A bioinformatics approach to defining the AU-Rich Element (AURE)
Identification of novel motifs flanking AU-Rich Elements

Josh Blake, Timothy J. Gingerich, Jonathan LaMarre

This manuscript was prepared under the supervision of Jonathan LaMarre, Department of Biomedical Sciences, Ontario Veterinary College

Post-transcriptional regulation of genes is critical for the proper temporal and spatial expression of gene products. mRNA transcripts contain cis AU-rich elements (AUREs) in their 3’ untranslated region (UTR) that facilitate the binding of trans acting stability factors. AUREs however remain poorly characterized. Using regular expressions inherent in the PERL programming language, coupled to the Ensemble Human Transcriptome database, we confirm trends observed in already characterized 3’ UTRs as well as identify several significant and novel 3’ UTR flanking sequences that may have biological significance.

The temporal and spatial expression pattern of any gene product is highly regulated by transcriptional, post-transcriptional and post-translational processes. While significant effort has been directed at understanding the mechanisms governing transcriptional and post-translational regulation, the post-transcriptional regulation of mRNA transcripts, a mechanism restricted to higher eukaryotes, has only recently become the focus of intensive research. Within the eukaryotic cell, mRNA transcripts contain two essential elements that prevent their degradation: the 5’ cap and the poly-A tail. Removal of the 5’ cap allows the mRNA transcript to be degraded in a 5’→3’ manner through XRN1 exonuclease. Deadenylation of the poly-A tail, the most common form of mRNA degradation, leaves the mRNA transcript vulnerable to rapid degradation in the 3’→5’ direction via a complex of exonucleases called the exosome. Many important mRNA transcripts contain cis adenylate, uridylate (AU) rich sequences in their 3’ untranslated region (UTR) that coordinate the binding of trans acting factors to facilitate, or impede 3’→5’ exosomal degradation, and was first characterized in granulocyte macrophage colony stimulating factor (GM-CSF) mRNA. Since then, a number of cytokines, growth factors, proto-oncogenes, and transcription factors have also been shown to be regulated by these AU-rich elements (AURE).

A number of AU binding proteins (AUBPs) that interact with AUREs have been identified, including: K homology-type Splicing Regulatory Protein (KSRP), AU Rich Element Degradation Factor 1 (AUF-1), Tristetraprolin (TTP), and finally Embryonic Lethal Abnormal Vision 1 (ELAV-1 / HuR). Of these, KSRP, AUF-1 and TTP mediate the destabilization of mRNA by targeting them for exosome-mediated degradation. Conversely, HuR has been shown to promote the stabilization of mRNA transcripts. Ultimately, the balance between destabilizing and stabilizing AUBPs regulates the overall half-life of mRNA transcript and subsequent temporal and spatial expression of a gene product. AU-rich elements are highly variable in nucleotide sequence, and a clearly defined motif remains to be proposed, however, they prototypically contain an AUUUA pentamer sequence (referred to interchangeably as the pentamer). Classification has typically been based on the pentamer and its flanking sequences. Furthermore, it has been shown that AUREs containing tandem repeats of the pentamer exhibit a higher binding affinity for AUBPs. It has also been considered that 3’ UTRs might facilitate the simultaneous binding of both destabilizing and stabilizing factors. While these mechanisms suggest a complex level of regulation, when considering the multitude of transcripts that are regulated through AUREs, the question arises as to...
what may confer increased AUBP binding specificity to one mRNA transcript over another, and whether this specificity is due to the variation observed within the AURE nucleotide sequence itself.

In an effort to provide a complete and searchable repository for genomic, transcriptomic, and proteomic data, the Ensembl Genome Browser was created in 1999 to coincide with the completion of the human genome project. The present Ensembl database build (build 52) contains the most recent and annotated edition of the human transcriptome (NCBI36). Equally so, Ensembl provides advanced database functionality through the use of the BioMart biological database engine, and allows for the generation of files containing only desired data.

