In silico analysis of EST and genomic sequences allowed the prediction of cis-regulatory elements for Entamoeba histolytica mRNA polyadenylation

Absalom Zamoranoa, César López-Camarillob, Esther Orozco, Christian Weberde, Nancy Guillenf, Laurence A. Marchata

a ENMH-IPN, Programa Institucional de Biomedicina Molecular, Guillermo Massieu Heguera #239, Ticoman, CP 07320, México, D.F., Mexico
b Universidad Autónoma de la Ciudad de México, Posgrado en Ciencias Genómicas, Av. San Lorenzo #290, Col. del Valle, CP 03100, México, D.F., Mexico
c CINVESTAV-IPN, Departamento de Patología Experimental, Av. IPN # 2508, Col. San Pedro Zacatenco, CP 07360, México, D.F., Mexico
d Instituto Pasteur, Unité Biologie Cellulaire du Parasitisme, 28 rue du Dr. Roux, Paris F-75015, France
e INSERM U786, Paris F-75015, France

A R T I C L E   I N F O

Article history:
Received 12 September 2007
Received in revised form 24 March 2008
Accepted 24 March 2008

Keywords:
Bioinformatics analysis
Pre-mRNA 3′-end processing
Cis-regulatory elements
Protozoan parasite

A B S T R A C T

In most eukaryotic cells, the poly(A) tail at the 3′-end of messenger RNA (mRNA) is essential for nuclear export, translatability, stability and transcription termination. Poly(A) tail formation involves multi-protein complexes that interact with specific sequences in 3′-untranslated region (3′-UTR) of precursor mRNA (pre-mRNA). Here we have performed a computational analysis of a large EST and genomic sequences collection from Entamoeba histolytica, the protozoan parasite responsible for human amoebiasis, to identify conserved elements that could be involved in pre-mRNA polyadenylation. Results evidenced the presence of an AU-rich domain corresponding to the consensus UA(A/U)UU polyadenylation signal or variants, the cleavage and polyadenylation site that is generally denoted by U residue and flanked by two U-rich tracts, and a novel A-rich element. This predicted array was validated through the analysis of genomic sequences and predicted mRNA folding of genes with known polyadenylation site. The molecular organization of pre-mRNA 3′-UTR cis-regulatory elements appears to be roughly conserved through evolutionary scale, whereas the polyadenylation signal seems to be species-specific in protozoan parasites and the novel A-rich element is unique for the primitive eukaryote E. histolytica. To our knowledge, this paper is the first work about the identification of potential pre-mRNA 3′-UTR cis-regulatory sequences through in silico analysis of large sets of cDNA and genomic sequences in a protozoan parasite.

1. Introduction

The knowledge of parasite genome sequences is an extraordinary tool to identify and functionally characterize new genes that are important for pathogen survival and host infection. It also makes possible the determination of genes organization and the detection of conserved cis-regulatory elements for gene expression. Experimental characterization of 5′- and 3′-untranslated regions (5′-UTR and 3′-UTR, respectively) successfully allows the identification of control sequences, but they are usually limited to a small number of genes and specific contextual scopes. Moreover, they often consider individual interactions without taking surrounding sequences or factors into account. Computational analysis of genomic, full-length cDNA and expressed sequence tag (EST) databases represents an attractive alternative to detect conserved signals that can be important for the expression of numerous genes.

Bioinformatics approach has been largely used to predict regulatory elements for precursor messenger RNA (pre-mRNA) 3′-end polyadenylation (Beaudoin et al., 2000; Hajarnavis et al., 2004; Loke et al., 2005) that is a crucial maturation step for most eukaryotic mRNA, affecting stability, translatability, and nuclear-to-cytoplasmic export (Zhao et al., 1999), representing therefore an integral part of gene expression regulation. The polyadenylation reactions involves about 25 nuclear proteins that recognize poly(A) sequences in pre-mRNAs 3′-UTR and act in a coordinated way to perform polyadenylation reaction (Proudfoot, 2004). In human cells and yeast, the polyadenylation site (poly(A) site) is generally flanked by the upstream polyadenylation signal (A(A/U)UAAA) and the downstream U/GU-rich element. Additional U-rich and G/C-rich elements are also associated with poly(A) site (Zhao et al., 1999; Graber et al., 2002; Hu et al., 2005).

