in the library. Results are grouped in a range of 10 units of percents (e.g., 0 to 10 %, 11 to 20 %, 21 to 30 % etc.)

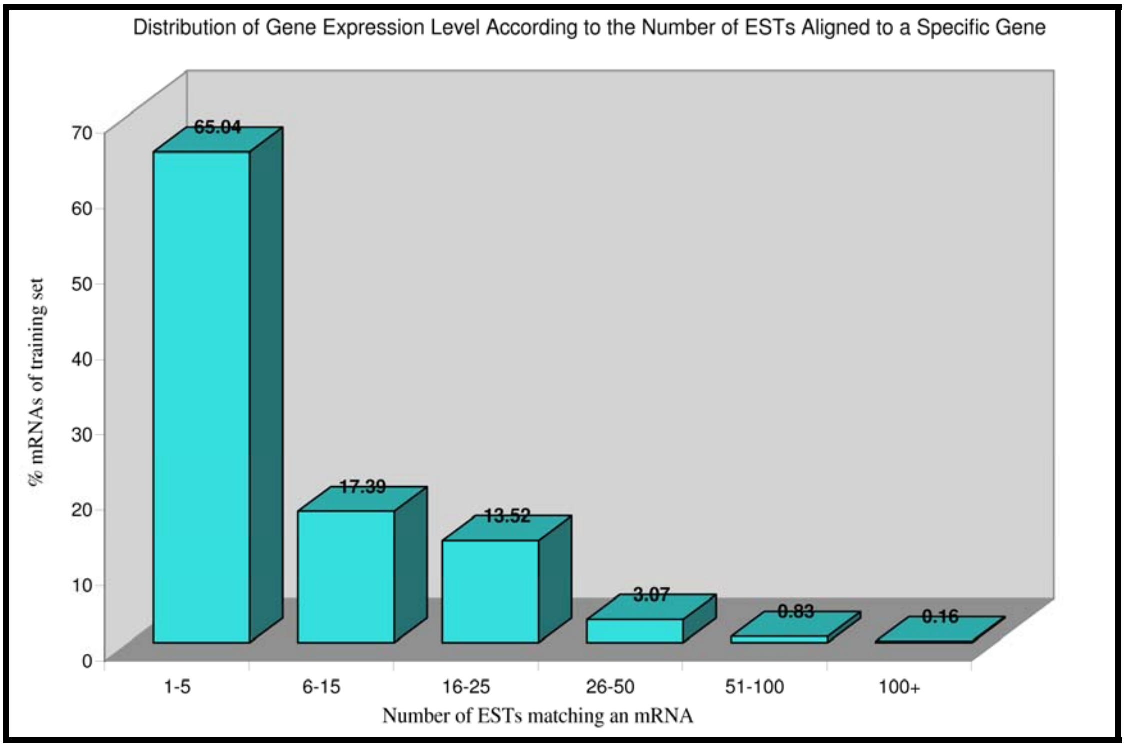

Figure 2: Figure 2 summarizes the distribution of gene expression level according to the number of ESTs which are aligned to a given gene. For each gene we calculated the number of specifically aligned ESTs. The genes are grouped according to the number of ESTs aligned, and the size of each group is represented in percentage from a total of 546 genes that were analyzed

The expression level of most genes is represented by small number of ESTs

We calculated the number of ESTs aligned to a specific gene, and thus considered to represent the expression level of the gene. Figure 2 summarizes the distribution of gene expression level according to the number of ESTs which are aligned to a given gene. 65% of the 546 genes analyzed, aligned only 1 to 5 ESTs in human ESTs libraries screened. Additional 17.4% of the analyzed genes aligned 6 to 10 ESTs and 13.5% aligned 11 to 25 ESTs, the remainder of 4.2 % aligned 26 ESTs and more.

Since the frequency of PolyA/PolyT tailed ESTs in ESTdb libraries is low and since the expression level of most genes is represented by small numbers of ESTs, it is very likely that polyadenylation signal usage analysis, that is based on PolyA/PolyT tailed ESTs dataset, does not represent a truthful picture of the physiological situation. Consequently it is probable that many alternatively polyadenylated variants are not detected and reported, though produced.

Under-representation of PolyA/PolyT tailed ESTs is responsible for a high percentage of false-negative results in the evaluation of alternative polyadenylated variants' expression

PolyA/PolyT tailed ESTs are commonly being used to assess the 3' end of mRNA variants, of a given tissue. The position of the PolyA/PolyT tailed EST alignment specifically indicates a given polyadenylation signal usage. Thus, if no PolyA/PolyT tailed EST aligns to a given polyadenylation signal position, this 3' end variant is considered unexpressed. Prior to the BLAST analysis with PolyA/PolyT tailed ESTs, we analyzed the expression level of our training-set genes, as indicated by the alignment of the general population of ESTs. We then calculated the degree of false-positive results; i.e., alternatively polyadenylated genes which according to BLAST analysis with the general population of ESTs showed measurable expression in a given tissue, but showed no expression when analyzed utilizing PolyA/PolyT tailed EST solely. Unsurprisingly, a substantial percentage of 55.7% events of false-negative results were counted (false-negative rate for all human genes was calculated and stands on 35%).