Regular expressions (regex) are abstract terms commonly used to recognize arbitrary data that conforms to a specific pattern (Christiansen). Rather than matching the literal value of a string of characters, regular expressions match a structure. To give an example, the following regex: \w+? will search for and match every word within this document irrespective of the length of each word, or what letters it contains. In the context of bioinformatics, regular expressions can be applied to recognize known motifs and unknown motifs in a quantitative and qualitative manner through the creative employment of regex terms.

Using the PERL programming language and its implementation of regular expressions, coupled to the Ensembl database, we are able to show that 3' UTRs are rich with AUUUA pentamers while simultaneously identifying several significant motifs flanking a pentamer.

METHODS

Ensembl BioMart Data Retrieval
The Homo sapiens NCBI Revision 36 3' UTR transcriptome was defined from the Ensembl Database (Revision 52) by selecting ‘Sequences’ and then by enabling ‘Sequences:3’ UTR’ and ‘Transcript Attributes:Ensembl Transcript ID’. This was exported as a FASTA sequence file with the ‘Unique results only’ tab checked.

Developing A Search Algorithm
All scripts were written using the PERL programming language and run on an Apple MacBook (Model 2,1: 2GHz Intel Core 2 Duo, 1GB RAM) running Mac OS X 10.5.6, and PERL v5.8.8. Statistical calculations were performed in Microsoft Excel 2008 for Mac v12.1.0. An annotated prototype script is presented below to illustrate the rationale (note that minor variations of this script were used depending on how the data needed to be iterated over):

Open the file containing the data, and store the contents as the variable $data
open INPUT, "<ensembl-utr.txt" or die "Can't open file : $!";
while (<INPUT>) { $data.=$_; }
close INPUT;

Split the variable into an array called @data, recognizing '>' as a new entry
@data_array = split /\>/, $data;

Count how many entries are in the array
$count = @data_array;

Iterate over each entry in the @data array
for($i=0; $i<$count; $i++) {
    Split each entry up, recognizing '|' as a delimiter, capture the data as a RegEx
    if ( @data_array[$i] =~ m/(.+)(.+)(.+?)\n(.*)/s ) {
        Save the captured data in variables
        $geneid = $1;
        $transcriptid = $2;
        $nucleotide = $4;

        Remove all whitespace from nucleotide data, including line breaks
        $nucleotide =~ s/\W//gs;

        Match the SEARCHPHRASE and tally it up
        while ( $nucleotide =~ m/$SEARCHPHRASE/g ) {
            $freq{$1}++;
        }
    }
}

Save the output data
open OUTPUT, ">output.search" or die "Cannot open file : $!";
print OUTPUT map { "$_|$freq{$_}

A list of common search phrases is found in Table 1.
Positional Nucleotide Analysis
The nucleotide content (global nucleotide frequency) in all 3’ UTRs was determined, as well as the nucleotide content in only AUUUA pentamer containing 3’ UTRs (pentamer specific nucleotide frequency). If a given nucleotide has no positional relevance, then it is expected to be randomly distributed through all 3’ UTRs at the observed global nucleotide frequency. Since only pentamer containing 3’ UTRs were selected from the entire 3’ UTR dataset, the pentamer specific nucleotide frequency was instead used to estimate the frequency of random distribution of a given nucleotide.

Positional Nucleotide Analysis is performed by identifying each pentamer in a given 3’ UTR, and then tallying up the nucleotides found at a given distance on each side flanking the pentamer. Nucleotide counts and positions were determined for all pentamers in the human transcriptome, up to 200 nucleotides immediately 5’ and immediately 3’ to the AUUUA pentamer.

Estimating Sequence Probabilities
If the individual nucleotide frequencies at any given position are independent of each other, then the probability of finding a given combination of nucleotides is the product of their individual probabilities (termed the expected probability). A given flanking sequence has a length of \( n \) nucleotides. The expected number of transcripts is the product of the actual number of transcripts that have \( n \) flanking nucleotides and the expected probability with 3’ degrees of freedom.