Entamoeba histolytica is the intestinal protozoan parasite responsible for amebic colitis and liver abcess, which cause
mortality in many developing countries (Jackson, 2000). The sequencing of the parasite genome provides new insights into the cellular workings and genome evolution of this major human pathogen (Loftus et al., 2005). As a first step towards characterizing 3′-UTR cis-regulatory sequences that could be important for pre-mRNA 3′-end processing in E. histolytica, we performed a small-scale in silico analysis of genomic regions and identified four conserved motifs (López-Camarillo et al., 2005). Here, we extended the computational analysis to a larger E. histolytica EST collection and genomic sequences dataset and proposed a hypothetical array of UA(U/A)UU, U-rich and A-rich elements in pre-mRNA 3′-UTR. Finally, our model was validated by studying genomic sequences and secondary RNA structures of E. histolytica genes with known poly(A) site.

2. Materials and methods

2.1. Sequences selection

E. histolytica EST were selected from a set of 2348 raw cDNA sequences obtained by retrotranscription of mRNA isolated from the virulent HM1:IMSS strain. Briefly, cDNA sequences were generated by reverse transcription initiated by an oligodeoxythymidylate primer, cloned into the TriplRx2M plasmid (Clontech) and automatically sequenced (Weber et al., 2006). As the cloning strategy was not oriented, sequences were searched for the presence of at least ten A or T residues tracts that were assumed to correspond to EST 3′-ends. For sequences with poly(T) tract, we obtained complementary strands and inverted them, so that they were in sense orientation. To eliminate sequences with internal poly(A) tracts and make certain of working with 3′-ends, we selected sequences having at least 400 nucleotides (nt) upstream the 10-A tract and cut them to conserve the last 100 bases at the 3′-end. Finally, sequences were checked for redundancy by intersequence comparisons. Any pair of sequences with identity greater than 90% was considered as the same transcript and one was eliminated from the data set. Therefore, assuming no errors, the 3′-end of the selected EST sequences were considered as the poly(A) site.

Genomic sequences were obtained from E. histolytica genome databases (http://www.tigr.org/tdb/e2k1/eha1/ and http://www.sanger.ac.uk/Projects/E_histolytica/). Loci were randomly chosen throughout the parasite genome to retrieve nt sequence of selected open reading frames (ORFs) including 3′-UTR. This information was used to select DNA sequences corresponding to 180 nt around the predicted stop codon in order to conserve 60 bases of ORF and 120 nt of 3′-UTR for each locus. Finally, sequences were checked for redundancy as described above.

2.2. Sequences analysis of EST and genomic sequences

As we were looking for elements involved in pre-mRNA 3′-end formation, T residues were substituted by U residues in both EST and genomic sequences. Entropy and single nt frequencies were determined using the bioinformatics tools of the BioEdit biological sequences alignment editor (version 7.0.5.3) (http://www.mbio.ncsu.edu/BioEdit/bioedit.html). To identify potential cis-regulatory elements, we obtained all the strings of size four, five and six, formed by A, C, U, G and N (where N is any nt) and sequences were screened for all 4-nt, 5-nt and 6-nt words by brute force string search to find those with the largest frequency. In EST and genomic sequences, motifs position was expressed in relation to the poly(A) site and the stop codon, respectively, which were arbitrarily fixed at position 0. Finally, we used the Quikfold program at http://www.bioinfo.rpi.edu/applications/hybrid/quikfold.php (Zuker, 2003; Markham and Zuker, 2005) to predict transcripts folding for genes with known poly(A) site using defaults parameters.
(37°C, 1 M [Na\(^{+}\)] and 0 M [Mg\(^{2+}\)]); top free energy value was considered to select the most probable structures.

3. Results

3.1. Identification of conserved regions in 3′-end of E. histolytica cDNA sequences

The initial set of 2348 raw cDNA sequences was submitted to successive filtering processes to select EST sequences whose 3′-end was considered as the poly(A) site (position 0). The last 100 nt of the resulting 512 cDNA sequences were then examined for information content through the determination of entropy that is a measure of the uncertainty associated with a random variable. In this approach, information entropy is thought to provide data about the amount of nt variability at each position relative to the poly(A) site in the set of aligned sequences, allowing the measure of the information content at each nt position in the alignment. As shown in Fig. 1A, cDNA sequences could be roughly divided in two distinct regions. From −100 to −50 nt, entropy values were high (mean value: 1.3286), indicating a lack of predictability at an aligned position that is probably due to the fact that this region corresponds to distinct ORF without any conserved motifs. In contrast, entropy values decreased in the last 50 nt towards the poly(A) site (mean value: 1.2887), indicating a higher information content. Remarkably, the entropy plot evidenced three windows spanning from −50 to −25 nt (minimum value: 1.2449), −25 to −15 nt (minimum value: 1.1592) and −15 nt to poly(A) site (minimum value: 1.1991), respectively (Fig. 1A), that could correspond to three conserved domains.

