Bioinformatic analysis of post-transcriptional regulation by uORF in human and mouse

Motomu Matsui[a,b], Nozomu Yachie[a,b], Yuki Okada[a,b], Rintaro Saito[a,c,*], Masaru Tomita[a,b,c]

[a] *Institute for Advanced Biosciences, Keio University, Tsuruoka 997-0035, Japan*
[b] *Bioinformatics Program, Graduate School of Media and Governance, Keio University, Fujisawa 252-8520, Japan*
[c] *Department of Environment and Information Studies, Keio University, Endo 5322, Fujisawa, Kanagawa 252-8520, Japan*

Received 11 January 2007; revised 3 July 2007; accepted 24 July 2007

Available online 3 August 2007

Edited by Robert B. Russell

Abstract RNA decay is thought to exert an important influence on gene expression by maintaining a steady-state level of transcripts and/or by eliminating aberrant transcripts. However, the sequence elements which control such processes have not been determined. Upstream open reading frames (uORFs) in the transcripts of several genes are reported to control translational initiation by stalling ribosomes and thereby promote RNA decay. We therefore performed bioinformatic analysis of the tissue-wide expression profiles and mRNA half-life of transcripts containing uORFs in humans and mice to assess the relationship between RNA decay and the presence of uORFs in transcripts. The expression levels of transcripts containing uORF were markedly lower than those not containing uORF. Moreover, the half-life of the uORF-containing transcripts was also shorter. These results suggest that uORFs are sequence elements that down-regulate RNA transcripts via RNA decay mechanisms.

Keywords: Upstream open reading frames (uORF); RNA decay; Post-transcription control; 5′ Untranslated region; Kozak consensus

1. Introduction

RNA decay plays a major role in the post-transcriptional control of gene expression, maintaining the balance between the synthesis and degradation of RNA transcripts [1]. Although the complex processing pathways involved in RNA degradation have been described, the key determinants of instability remain unclear. Previous research has suggested that longer mRNA transcripts are less stable [2]. However, transcript stability was recently shown to not be correlated with overall transcript size, length of poly (A) tract, number of ribosomes, expression level, or codon usage [3].

The nonsense-mediated mRNA decay (NMD) pathway is thought to be an important surveillance mechanism that promotes the degradation of aberrant transcripts coding for non-functional or harmful proteins [4–6]. Nonsense or frameshift mutations introduce premature translation termination codons (PTCs) into the open reading frames (ORFs) of mRNAs and are a common cause of genetic disorders. PTCs usually lead to rapid mRNA degradation by NMD, which affects decapping, deadenylation, and 5′ → 3′ exonucleolytic activities [7]. Upstream open reading frames (uORFs) are small open reading frames located in the 5′ untranslated regions (5′ UTR) of mRNA and also have important post-transcriptional effects. They are believed to function via *cis*-acting peptide products that reduce the initiation of translation of downstream ORFs by stalling the ribosome at the end of the uORF, thereby exposing the mRNA to degradation [8,9]. Genome-wide comparison of the human and mouse genomes suggest that the majority of uORFs are strongly conserved at the peptide level [10,11]. Furthermore, expression of the proteins encoded by human uORFs has been confirmed by mass spectrometry [9]. However it is not clear how pervasive a role the uORFs play in RNA decay. Thus, for example, a small peptide encoded within the 5′UTR of Yap2 mRNA modulates NMD in yeast [12] whereas the uORF encoded by the upstream region of the cytokine thrombopoietin transcript has been shown to not induce NMD in humans [13].

In this study we used a bioinformatics approach to assess whether uORFs in general affect RNA degradation. We first predicted the uORFs in the human and mouse transcriptomes and then compared the tissue-wide expression profile and decay rate of uORF-containing and non-uORF-containing transcripts. We found that the average level of expression of uORF-containing transcripts was markedly lower than that of the non-uORF-containing transcripts, and their decay rates were higher.

