ABSTRACT

Motivation: The translation start site plays an important role in the control of translation efficiency of eukaryotic mRNAs. However, mRNAs with a suboptimal context of start AUG codon are relatively abundant. It is likely that at least some mRNAs with suboptimal start codon context contain the other signals providing additional information for efficient AUG recognition.

Results: Frequency of AUG codons at the beginning of the coding part of eukaryotic mRNAs was analyzed in relation to the context of translation start codon. It was found that the observed downstream AUG content in the mRNAs with optimal start codon context was close to the expected value, whereas it was significantly higher in the mRNAs with a suboptimal context. It is likely that downstream AUG codons can often be utilized as additional start sites to increase translation rate of mRNAs with a suboptimal context of the annotated start codon and many eukaryotic proteins can be characterized by some N-end heterogeneity.

Contact: ak@bionet.nsc.ru

INTRODUCTION

Translation of most eukaryotic mRNAs is likely to be initiated by a linear scanning, although some other mechanisms are also possible (Kozak, 2002). According to the scanning model, 40S ribosomal subunits can either initiate translation at 5′-proximal AUG codon or miss it and initiate translation at downstream AUG(s). The initiation/scanthrough ratio depends on both the AUG nucleotide context and the features of downstream mRNA fragment (Kozak, 2002; Wang and Rothnagel, 2004). For mammalian and plant mRNAs, the most crucial elements of AUG context are purine at position −3 and guanine at position +4 (Łukaszewicz et al., 2000; Kozak, 2002).

One might expect that mRNA should possess the features providing efficient translation, including the recognition of a genuine translation start site (TSS). However, the fraction of eukaryotic mRNAs with the start AUG codon in a suboptimal context is relatively large as well as the fraction of mRNAs with the AUG-containing 5′-untranslated regions (5′-UTRs) (Rogozin et al., 2001). It is likely that at least some mRNAs with a suboptimal start codon context contain the other signals providing additional information for efficient TSS recognition (Kozak, 2002; Kochetov et al., 2003; Shabalina et al., 2004). To verify this assumption, we performed the comparative computational analysis of eukaryotic genes with either optimal or suboptimal start codon contexts. It was found that eukaryotic mRNAs with suboptimal TSSs were characterized by a significantly higher occurrence of in-frame AUG codons at the beginning of mRNA coding part. It is likely that closely located additional start codons can often be used to increase the translation initiation efficiency and the corresponding proteins might display certain N-end heterogeneity.

METHODS

Overall, 26,225 EMBL entries were obtained at http://srs.ebi.ac.uk/ using the following search fields and terms: ‘Organism’, Arabidopsis thaliana; ‘Molecule’, mRNA; ‘FtKey’, CDS (coding DNA sequence); and ‘Description’, complete CDS. Of these, 12,632 sequences contained both complete coding parts and 5′-UTRs longer than 20 nucleotides. mRNAs of Homo sapiens (29,632), Mus musculus (17,298), Aves (1506), Liliopsida (3423) and Arthropoda (5131) were extracted in a similar way. AUG frequencies were calculated in the 40 5′-proximal codons of annotated coding sequences. Statistical difference in positional AUG frequencies between the mRNA samples with optimal and suboptimal start codon contexts was evaluated with t-test.

RESULTS

Nucleotide frequencies at start codon context positions (not shown) coincided well with the data reported earlier (Rogozin et al., 2001). It was shown that pyrimidine at position −3 upstream AUG decreased the efficiency of its recognition by 40S ribosomal subunits and resulted in an alternative initiation of translation at downstream AUG(s) in mammalian and plant cells, although the recognition/scanthrough ratio varies considerably and depends on the nucleotides at the other context positions, namely, −2, −1, and +4 (Łukaszewicz et al., 2000; Kozak, 2002). Thus, two classifications of AUG contexts were used: (1) based on the important position −3 [pyrimidine (Py) corresponding to suboptimal and purine (Pu) to optimal contexts] and (2) based on the extended context (consensus versus anticonsensus, corresponded to extended optimal and extended suboptimal contexts, respectively; Table 1).

The frequencies of in-frame AUG codons were calculated at the beginning of coding sequences in samples with optimal and suboptimal start codon contexts; the results for the largest human mRNA set are shown in Figure 1. It was found that human mRNAs with suboptimal start codon contexts were characterized by a
Table 1. Consensus and anticonsensus sequences of start codon contexts of eukaryotic mRNAs

<table>
<thead>
<tr>
<th>Organism/taxon</th>
<th>A.thaliana</th>
<th>Lilopsida</th>
<th>Arthropoda</th>
<th>H.sapiens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pos.</td>
<td>−3</td>
<td>−2</td>
<td>−1</td>
<td>+4</td>
</tr>
<tr>
<td></td>
<td>+4</td>
<td>−3</td>
<td>−2</td>
<td>−1</td>
</tr>
<tr>
<td>Consensus</td>
<td>Pu AAGP u CCGP u AA— Pu CCG</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anticonsensus</td>
<td>Py notA notA notG Py notC notC notG Py notA notA — Py notC notC notG</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Context organization of *M.musculus* and *Aves* was identical to *H.sapiens* (not shown).