Flanking sequence probabilities were calculated for up to 10 nts 5’ and 10 nts 3’ of a pentamer by examining each flanking sequence as it was found rather than by iterating through \( 2 \times 4^{10} \) combinations. Simultaneously, the number of times each combination was found was added up. The observed frequency and expected frequency were then compared using the CHIDIST() function with 3p degrees of freedom (4 – 1 degrees of freedom per position).

RESULTS
AUUUA-Pentamers Are Found Frequently Throughout the 3’UTR
The initial search of the transcriptome revealed 23 373 transcripts out of 63 281 containing at least one pentamer in their 3’ UTR. Out of these transcripts, 92 815 pentamers were found. The frequency of each nucleotide in all 3’ UTRs was determined, as well as the frequency of each nucleotide in only pentamer-containing 3’ UTRs, in order to establish an average nucleotide frequency (Table 2). The expected frequency of AUUUA pentamers is the combinatorial product of the individual nucleotide frequencies and was calculated to be 1.874 per 1000 nts for all 3’ UTRs, and 2.204 per 1000 nts for pentamer containing 3’ UTRs.

Table 1: Average nucleotide frequency in all 3’ UTRs, and only pentamer containing 3’ UTRs

<table>
<thead>
<tr>
<th>Nucleotide</th>
<th>All 3’ UTRs</th>
<th>Pentamer Containing 3’ UTRs</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Count</td>
<td>Frequency</td>
</tr>
<tr>
<td>A</td>
<td>118620</td>
<td>0.270472111</td>
</tr>
<tr>
<td>G</td>
<td>953705</td>
<td>0.217476508</td>
</tr>
<tr>
<td>U</td>
<td>129289</td>
<td>0.294799515</td>
</tr>
<tr>
<td>C</td>
<td>952795</td>
<td>0.217251867</td>
</tr>
<tr>
<td>(Unknown) N</td>
<td>27</td>
<td>6.15641E-07</td>
</tr>
<tr>
<td>Total</td>
<td>43856714</td>
<td>36234864</td>
</tr>
</tbody>
</table>
pentamer density as high as 62.5 per 1000 nt. Taken together, this data suggests that there are many transcripts containing pentamer-rich 3’ UTRs.

**AUUUA-Pentamer Flanking Nucleotides Are Significant**

Positional nucleotide analysis was performed to determine whether any nucleotides flanking a pentamer were conserved. All pentamers were identified within each 3’ UTR and then up to 200 nucleotides, flanking each of the 5’ and 3’ borders of the core AUUUA pentamer, and were tallied up. The frequencies of each nucleotide were then calculated by dividing the observed count of each nucleotide at each position by the total number of nucleotides at each position. The frequencies were then plotted by position relative to a pentamer (Figure 1) to illustrate positional significance. Intriguingly, large fluctuations are observed within 10 nts of the pentamer (Figure 2). Notably, each nucleotide frequency trends towards the average 3’ UTR nucleotide frequency as the distance from the pentamer increases. In addition, each nucleotide base was plotted separately to identify statistically significant trends specific to each nucleotide frequency and position (Figure 3).

Significant flanking nucleotides were identified as far as 8 nts from the pentamer (3σ) (Table 3). It was found that U at the -1, +1 position occurs more frequently and less due to chance alone (3σ), consistent with already reported data. Intriguingly, it was also found that C at the -1 position occurs more frequently than anticipated (3σ), suggesting a novel yet conserved motif associated with pentamers.