The 512 ESTs were then examined for single nt frequencies at each position relative to the poly(A) site (Fig. 1B). Sequences upstream the poly(A) site generally presented an AU-rich content, with an average A and U residues frequency around 37% and 32%, respectively. A detailed examination of curves revealed four distinct regions: from −100 to −50 nt, nt content was almost constant; from −50 to −25 nt, the abundance of A increased, whereas the U content was kept almost unchanged and GC content decreased; from −25 to −15 nt, nt sequences displayed a combination of A and U residues that could correspond to the UA(U/A)UU polyadenylation signal; and the last 15 nt-region preceding the poly(A) site exhibited a clear peak of U residues. Finally, the last nt at 3′-end was a U residue in more than 40% of EST sequences, suggesting that it is the preferred poly(A) site (Fig. 1B). Interestingly, these four regions were in agreement with those previously determined from the information content analysis (Fig. 1A), confirming their significance.

3.2. Identification of conserved regions in 3′-end of E. histolytica genomic sequences

363 distinct loci were obtained from the parasite genome databases and filtered before evaluating information content and single nt frequencies throughout a180 nt-region around the predicted stop codon (position 0) (Fig. 1C). Entropy profile confirmed the existence of differences between codifying and not codifying regions, the mean entropy value being higher (1.2609) in the 60 nt region upstream the stop codon than in the last 120 nt (1.1549). This confirmed the lack of predictability at an aligned position in distinct ORF and indicated an higher information content in 3′-UTR. Interestingly, entropy values dramatically decreased down to 0.04786 at position 0, in agreement with the relevance of the stop codon. Although the entropy plot did not evidence any specific regions with high information content in 3′-UTR, values were lower throughout the first 60 nt, suggesting the presence of informative elements near the stop codon (Fig. 1C).

Next, we examined the same genomic sequences for single nt frequencies at each position relative to the stop codon (Fig. 1D). The AU content was higher in 3′-UTR (81%) that in coding regions (70%) as previously reported (López-Camarillo et al., 2005). After the stop codon, we observed a short strength of AU residues (4–10 nt) followed by a 50–nt region with a disproportionate frequency of U residues, whereas there was a combination of AU residues and a low GC content throughout the last 60 nt. Remarkably, the A-rich region previously identified from EST analysis was detected upstream the stop codon showing that it did not correspond to a 3′-UTR conserved motif but was located inside codifying sequences (Fig. 1D). Based on results obtained from EST and genomic DNA sequences analysis, we proposed that potential regulatory elements could be located in the 25 nt-region upstream the poly(A) site and in the 60 nt-region downstream the stop codon.

3.3. Identification of conserved elements in 3′-end of E. histolytica cDNA sequences

In order to identify the conserved motifs located before the poly(A) site, we searched for the most representative words located in the last 25 nt of the EST collection. We initially focused on 5-nt motifs because the only reported 3′-end processing element in E. histolytica is the consensus UA(U/A)UU polyadenylation signal (Bruchhaus et al., 1993) (Table 1). A single UAAUU or UAUUU motif was present in 136 (27%) and 57 (11%) sequences, respectively. 68 (13%) sequences presented various UA(U/A)UU motifs, while 252 (49%) sequences did not present UAAUU nor UAUUU motifs.