Pu, purine; Py, pyrimidine.

Fig. 1. AUG codon frequency at the beginning of coding parts of the human mRNAs with optimal and suboptimal start codon contexts (Opt, purine at position −3; Sub, pyrimidine at position −3; Ext Opt, Pu C AUG G; Ext Sub, Py notC notC AUG notG). AUGexp (expected AUG frequency) was calculated in the assumption of a random codon choice (0.016).

considerably higher frequency of in-frame AUGs within 5′-proximal CDS fragment spanning ∼14 codons. The difference was greater between the mRNA samples with extended optimal and extended suboptimal contexts. Note that downstream AUG frequencies in mRNAs with optimal TSSs were close to (or less than) the expected value calculated in the assumption of a random codon choice (Fig. 1).

Average frequencies of downstream in-frame AUG codons lying between third and ninth codons in either optimal or suboptimal contexts were calculated separately. It was found that a suboptimal TSS context correlated with the higher frequency of in-frame downstream AUG codons in the optimal context [71% versus 64% for mRNAs with the optimal TSS; further downstream (between 30th and 40th codons) the difference between the samples was negligible (54% versus 53%)].

I calculated the average frequency of AUG codons at the beginning of CDS (from 3rd to 9th codons) of mRNAs of well-investigated organisms (*H.sapiens*, *M.musculus* and *A.thaliana*) and samples of some other taxa (*Aves*, *Lilopsida* and *Arthropoda*). In general, the mRNAs with suboptimal TSSs were characterized by a significantly higher average downstream AUG frequency than the mRNAs with optimal TSSs or the average AUG frequency either expected in the assumption of a random codon choice (0.016) or calculated in mRNA fragments located further downstream (from 30th to 40th codons; Table 2). Unlike that of downstream AUGs located in-frame with CDS, the average frequency of out-of-frame AUG triplets was lower at the beginning of CDS and did not correlate with the context of translation start site (Table 2).

DISCUSSION

It may be assumed that downstream in-frame AUG codons can be used as additional start sites to recruit 40S ribosomal subunits missing the proximal start codon in a suboptimal context (Kozak, 2002). The usage of a closely located downstream in-frame AUG codon as an alternative TSS could result in additional synthesis of a slightly truncated protein variant with the same functions as its annotated counterpart: it could increase the rate of protein synthesis and prevent the initiation of translation at out-of-frame AUGs (Kozak, 2002; Kochetov et al., 2003). N-truncated protein forms can also possess new functional properties (Kochetov and Sarai, 2004). However, the actual role of this mechanism was not systematically studied, and the AUG codons at the beginning of CDS were not commonly taken into account in the evaluation of mRNA translation efficiency and coding potential.

To test this assumption, I analyzed the samples of eukaryotic mRNAs with optimal and suboptimal contexts of the annotated start codons. If the downstream AUG codons are utilized as alternative TSSs, their occurrence has to correlate with the start codon context: suboptimal context should be accompanied by a higher frequency of downstream AUGs. It was found that mRNAs with optimal start codon contexts were characterized by downstream AUG frequencies...
close to the expected values, whereas suboptimal context of translation start codon correlated with a significantly higher frequency of downstream AUGs within the region spanning ~14 5′-terminal codons (Fig. 1; Table 2). Note that the average frequency of out-of-frame AUG triplets in human mRNAs was lower within 5′-terminal CDS region and did not depend on the start codon context (Table 2).

This observation coincides well with the hypothesis about functional significance of downstream AUGs: it is likely that such a mechanism is used to increase translation initiation of some open reading frames with a suboptimal context of the 5′-proximal AUG codon, and some eukaryotic proteins are heterogeneous at their N-ends. For example, 903 of 5122 human mRNAs with suboptimal start codon contexts contained at least one AUG codon within the eight downstream positions.

Unlike other taxa, Arabidopsis mRNAs with either optimal or suboptimal start codon contexts were characterized by the downstream AUG frequencies exceeding the expected value (note that the difference between mRNA samples with optimal and suboptimal TSSs was significant in a few positions; Table 2). It can be assumed that some other mRNA features influence the Arabidopsis AUG recognition (Pesole et al., 1999; Sawant et al., 2001; Kochetov et al., 2003; Shabalina et al., 2004) and the classification of Arabidopsis TSSs into ‘optimal’ and ‘suboptimal’ used in this study was not fully correct.

