Estimation of translational importance of mammalian mRNA nucleotide sequence characteristics based on ribosomal profiling data

O.A. Volkova[1,*], Y.V. Kondrakhin[2,3], R.N. Sharipov[2,3,4]

[1]Institute of Cytology and Genetics, SB RAS, Novosibirsk, Russia

[2]Design Technological Institute of Digital Techniques, SB RAS,

Novosibirsk, Russia

[3]Institute of Systems Biology, Ltd, Novosibirsk, Russia

[4]Novosibirsk State University, Novosibirsk, Russia

*ov@bionet.nsc.ru

Keywords: ribosome profiling, RiboSeq, 5'UTR, mRNA, translation initiation

Background

Regulation of eukaryotic genes expression at the translational level mainly occurs at the initiation stage. It is known that 5' untranslated region (5'UTR) of mRNA is involved in the interaction with the translation initiation factors and 40S ribosomal subunits. The 5'UTR mRNA characteristics can influence on the translation initiation efficiency and specificity. Previous knowledge about 5'UTR characteristics were obtained theoretically and *in vitro* for mRNA of individual genes. It did not allow systematic analysis of the mRNA translationally important parameters. For identifying 5'UTR characteristics mentioned above, it is necessary to analyze their relationships with the translational activity of the corresponding mRNAs. Until recent time there were no experimental data on the translation efficiency. Owing to ribosome profiling technology (RiboSeq) genome-wide experimental data of translation efficiency were obtained for many eukaryotic mRNAs. Now it seems to be possible to reveal the translationally important mRNA parameters and predict translation efficiency based on their nucleotide sequences. The aim of this study was to determine the translational significance of individual 5'UTR characteristics in accordance to experimental ribosome profiling data.

Materials and Methods

We extracted RiboSeq data GSE30839 (*Mus musculus*, embryonic stem cells) (Ingolia *et al*., 2011) from the Gene Expression Omnibus database (http://www.ncbi.nlm.nih.gov/geo/). In this work we used the following BioUML (http://www.biouml.org) capabilities: import/export data in different formats; work with tables and samples; access to the sequences and their annotations by DAS protocol (http://www.biodas.org); genome browser for interactive sequences visualization, annotation and work with NGS data; BWA; Bowtie; integration with R / Bioconductor and Galaxy (https://main.g2.bx.psu.edu).

Regression analysis was carried out for revealing relationships between the mouse mRNA nucleotide sequence characteristics and ribosome profiling data. In the frame of regression analysis we have considered the following three different models of regressions: classical least squares regression, random forest regression and support vector machine epsilon-regression.

Results

1. We processed raw RiboSeq data taken from the GSE30839 (Ingolia et al, 2011) using a BioUML workflow (Fig. 1) and compared obtained results with the results from the article. Correlation between the results was ~ 0.98.