2. Materials and methods

2.1. Prediction of uORF-regulated transcripts

The human and mouse transcripts in the RefSeq database (release 23, accessed 23 May 2007) and human UniGene database (build 202, accessed 23 May 2007) were obtained via the National Center for Biotechnology Information (NCBI) ftp server (ftp://ftp.ncbi.nlm.nih.gov). Then all the transcripts were categorized into four levels through the following steps, as summarized in Fig. 1.

First, transcripts having no definite CDS annotations were eliminated. Then all of the longest ORFs satisfying the following three conditions were defined as uORFs; (1) the ORF (AUG) started in the 5′UTR, (2) the end of the ORF was not identical to the stop codon of the annotated downstream CDS, and (3) the end of the ORF was

[*] Corresponding author. Address: Department of Environment and Information Studies, Keio University, Endo 5322, Fujisawa, Kanagawa 252-8520, Japan. Fax: +81 466 47 5099.
E-mail address: rsaito@sfc.keio.ac.jp (R. Saito).

Abbreviations: ORF, open reading frame; uORF, upstream open reading frame; NMD, nonsense-mediated mRNA decay; PTC, premature termination codon

0014-5793/$32.00 © 2007 Federation of European Biochemical Societies. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.febslet.2007.07.057

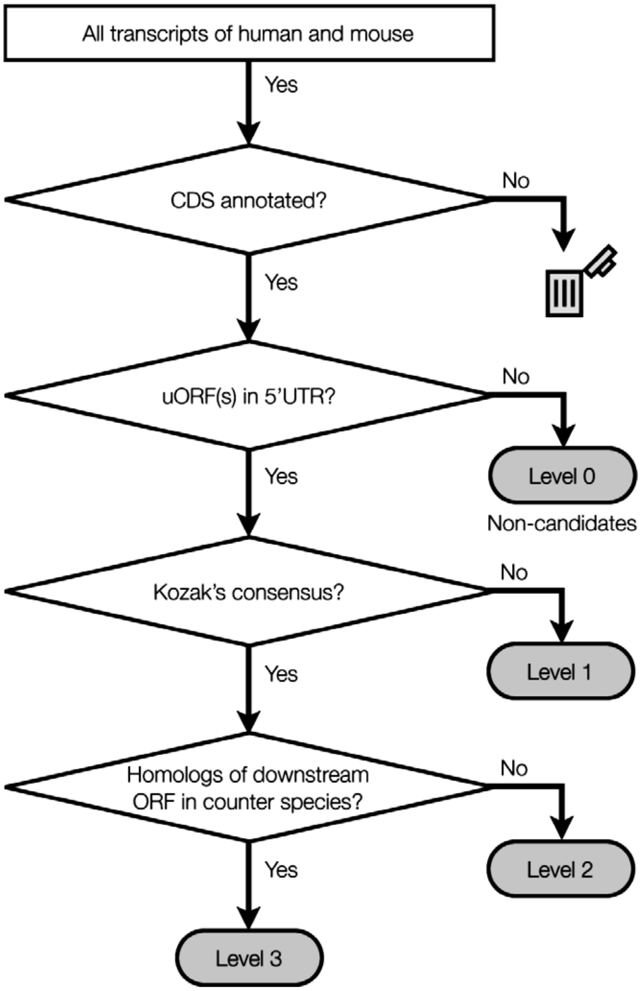

Fig. 1. Schematic representation of the categorization of the uORF-regulated candidates into four levels.

not in the 3′UTR [14]; the transcripts having no uORF were defined as "level 0" uORF candidates (non-candidates) and those having uORF(s) were defined as "level 1" candidates.

The efficiency of translation initiation from a given AUG codon is determined in part by the local sequence context around that codon [15,16]. Therefore, we selected those which had the Kozak consensus motif, 5′-[G/A]CC(AUG)G-3′ (the sequence in parentheses denotes the start codon), which is the most efficient context for the start codons of true ORFs, and re-defined them as "level 2" candidates.