Many eukaryotic mRNAs collected in databanks are characterized by AUG-containing 5′-UTRs and a suboptimal context of annotated start codon (Rogozin et al., 2001). It is suggested that in some mRNAs TSSs were annotated incorrectly since the accurate start codon prediction is a complex and still unresolved problem (Casadei et al., 2003; Porcel et al., 2004). Suboptimal context of annotated start codon and the frequency of downstream in-frame AUG codons can be associated with the reasons other than the usage of downstream in-frame AUGs as alternative TSSs. In the context of this assumption, some mRNAs can be characterized by complex translation initiation signals containing two or more AUG codons and their translational efficiencies and coding potentials should be re-evaluated. Recent investigation showed that many eukaryotic genes yielded transcript(s) that translate into several and often very numerous families of polypeptide species (Kettman et al., 2002). Initiation of translation at downstream AUG codons could also result in a synthesis of additional protein variants possessing new functional properties. Study of the mRNA features correlating with the TSS ‘strength’ (including those described here) could be useful for evaluation of translation initiation rate and new functional forms of eukaryotic proteins.

It is known that mRNAs with AUG-containing 5′-UTRs can be translated [albeit at a lower level (Kozak, 2002; Wang and Rothnagel, 2004)] and the negative 5′-UTR and TSS characteristics can be of functional importance (Rogozin et al., 2001). In my opinion, it is unlikely that a correlation between the context of annotated translation start codon and the frequency of downstream in-frame AUG codons was calculated between 30th and 40th triplets. mRNA samples with optimal and suboptimal start codon contexts were characterized by similar out-of-frame AUG frequencies in this region (0.28 and 0.26, respectively).

### Table 2. Average frequencies of AUG codons within positions 3–9 of eukaryotic mRNA coding parts displaying either optimal (Opt) or suboptimal (Sub) start codon contexts

<table>
<thead>
<tr>
<th>Organism (taxon)</th>
<th>TSS context (Pos. –3)</th>
<th>Significant positions</th>
<th>Extended TSS context</th>
<th>Significant positions</th>
<th>AUGe</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sub</td>
<td>Opt</td>
<td>Sub</td>
<td>Opt</td>
<td></td>
</tr>
<tr>
<td>H. sapiens</td>
<td>0.025</td>
<td>0.016</td>
<td>2–9, 11, 13, 14</td>
<td>0.031</td>
<td>0.012</td>
</tr>
<tr>
<td>M. musculus</td>
<td>0.029</td>
<td>0.016</td>
<td>2–11, 14</td>
<td>0.038</td>
<td>0.015</td>
</tr>
<tr>
<td>A. thaliana</td>
<td>0.042</td>
<td>0.019</td>
<td>2–4</td>
<td>0.063</td>
<td>0.020</td>
</tr>
<tr>
<td>Arthropoda</td>
<td>0.038</td>
<td>0.019</td>
<td>2–5, 9, 14</td>
<td>0.042</td>
<td>0.019</td>
</tr>
<tr>
<td>A. halobateria</td>
<td>0.022</td>
<td>0.023</td>
<td>13</td>
<td>0.031</td>
<td>0.022</td>
</tr>
<tr>
<td>Liliopsis</td>
<td>0.027</td>
<td>0.021</td>
<td>3, 5</td>
<td>0.041</td>
<td>0.017</td>
</tr>
<tr>
<td>H. sapiens</td>
<td>0.018</td>
<td>0.017</td>
<td>4</td>
<td>0.019</td>
<td>0.012</td>
</tr>
</tbody>
</table>

*Average frequency of AUG codons in CDS fragments located further downstream of the translation start site (from 30th to 40th codons).

**Positions within first-proximal 15 codons where the AUG frequencies in the mRNAs with suboptimal start codon contexts were significantly higher than in the mRNAs with optimal start codon contexts (according to t-test, P < 0.05).

**Average frequencies of AUGs located in +1 frame in respect to CDS (e.g., NAT GNN) in the region from third to ninth triplets.

*Average frequency of AUGs located in +1 frame in respect to CDS (e.g., NAT GNN) calculated between 30th and 40th triplets. mRNA samples with optimal and suboptimal start codon contexts were characterized by similar out-of-frame AUG frequencies in this region (0.28 and 0.26, respectively).

**Correlation of the start AUG codon in some cases can be doubtful and a downstream AUG can be a more appropriate TSS than the annotated start site. This could result in some overestimation of in-frame AUG frequency downstream of a suboptimal TSS. However, in my opinion such closely located upstream in-frame AUG codons lying in a suboptimal context can also be considered alternative translational start sites.

### ACKNOWLEDGEMENTS

I thank N.A. Kolchanov and I.B. Rogozin for helpful discussions and the Reviewers for their useful comments. This work was supported by the Russian Foundation for Basic Research (grant no. 02-04-48508) and the Programs of Russian Academy of Sciences (Origin and Evolution of Biosphere and Dynamics of Plant, Animal, and Human Gene Pools). I thank SD RAS Complex Integration Program (No. 59), Ministry of Education (grant no. PD02-1.4-464), and Ministry of Industry, Science and Technologies of the Russian Federation (grant no. 2275.2003.4) for partial support.

### REFERENCES