Discussion:

We screened normal human EST libraries for PolyA/PolyT tailed ESTs fraction in order to assess whether they can serve as a means to evaluate polyadenylation signal usage in alternatively polyadenylated genes. Our results clearly demonstrate that PolyA/PolyT tailed ESTs are vastly under-represented in human EST libraries. Though the vast majority of EST libraries are primed with an oligo(dT) primer, hardly any library in the database contains 100% PolyA/PolyT tailed ESTs. On the contrary, for the most part PolyA/PolyT tailed ESTs constitute only 10% of most libraries. Under-representation of PolyA/PolyT tailed ESTs is the result of sequences being processed prior to submission (vector and linker sequences removed as well as polyA/polyT evidences); due to the conditions sequencing analysis is being performed at (i.e., low concentrations of analogs in order to assure long chain sequencing, usually is responsible for skipping the first tens nucleotides of the molecule being sequenced); and a dominantly preference to sequence the 5'-end of the molecules in order to discover protein coding sequences. Under-representation of PolyA/PolyT tailed ESTs is the main cause for the detected ratio of more than 50% false-negative results. Due to this problematic ratio we claim that PolyA/PolyT tailed EST-based inference in terms of polyadenylation signal usage is unreliable.

Our results question both the reports on the widespread of alternative polyadenylation phenomenon, and on the distribution of alternative polyadenylation signal utilization. ESTs-based analyses estimated alternative polyadenylated genes abundance to include at least half of human genes. **[5,7-10]** However, under-representation of PolyA/PolyT tailed ESTs in the database may lead to a biased incorrect estimation of alternative polyadenylation scope, depending on whether PolyA/PolyT tailed ESTs are randomly under-represented or not.

Moreover, any conclusions regarding polyadenylation signal usage in physiological or tissue-specific contexts are controversial. Zhang et al reported on biased alternative polyadenylation in human tissues, indicating the existence of tissue-specific strong and weak polyadenylation site usage. **[11]** However, based on our results, the validity of their deductions is open for discussion. Low expressed, alternatively polyadenylated variants of genes will appear as if given signals are solely being used, despite the fact that physiologically, additional polyadenylation signal is utilized. Therefore, the ratios of PolyA/PolyT tailed ESTs represented in the database do not reflect the true-picture of mRNAs 3' end variants of a given physiological situation.

Conclusion:

In conclusion, the under-representation of PolyA/PolyT tailed ESTs in normal human adults' EST libraries, low numbers of ESTs representing expressed genes, and consequently high rates of false-negative results all together make the EST database unreliable as a source of alternative polyadenylation signal usage inference. Though we believe that alternative polyadenylation mechanism significantly contributes to the diversity of genome content expression, the scope and nature of this phenomenon can not be assessed by using the available dataset. Alternative-polyadenylation specific microarrays should be designed in order to assess polyadenylation signal usage of human genes. Alternatively, one will need access to the original chromatograms or the unedited sequences deduced from it. Only then can the contribution of alternative polyadenylation to human transcriptome diversity be evaluated, and the role of this process in different physiological contexts be assessed.

Acknowledgement:

We thank Dr. Nili Beck for statistical advice; Dr. Tamar Kahan for critical reading of the manuscript. This research was supported by The Academic College of Tel Aviv- Yaffo research grant.

References:

[01] B. Mazumder, *et al*., *Trends Biochem Sci.,* 28:91 (2003) [PMID: 12575997]

[02] C. J. Wilusz, *et al., Nat Rev Mol Cell Biol.,* 2:237 (2001) [PMID: 11283721]

[03] R. P. Jansen, *Nat Rev Mol Cell Biol.,* 2:247 (2001) [PMID: 11283722]

[04] C. J. Decker & R. Parker, *Curr Opin Cell Biol.,* 7: 386 (1995) [PMID: 7662369]

[05] E. Pauws, *et al*, *Nucleic Acids Res.,* 29:1690 (2001) [PMID: 11292841]

[06] E. Beaudoing & D. Gautheret, *Genome Res.,* 11:1520 (2001) [PMID: 11544195]

[07] E. Beaudoing, *et al*., *Genome Res.,* 10:1001 (2000) [PMID: 10899149]

[08] C. Iseli, *et al*., *Genome Res.,* 12:1068 (2002) [PMID: 12097343]

[09] B. Tian, *et al*, *Nucleic Acids Res.,* 33:201 (2005) [PMID: 15647503]

[10] J. Yan & T. G. Marr, *Genome Res.,* 15:369 (2005) [PMID: 15741508]

[11] H. Zhang, *et al*., *Genome Biol.,* 6:R100 (2005) [PMID: 16356263]

[12] A. Lindlof, *Appl Bioinformatics,* 2:123 (2003) [PMID: 15130797]

[13] Q. Xu, *et al*, *Nucleic Acids Res.,* 30:3754 (2002) [PMID: 12202761]

[14] M. D. Adams, *Bioessays,* 18:261 (1996) [PMID: 8967892]

[15] H. H. Chou & M. H. Holmes, *Bioinformatics,* 17:1093 (2001) [PMID: 11751217]

[16] J. S. Aaronson, *et al*., *Genome Res.,* 6:829 (1996) [PMID: 8889550]

[17] J. Burke, *et al*., *Genome Res.,* 8:276 (1998) [PMID: 9521931]

[18] R. Sorek & H. M. Safer, *Nucleic Acids Res.,* 31:1067 (2003) [PMID: 12560505]

[19] Polyadenylation signal indication in RefSeq records annotation is based on integration of data from several sources: consensus sequence Poly(A) signals in conjunction with strings of A's; EST based indications on poly(A) site; experimentally reported records. Personal communication with Reference Sequence Group at NCBI

[20] M. S. Boguski, *et al*., *Nat Genet.,* 4:332 (1993) [PMID: 8401577]

[21] A. F. A. Smit, *et al*., unpublished data http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker

[22] J. Jurka, *et al*., *Cytogenet Genome Res.,* 110:462 (2005) [PMID: 16093699]

[23] S. F. Altschul, *et al*., *J Mol Biol*., 215:403 (1990) [PMID: 2231712]

Edited by P. Kangueane

Citation: Gilat *et al.*, Bioinformation 1(6): 220-224 (2006)