When we analyzed the global frequency score of the top-10 motifs present at least once in the EST set (Fig. 2A), the UAAUU signal appeared at the 2nd place, whereas the UAAUU motif was at the 13th place (Supplementary Fig. 1). Other recurrent pentamers, such as UUAUU (7th) and AUAUU (9th), could be variants of the reported polyadenylation signals with a single nt changed. The other AU-rich motifs (UUAUU, AUUAU and AUAUA), have more than one substitution. The UUUUU word appeared as the most global common motif, in agreement with the high U content in EST sequences. Considering that words with 4 or 5 U residues are U-rich motifs, we proposed that UUUAUU (7th), AUUUU (8th) and UUUUA (10th), together with the UUUAUU pentamer, could contribute to the formation of the U-rich region identified in Fig. 1B. Although its high abundance, the AAAA motif (4th) was not considered as relevant since it is the only A-rich word detected (Fig. 2A and Supplementary Fig. 1). Analysis of

<table>
<thead>
<tr>
<th>UAAUU</th>
<th>Occurrence</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>252</td>
<td>136</td>
<td>22</td>
<td>2</td>
<td>412</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>57</td>
<td>17</td>
<td>19</td>
<td>2</td>
<td>95</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>313</td>
<td>154</td>
<td>41</td>
<td>4</td>
<td>512</td>
<td></td>
</tr>
</tbody>
</table>

White cell, sequences without any consensus polyadenylation signal; light grey cell, sequences with a single consensus polyadenylation signal; dark grey cell, sequences with multiple consensus polyadenylation signals.
4-nt and 6-nt words occurrence frequencies did not identify other relevant motifs (data not shown).

We next analyzed the relative position of the most representative 5-nt words identified above (Fig. 2B). Interestingly, AU-rich motifs were centered on \(-18\) to \(-16\) nt positions, except for the UUAUU motif that clustered with U-rich motifs around \(-13\) to \(-9\) nt positions. Both regions matched with those previously identified (Fig. 1A and B).

### 3.4. Identification of conserved elements in 3′-end of *E. histolytica* genomic sequences

To characterize 3′-UTR after the poly(A) site, we searched for the most representative 5-nt words in the 60 nt region downstream the stop codon in genomic sequences (Table 2). A single consensus polyadenylation signal, UAAUU or UAUUU, was detected in 46 (13%) and 69 (19%) sequences, respectively. Multiple consensus polyadenylation signal were detected in 188 (52%) genomic sequences, whereas 61 (17%) genes did not have UAAUU nor UAUUU motifs.

Analysis of the global frequency score of the top-10 motifs present at least once in genomic sequences (Fig. 2C), showed that the consensus UAAUU and UAUUU motifs appeared at the 5th and 10th place, respectively. The UAUUU motif that was proposed above to be a variant of the consensus polyadenylation signals with a single nt changed, was at the 4th place, whereas the other recurrent AU-rich motifs with various substitutions, UUAUU and UUAUU, were at the 7th and 9th position, respectively. The UUAUU word (1st) together with other U-rich words, UUUUA, AUUUU, UUAUU, UUAUU and UUAUU (2nd–6th place), contributed to the high U

### Table 2

<table>
<thead>
<tr>
<th>Classification of genomic sequences according to the number of consensus UAAUU and UAUUU polyadenylation signals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Occurrence</td>
</tr>
<tr>
<td>------------</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>

White cell, sequences without any consensus polyadenylation signal; light grey cell, sequences with a single consensus polyadenylation signal; dark grey cell, sequences with multiple consensus polyadenylation signals.
3.6. Predicted secondary structures of 3′-end regions from E. histolytica genes

We explored the formation of higher order structure for E. histolytica transcripts with known poly(A) site using the Quikfold program (Zuker, 2003; Markham and Zuker, 2005) (Fig. 4). For analysis purposes, the input was a 400-nt region where the poly(A) site was at the 300 nt position. Interestingly, secondary structures corresponding to a minimal free energy value (from −56.2 to −82.8 kcal mol$^{-1}$) could be clustered into two previously reported groups (Loke et al., 2005): the poly(A) site was located on a cluster of stem loop structures in 53% of transcripts (group I, represented by the Ehadh3 gene), whereas it was on or around the stem loop, but not flanked by a cluster of secondary structures, in 47% of transcripts (group II, represented by the Ehub gene). Consequently, the stem loop structures found around the poly(A) site were obtained from base pairing of adjacent regions that contain cis-regulating elements for mRNA 3′-end formation (Fig. 4).

4. Discussion

In eukaryotic cells, mRNA polyadenylation is an important gene expression control point that depends on trans-acting factors interacting with specific 3′-UTR cis-acting elements. Considering that it is possible to predict regulatory elements based on their conserved position and nt content in a large set of sequences, several groups have reported the in silico identification of conserved elements that could be involved in mRNA poly(A) tail formation in A. thaliana (Loke et al., 2005), H. sapiens (Hu et al., 2005) and S. cerevisiae (Graber et al., 2002). To our knowledge, the present paper is the first report about the prediction of pre-mRNA 3′-UTR regulatory sequences in a protozoan parasite.

