Context-dependent selection of upstream start codons in *Homo sapiens* lineage

Svetlana Iarovenko
Faculty of Bioengineering and Bioinformatics Lomonosov Moscow State University, e-mail: svetyalana@fbb.msu.ru

Stepan Denisov
IITP RAS, e-mail: stepadenisov@gmail.com

Originally ribosomes were believed to translate only long protein-coding open reading frames (ORF). Yet more than 50% transcripts according to ribosome profiling data undergo translation in 5' untranslated regions (5'UTR). These translated upstream regions are called upstream ORF (uORF). Moreover, over a quarter of such upstream translations appeared to be initiated by AUG triplet (upstream AUG or uAUG) [1]. This start codon can be found in three main configurations: uAUG can have an upstream stop codon pair (uSTOP) and be a part of uORF, be in-frame with the main start of translation (in-frame ORF or iORF) or overlap with main coding sequence (overlapping ORF or oORF). Nearly 100 uORF were experimentally shown to have a regulatory function, assuming a crucial role of this region in translation regulation [2].

The main aim of this research work was to assess the influence of different uAUG configurations on protein expression using evolutionary approach. This approach is comprised of measuring frequency of uAUG acquisition, estimating negative selection on uAUG and evaluating frequency of transitions between different configurations.

Frequency of uAUG "birth" is described by uAUG "birth rate" measure, i.e. how many preAUG (codons that differ from AUG in one nucleotide) mutated to AUG in human genome normalized on total number of preAUG in ancestor genome (see below). This measure describes the "harmfulness" of initiation codon acquisition in 5'UTR.

$$\nu_{uAUG-birth} = \frac{preAUG \rightarrow uAUG}{preAUG}$$

For assessing negative selection on different uAUG configurations the similar measure called uAUG "death rate" is used (number of mutated AUG in human normalized on total number of AUG in ancestor).

Sequences of *Homo sapiens*, *Macaca mulatta* and *Callithrix jacchus* from multiple

alignment of 46 vertebrates from UCSC were selected for this comparison analysis. Hg19 assembly of human genome was used. Annotation of transcription and translation starts was taken from RefSeq. The search procedure of uAUG was performed on protein-coding genes with first exon that contains whole 5'UTR.

All three possible events (emergence and death of uAUG in human genome as part of oORF, iORF or uORF) were studied separately. In each case the concurrence of a triplet (preAUG or AUG) at the same position in *Macaca mulatta* and *Callithrix jacchus* sequences was considered as ancestral state. Other possible codons except for stop-codons were used for control.

The "birth rate" of uAUG as a part of uORF has been found to be higher than in cases with no uSTOP (0,003 versus 0,002 and 0,001 for iORF and oORF). Thus, uSTOP may reduce the "harmfulness" of uAUG. Furthermore, the decreased emergence of uAUG in general compared to other codons has been observed, assuming the selection against this triplet in 5'UTR region (fig. 1).

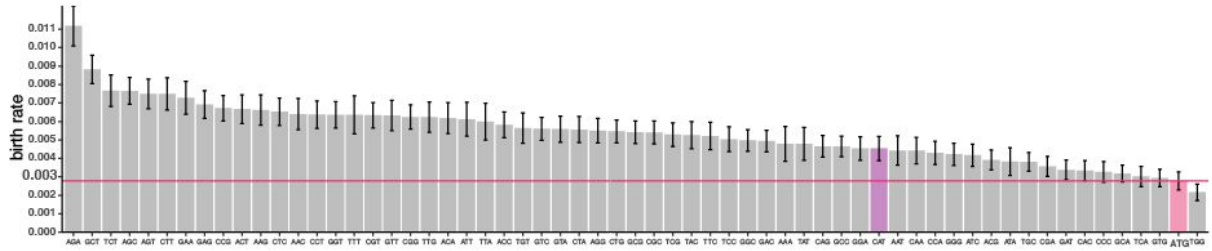

Figure 1. uAUG "birth rate" as a part of uORF.
Pink color highlights the AUG birth rate , violet — birth rate of CAU (a complementary to AUG triplet)

Negative selection has been the most vivid for uAUG in iORF configuration (fig. 2), assuming negative influence on the main protein product.

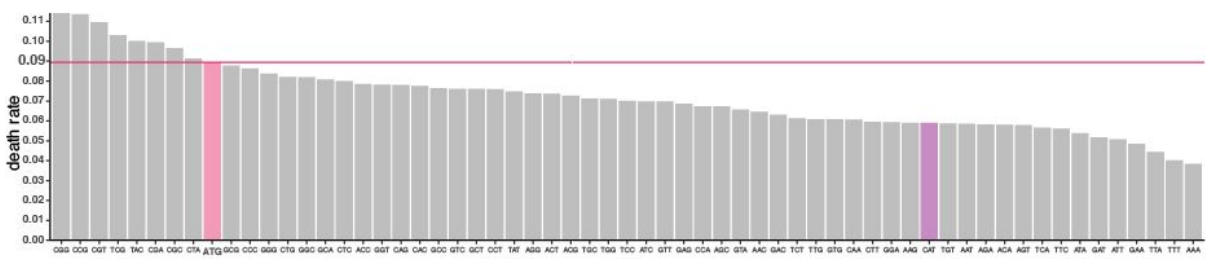

Figure 2. uAUG "death rate" as a part of iORF.
Pink color highlights the AUG birth rate , violet — birth rate of CAU (a complementary to AUG triplet).

Further it is planned to analyze the strength of selection against uAUG depending on the distance between uAUG or uSTOP and main start-codon and to reconstruct evolution dynamics of transitions between different uAUG configurations.

1. S. Lee, B. Liu, S. Lee, S.-X. Huang, B. Shen, and S.-B. Qian (2012) "Global mapping of translation initiation sites in mammalian cells at single-nucleotide resolution," *Proc. Natl. Acad. Sci.*

2. K. Wethmar (2014) "The regulatory potential of upstream open reading frames in eukaryotic gene expression," *Wiley Interdiscip. Rev. RNA*.