The analysis of upstream open reading frame using comparative genetics.

Prosvirov K.A, Mironov A.A., Soldatov R.A.
Faculty of Bioengineering and Bioinformatics, Lomonosov MSU
IITP

According to classical translation initiation mechanism, ribosome scans iRNA starting from cap and initiate translation when meets the first start codon – AUG. However, half of genes have start codons in 5'UTRs known as upstream AUG or shortly uAUG. Ribosome profiling shows the active translation of the vast majority of upstream reading frames starting from uAUG. ~80 % of uAUGs starts with inframe upstream STOP (uSTOP), forming upstream open reading frame – uORF. There are several examples, when uORF has a regulatory role in different cell processes like cell stress.

Although ~40 % of genes have uORF, only ~150 of them have selection of aminoacid sequences which can be seen withing interspecies comparison. From the other hand, AUG is the most conservative triplet in 5'UTR, therefore there are ~10 % of uAUG which are conserved due to the reason of translation. Thereby, our hypothesis is that there exist more conserved uORFs and uAUGs, than there are uORFs with the selection of aminoacid sequence. It is also proven by the fact that there are functionall active uORF but with no cosnerved sequence. Our goal is to estimate the number of functional uORFs and if it is possible to create the list of conserved uORFs and the ratio of functional active among them.

When it comes to the relative position of uAUG in 5'UTR, there are 3 possible variants:

- it has inframe uSTOP therefore forming uORF
- it has stop codon which is located in CDS therefore forming overlapping reading frame – oORF
- uAUG is inframe with the actual translation starting codon – iORF

Before making any predictions, transcripts from USCS Genome Browser were filtered

according to the following rules:

- all the transcripts for one gene should have the same TSS and translation starting site
- the length of 5’UTR lays between 15 and 500
- whole 5’UTR is located inside the first exon

So after the filtration we have 6170 genes. Results obtained on this data set is presented below:

- There is a negative selection on uAUG and uORF
- The longer 5’UTR is, the more it is likely to have single uUAG which is the part of functional element (e.g. uORF/oORF)
- The probabilites of uAUG to be a part of uORF/oORF/iORF were calculated
- Lists containing functional uORFs/oORFS/iORFs were obtained
- All types of upstream open reading frames tend to save their frame
- Different examples of deeply conserved uORFs/oORFs/iORFs were analyzed.