We further selected those whose downstream ORFs were conserved in the human and mouse from the candidates in level 2, and re-defined these as "level 3" candidates. We used the BLAT program [17] with the E-value cutoff value 1e–50 to search for homologous gene pairs in the human and mouse. In other words, both transcripts in the human and mouse homologous pairs have uORFs in level 3.

In order to consider the effect of natural sense–antisense transcripts to gene expression, we checked the presence of overlapping transcripts in the antisense strand of the transcripts in each of the four levels. The dataset of 1233 natural antisense transcripts of humans and 4398 of the mouse were downloaded via the NATsDB website (http://natsdb.cbi.pku.edu.cn/, accessed 23 May 2007).

2.2. Preparation of data for comparing the expression intensities and half-lives of mRNA transcripts

We obtained 4935 RNA expression profiles from 79 different human tissues and 16617 profiles from 61 different mouse tissues, annotated with RefSeq IDs, from the SymAtlas database (http://symatlas.gnf.org/, accessed 23 May 2007) [18]. The expression profiles in the SymAtlas database were examined using high-density oligonucleotide arrays, and the custom arrays were generated using a non-redundant set of documented and predicted genes compiled from RefSeq [19], Celera [20], Ensembl, and RIKEN [21]. We used the expression data normalized by the gcRMA algorithm [22,23].

The decay rates of human mRNAs were obtained from a previous study which investigated the decay rates of 5245 individual transcripts in human cells [24]. The data were downloaded from the Genome Research website (http://www.genome.org/) and we used decay data on 2948 transcript IDs matching those in the UniGene database.

3. Results and discussion

We divided mRNAs into four categories, i.e. level 0, level 1, level 2 and level 3, according to the uORF predictions, the presence or absence of Kozak consensus motifs around the start codons and the conservation of uORF between the human and mouse (conservation of sequence patterns of uORFs were not considered). Among the 38927 human and 46627 mouse mRNAs obtained from the RefSeq database, and the 6731038 human mRNAs from the UniGene database, the CDSs of 33670, 42934, and 58745 mRNAs, respectively, were unambiguously annotated. Using the RefSeq data, we extracted candidates for uORF-regulated transcripts; 13174, 12711, 242 and 365 of these were classified into levels 0–3, respectively, in the human case. In the mouse, 15198, 14424, 263 and 440 candidates were classified into each of these four levels. Similarly, we extracted candidates from the UniGene database and the number of transcripts which were classified into levels 0–3 were 53137 39429, 820 and 1146, respectively. In addition, we extracted mRNAs which had a natural antisense transcript (NAT) in the RefSeq and UniGene databases to investigate the relationship between the presence of NATs and gene expression. Approximately, one-fourth of mRNAs in the human and mouse exhibited a NAT regardless of the level they were classified. In other words, the presence of NATs and uORFs in the transcripts were independent. Approximately, half of all the mRNAs were predicted to be uORF-regulated, in agreement with previous work using human, mouse, and rat mRNA sequences in the RefSeq database [10,11]. We speculate that the majority of the predicted uORFs possess the potential to be scanned by ribosomes, and so to control the translation reinitiation of downstream ORFs [12] by stalling and occupying ribosomes at their stop codons [12,25], acting in *cis* at the peptide level [11], or promoting NMD [6]. The RefSeq database contains entries whose transcription initiation sites were not defined [21,26], and many of the 5′UTRs were not completely identified. It is possible therefore that some of the transcripts categorized as level 0 candidates might actually contain uORFs. However, as the aim of this study was to compare the overall trends of the expression intensities and half-lives of the different categories using massive and statistical approaches, we believe that some small number of mis-predictions of uORF-regulated genes should not have a major impact on the results. Similarly, some of the candidate transcripts (levels 1–3) may not be regulated by uORFs but this again should not influence our results.

We compared the amount of transcripts within the corresponding categories of human and mouse mRNA using the SymAtlas database. The numbers of transcripts in each category (that were) cross-linked between the RefSeq and the SymAtlas database are given in Table 1. There were no marked