Our analysis of E. histolytica EST sequences evidenced that the last 25-nt region upstream the poly(A) site contains a AU-rich domain corresponding to the consensus UA(A/U)UU polyadenylation signal or variants, followed by a U-rich region preceding the poly(A) site that is generally characterized the U residue. Data obtained from genomic sequences showed that the AU-rich region is located downstream the stop codon and upstream the poly(A) site, and evidenced the presence of a novel A-rich element that is located after the poly(A) site. Analysis of E. histolytica genes with known poly(A) site confirmed this hypothetical array and suggested that the U-rich region could surround the poly(A) site. The fact that the polyadenylation signal was included in the stop codon in
several genes may support the hypothesis that the stop codon is the polyadenylation signal ancestor, dividing the coding and not codifying genetic regions.

The molecular organization of pre-mRNA 3′-end processing signals is roughly conserved through evolutionary scale (Fig. 5). *E. histolytica* presents a U-rich region preceding the poly(A) site, as in plants, human and yeast. Moreover, we detected another U-rich tract downstream the poly(A) site, as in plants, human, yeast and *T. vaginalis*. In addition, distances between each motif are in the same range in all organisms. Remarkably, the novel A-rich region downstream the poly(A) site, seems to be unique for *E. histolytica* transcripts. Interestingly, the polyadenylation signal is the conserved A(A/U)UAAA hexanucleotide in plants, animals and yeast, whereas it appears to be specie-specific in the few protozoan parasites studied, corresponding to AGU(A/G)AAA in *Giardia lamblia* (Peattie et al., 1989; Que et al., 1996), UAAA in *T. vaginalis* (Espinosa et al., 2002), AUUAAA in the *Plasmodium pgs28* gene (Cann et al., 2004) and UA(A/U)UU in *E. histolytica* (Bruchhaus et al., 1993).

The high incidence of the canonical polyadenylation signal UA(A/U)UU in EST and genomic sequences, confirmed the functional relevance of this motif for pre-mRNA polyadenylation. The detection of UA(A/U)UU variants suggested that the core polyadenylation machinery of *E. histolytica* (López-Camarillo et al., 2005) could be able to interact with different polyadenyla-
tion signals. In human, CPSF160 binding to the consensus AUUAAA polyadenylation signal is stimulated by CstF, indicating that CPSF160 alone could be able to recognize divergent polyadenylation signals (Wilusz et al., 1990). The U1-snRNP A-protein (Gunderson et al., 1998; Lou et al., 1998) and hnRNP-F/H (Veraldi et al., 2001) proteins have been reported as modulators of polyadenylation, regulating the selection of the consensus polyadenylation signal by the core polyadenylation factors. The presence of multiple polyadenylation signals in several sequences also suggested that the use of distinct polyadenylation signal could be an important regulatory event for E. histolytica gene expression. Additionally, our data suggested the existence of UA(A/U)UU-independent polyadenylation processes in E. histolytica, adding more complexity to mRNA 3′-end formation mechanism in this protozoan pathogen.

The prediction of RNA secondary structure by energy minimization was used to explore the topological association with functional signals. E. histolytica transcripts folding brings regulatory sequences closer together, probably facilitating interactions between the distinct factors of the pre-mRNA 3′-end processing machinery to perform the polyadenylation reaction. In other organisms, it has been reported that mutations in poly(A) signal reduced polyadenylation efficiency (Mogen et al., 1990; Rothnie et al., 1994), which was associated with structural changes in mRNA folding (Loke et al., 2005). Therefore, structural studies of RNA–protein interactions are of particular interest to understand the pre-mRNA 3′-end cleavage and polyadenylation process.

The present study showed the useful of computer-based methods to determine sequences with a potential role in biological process. Our data will contribute to the understanding of gene expression regulation in E. histolytica, providing new insights into pre-mRNA 3′-end polyadenylation mechanisms in this pathogen. The functional relevance of E. histolytica sequence elements identified here is currently under experimental investigation.

Acknowledgments

We thank Christiane Boucher (Genopole at Pasteur Institute, France) for her help in automated cdNA sequencing. This work was supported by CONACyT, SIP-IPN, COFAA-IPN and UACM (Mexico).

Appendix A Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.compbiolchem.2008.03.019.

References


