ASSESSMENT OF TRANSLATION EFFICIENCY FROM RIBOSOME PROFILING AND MRNA-SEQ DATA

I. S. Yevshin[1,2], R. N. Sharipov[1,2], O. A. Volkova[3,*]

[1]*Institute of Computational Technologies, SB RAS, Novosibirsk, Russia*

[2]*BIOSOFT.RU, Ltd, Novosibirsk, Russia*

[3]*Novosibirsk State University, Novosibirsk, Russia*

[4]*The Federal Research Center Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia*,

**ov@bionet.nsc.ru*

***Keywords*:** mammalian mRNAs; translation efficiency; ribosome profiling; Ribo-Seq; mRNA-Seq

Motivation and Aim

Ribosome profiling or a Ribo-Seq technology allows evaluating the distinct mRNA translation level. Over the last years, many studies involving this technology have been performed on different cell types and in various physiological conditions. A wide-scale study of regulation of translation efficiency (TE) in a cell necessitates controlling input Ribo-Seq data and adequate assessment of TE. This paper is designed to assess the quality of raw Ribo-Seq data available in public databases and identify the optimal parameters of their processing and TE assessment.

Methods and Algorithms

Datasets containing both Ribo-Seq and mRNA-Seq data were selected from a RiboSeqDB database developed by the authors. Altogether, 9 datasets consisting of 39 matching mRNA-seq and ribo-seq pairs were found. Raw data were processed using special workflows developed for a BioUML platform. A data pre-processing workflow went on in stages: a) adaptor cutting, b) filtering out rRNA, snRNA and tRNA sequences, c) alignment to transcriptome, d) mappable length estimation and e) read density calculation. To control the quality of TE assessment, the TE values calculated were compared with respective TE values but resulting from mass-spectrometry and mRNA-Seq.

Results

1. ***TE assessment approach using ribo-seq***

After pre-processing, we obtained reads of different length and took into account that the number of mappable transcript positions depends on the read length. We assessed density of

ribo-seq reads as the sum of ratios as given in (1). A similar formula was used to calculate the mRNA-seq density (2).

$$Density_{ribo} = \sum_{L=Lmin}^{Lmax} \frac{rc_{CDS}(L)}{ml_{CDS}(L)}, (1)$$

$$Density_{mRNA} = \sum_{L=Lmin}^{Lmax} \frac{rc_{RNA}(L)}{ml_{RNA}(L)}, (2)$$

where $rc_{CDS}(L)$ - read count of length L mapped to CDS, $rc_{RNA}(L)$ - read count of length L mapped to RNA, $ml_{CDS}(L)$ - number of CDS positions mappable by the reads of length L, $ml_{RNA}(L)$ - number of RNA positions mappable by the reads of length L, Lmax and Lmin – maximal and minimal length of reads after adaptor was cut.

The translation efficiency is computed as a simple ratio of ribosome reads density (ribosome-covered reads) to mRNA-seq reads density, viz.

$$TE = \frac{Density_{ribo}}{Density_{mRNA}}. (3)$$

2. *mRNA-seq uniformity test*

Often mRNA-seq has a highly non-uniform read density profile. This can be explained by the mixture of mRNA isoforms, non-uniform mappability, etc. It is difficult to take into account all these factors in a TE estimation procedure. Currently, we have excluded such complex cases from analysis. Using Kolmogorov-Smirnov test, we checked whether the read density profile differed from the mappability profile. If so, we removed such cases from further consideration.

3. *Out-of-frame reads*

A ribosome recognizes a fixed start codon and moves discretely by 3 nucleotides, but ribo-seq has reads that map out of frame. The out-of-frame reads are those originated from ribosomes that have their A-sites located out of frame. To access the origin of these out-of-frame reads and investigate their influence on TE estimation, we should first identify out-of-frame reads. But the length of a ribosome-protected fragment is not fixed and may vary by several nucleotides due to imperfect RNA digestion. To find the optimal length of a ribosome-protected fragment, we analyzed distribution of 5’ read ends near the known translation start sites for different lengths of ribosome-protected fragments. We concluded that the reads of length 29 represent perfect digestion and the distance between 5’ end of read

and the first nucleotide of ribosome A-site equals 15 nucleotides. For further analysis of out-of-frame reads, we have selected only the reads of length 29 and transformed coordinates of 5' read ends to the coordinates of the first nucleotide located in the A-site of ribosome by adding 15. Knowing the ribosome A-site location, we can easily classify reads into in-frame and out-of-frame ones.

Then we computed TE using the reads from different frames and compared them with the TE resulting from mass spectrometry data. The TE obtained from all frames had the highest correlation (0.63) with the TE resulting from mass spectrometry data as compared to TE obtained from individual frames. The in-frame reads gave higher correlation (0.62) than the out-of-frame reads (0.45 for frame1, 0.47 for frame2). All the frames gave significant correlation. We supposed that emergence of out-of-frame reads can be connected to imperfect nuclease digestion (insufficient or extra) of ribosome-protected fragments during ribosome profiling or it can be explained by catching the ribosome in the intermediate state during transition between codons. Based on these results, we concluded that the reads from all frames should be taken into the TE calculation procedure.

4. ***Reads mapped to the CDS start and end.***

According to (Ingolia) study, cycloheximide treatment distorts ribosome occupancy in the first 15 and the last 5 codons. This begs the question whether or not we should include the reads mapped to the beginning and the end of CDS in the TE calculation procedure. Again, we calculated TE with and without reads from the beginning and the end of CDS and compared it to the TE resulting from mass spectrometry data. The correlation always decreases when we remove reads at the CDS start and end. We concluded that all reads overlapping CDS should be used for TE estimation.

5. ***TE dependency on CDS length***

We study the factors determining gene TE and found that TE depends on the gene length. Longer genes have lower translation efficiency. Further investigation reveals that the TE depends on the length of the CDS but not that of the UTR (untranslated region). We assumed such dependency to be explained by premature ribosome drop-off phenomena.

6. ***Correlation with mass spectrometry data***

Mass spectrometry coupled with mRNA-seq is another method used for TE assessment (Schwannhausser). In our study, we compared the TE values obtained for both riboseq and

mass spectrometry. Though the experiments were carried out in distinct cells and experimental conditions, we observed a relatively high correlation, viz. up to 0.7. We concluded that ribosome profiling can be widely used for TE measurement.

Conclusions

1. The TE assessment approach using ribo-seq data was developed.
2. The optimal parameters for adapter trimming and Ribo-Seq read alignment were adjusted.
3. The influence of out-of-frame Ribo-Seq reads was checked: they should be taken into account like other reads.
4. Reads mapped to the CDS start and end should be taken into account in the same way as reads from the inner part of CDS.
5. TE depends on the CDS length, but not on the UTR length.
6. High correlation of TE estimated from riboseq with mass spectrometry data was observed. Riboseq data can be used for reliable TE estimation.

This work was supported by the Russian Foundation for Basic Research № 14-04-01284.