In addition to individual nucleotides having positional significance, it was hypothesized that combinations of nucleotides flanking a pentamer might confer significance. In order to determine this, the combinatorial product was taken based on the observed nucleotide frequency at each position for each possible nucleotide combination up to 10 nucleotides flanking each side of a pentamer. A chi-square test statistic was then calculated by comparing the observed frequency and the expected frequency of a given combination of nucleotides. Theoretically, this could yield up to $4^{10}$ possible combinations flanking each side of the pentamer, however, only 128 759 combinations were observed on the 5’ side of the pentamer and 124 601 combinations on the 3’ side. Of these, 79 combinations were found to have statistical significance 5’ to the pentamer, and 97 were found to have statistical significance 3’ to the pentamer (3σ) as calculated by chi-square. The data was classified by likelihood based on the difference between the observed and expected frequencies and chi-square distribution probability. Consistent with known AU-rich elements, poly-U and poly-A tracts were found significantly more 3’ to a pentamer than expected. Intriguingly, CC was the most likely to be found 5’ to a pentamer than expected, and the most likely to be found 3’ to a pentamer after the poly-U and poly-A sequences. Interestingly, CG flanking either side of the pentamer was the least likely to be found when compared to its expected frequency.

**DISCUSSION AND CONCLUSION**

It should be noted that the Ensembl 52 database (Human NCBI Rev. 36) used as the working dataset for this project contained all annotated transcript variants currently reported. Consequently it must be acknowledged that one gene with many transcript variants may introduce bias through redundancy when considering conserved regions in the 3’ UTRs. However, it can be equally argued that such conserved regions are important in the definition of an AU-rich element in that they have remained conserved for a biological reason, and thus should not be considered bias. Further, it should be noted that 3’ UTRs were defined per the annotation provided by the ENSEMBL 52 database, and individual transcripts could be subject to mis-annotation and premature termination of the 3’ end of the transcript (an error encountered when generating cDNA from mRNA transcripts). To reinforce the findings noted above, and to eliminate bias due to scientific error, analysis of the ENSEMBL 52 (NCBI37) mouse 3’ UTR transcriptome should be performed. As previously discussed, the binding affinities of AU-rich binding proteins confer, at least in part, the biological half-life of the transcript. It is important to consider both the
Figure 1: Nucleotide frequencies within 200 nts, flanking an AUUUUA pentamer. Position 0 represents the pentamer.

Figure 2: Nucleotide frequencies within 50 nts, flanking an AUUUUA pentamer. Position 0 represents the pentamer.
absolute number of AUREs and their positional relationship to each other. Multiple AUREs can facilitate the simultaneous binding of both stabilizing and destabilizing factors while pentamer density confers a degree of competitive binding. It must be acknowledged that 3’ UTRs are rarely homogeneous, and thus pentamer density should only be used comparatively to indicate the average pentamer spacing in a 3’ UTR. Equally so, pentamer density may not accurately represent clustering of pentamers within the 3’ UTR, a phenomenon that may prove important to both the binding affinities and binding nature associated with that element. On the whole, it may prove important to classify transcripts by AURE composition, absolute pentamer quantity, and pentamer density in an effort to predict the biological half-life associated with a given transcript rather than just the AURE composition alone.

The above data shows consistency with previous definitions of AU rich elements in addition to identifying several significant and novel motifs associated with a pentamer. What remains to be explored is the biological significance in vivo associated with each of these motifs, and whether the discrepancy between the observed and expected frequencies of these flanking sequences are due to biological selective pressures, or to chance alone. Interestingly, the significance of CC flanking either side suggests that the symmetry inherent in an AUUUA pentamer may be disrupted or extended by the flanking sequences, and that the binding directionality of an AUBP may have a functional role in AUBP mediated mRNA regulation.

Arguably, the power inherent in regular-expression facilitated data mining has become apparent in the context of pattern recognition and quantification. When considering 3’ UTRs specifically, regular-expression facilitated identification of AUUUA pentamers produces a rapid and flexible way to search large datasets for variable nucleotide sequences. In order to extend this functionality to the
scientific community, the AU-Rich Element Analytical database (AURAdb), a publically available online database, has been created that takes advantage of these search features. The algorithms used in this analysis are not limited to the 3' UTR transcriptome, and can, with minor modifications, be used to search other datasets including the 5' UTR transcriptome, a region that has also been implicated in mediating RNA stability. In addition to confirming trends already accepted in known AU-rich elements, a number of novel AURE characteristics have been identified that can serve as the basis for further characterization.

REFERENCES


4. Christiansen, T. Perl regular expressions.