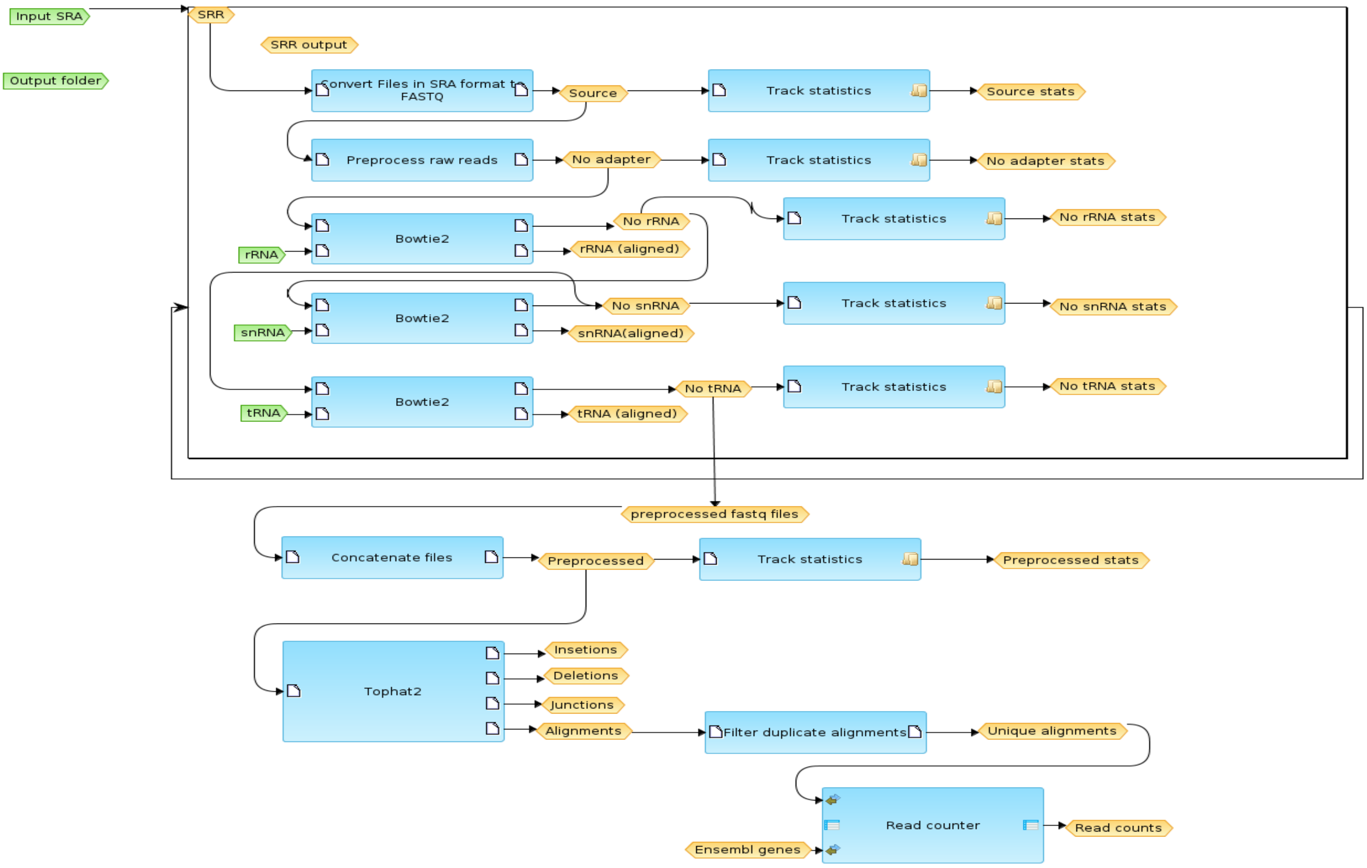

Figure 1. BioUML workflow for processing raw RiboSeq data

2. We carried out regression analysis of translation-relevant mRNA features such as: concentration of AUGs in 5'UTR, A or G on position -3, G on position +4, lg of 5'UTR length, lg of 3'UTR length, lg of transcript length , complementarity index on [-15, 1] positions, complementarity index on [1, 20] positions, C+G content in 5'UTR, C+G content in 3'UTR, C+G content in full transcript and translation efficiency (number of sequenced reads after harringtonin treatment; Ingolia et al, 2011) see Table 1. Correlation between predicted and observed results achieved 0.47, see Fig 2.

Table 1. Coefficients of least squares linear regression for prediction of y = lg(R1), where R1 is number of reads sequenced after harringtonin treatment

mRNA-features	**Regression coefficient**	**Statistic (Z-score)**	**p-value**
intercept constant	4.637	38.809	1.118E-293
AUG concentration in 5'UTR	-19.799	-21.733	4.429E-101
lg(5'UTR length)	-0.347	-16.837	1.870E-62
A or G on position -3	0.163	9.604	5.567E-22
complementarity index on [-15, 1] positions	-0.162	-4.523	3.107E-6
lg(3'UTR length)	-0.098	-4.509	3.321E-6
lg(transcript length)	-0.147	-4.042	2.685E-5
complementarity index on [1, 20] positions	-0.161	-3.882	5.228E-5
(C+G)-content in 5'UTR	-0.255	-3.315	4.605E-4
(C+G)-content in 3'UTR	-0.463	-3.101	9.686E-4
G in position +4	0.026	1.996	0.023
(C+G)-content in full transcript	0.029	0.133	0.447

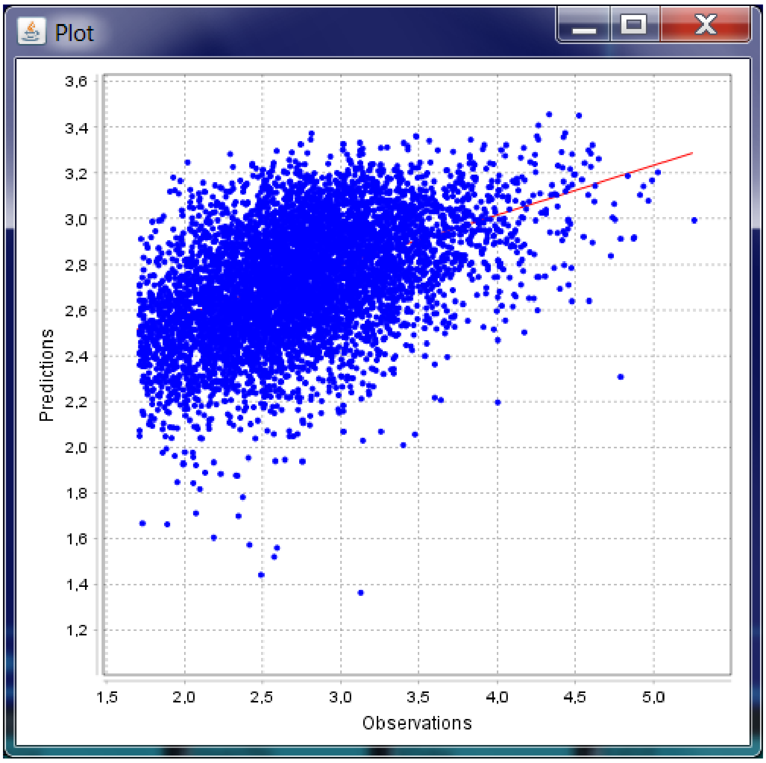

Figure 2. Relationship between the predicted and observed values of lg(R1), where R1 is number of reads sequenced after harringtonin treatment

Some of the mRNA parameters influencing on translation efficiency (see Table 1) were most significant: AUG concentration in 5'UTR, a purine in start codon context position -3, complementarity indices on [-15, 1] and [1, 20] positions, C+G contents in 5'UTR, 3'UTR and in full transcript, lg of 5'UTR and 3'UTR length reduced translation efficiency; whereas A or G in position -3 and G in position +4 increased translation efficiency. It is important to note that most of those mRNA parameters were analyzed in previous studies (Volkova PhD thesis, 2012; Kochetov et al, 2013), excepting complementarity indices on [-15, 1] and [1, 20] positions and C+G content in full transcript, but in our study we achieved essentially higher correlations between the mRNA parameters and translation efficiency (Table 2). Also, as far as we know, this is the first time when genome-wide experimental data (RiboSeq) were used for analysis of the translation-relevant mammalian mRNA features.

Table 2. Cross-validation of three regression models

Regression model	**Accuracy characteristics**	**Training set**	**Test set**
Least squares regression	Explained variance Pearson correlation Spearman correlation	22.2% 0.474 0.478	20.4% 0.455 0.462
Random forest regression	Explained variance Pearson correlation Spearman correlation	85.3% 0.957 0.954	24.9% 0.502 0.487
Support vector machine epsilon-regression	Explained variance Pearson correlation Spearman correlation	39.2% 0.634 0.641	20.0% 0.469 0.462

To assess the prediction abilities of the considered regression models we have performed their cross-validation. For this purpose, the total data set was splitted ‘half-and-half’ into training set and test set at random. Then the following three accuracy characteristics were calculated for each set: explained variance (in %), Pearson and Spearman correlations between predictions and observations, see Table 2. It is not difficult to see that Random forest and Support vector machine regression models seriously over-fitted the training data, therefore the least squares regression is more preferable in order to avoid over-fitting and bias.

This work was supported by the Russian Foundation for Basic Research (grant number 14-04-01284).

Authors are grateful to I.S. Yevshin for his help in this work.