Prediction of start codons and mRNA translation efficiency using RiboSeq-based weight matrices

Yu. V. Kondrakhin[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**:** weight matrices; mRNAs; translation efficiency; ribosome profiling; Ribo-Seq

Motivation and Aim

Ribosome profiling or Ribo-Seq technology allows estimating the level of translation of distinct mRNAs. A wide-scale study of regulation of translation efficiency (TE) in a cell has to be preceded by forming an adequate mRNA sample. A significant limitation of the mRNA analysis is that start codons in the databases, e. g. Ensembl, are mostly annotated with application of a canonical Kozak context. This paper is designed to assess heterogeneity of a start-codon consensus set of mammalian mRNA by RiboSeq-based weight matrices.

Material and methods

Qualitative recognition of functional sites with weight matrices needs: a qualitative site-score calculating method and a qualitative weight matrix or a set of matrices. If the known sites are homogeneous, it is enough to use one matrix. But if the functional sites are heterogeneous, we have to use several matrices.

We calculated site scores with an Individual Probability Score (IPS) method (Kondrakhin et al, 2014). Assuming that ribosome landing sites are heterogeneous, we developed our own matrix-construction method of start sites recognition (implemented in BioUML as a Matrix Derivation Module). The following two algorithms were implemented in Matrix Derivation:

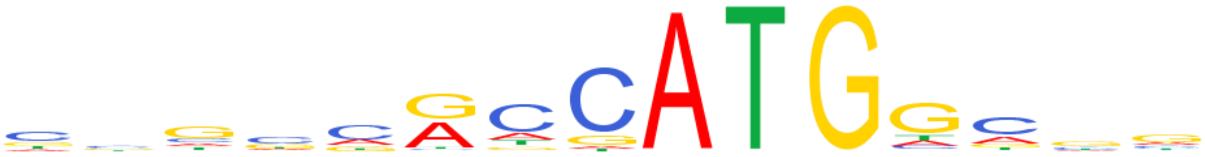
Algorithm 1. A sequential algorithm based on a two-component normal mixture;

Algorithm 2. An EM-like algorithm identifying several matrices at the same time.

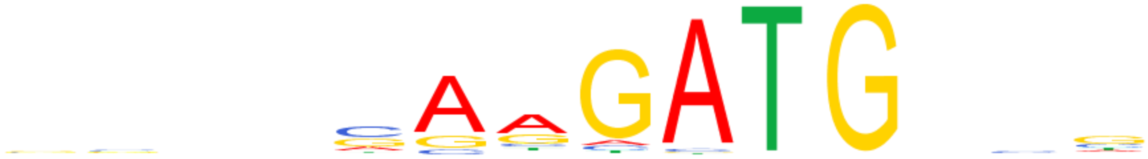
As a training sample of start sites, we used mouse mRNA sequences [-50, +50] around start codons that were annotated by Ensembl, on the one part, and a high score of harringtonin binding (Ingola et al., 2011), on the other part.

Results

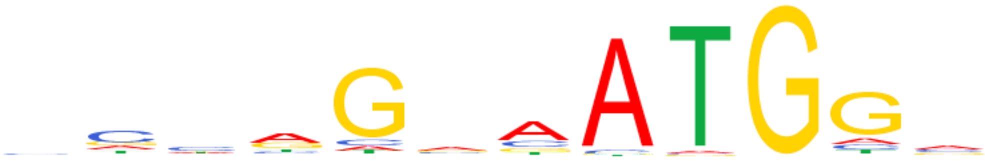
1. We applied algorithm 1 to this sample, and a well-known nnnCCRnnATGGnnn consensus was used as initial approximation. As a result, we obtained the following four matrices in the way independent from each other (Fig. 1).



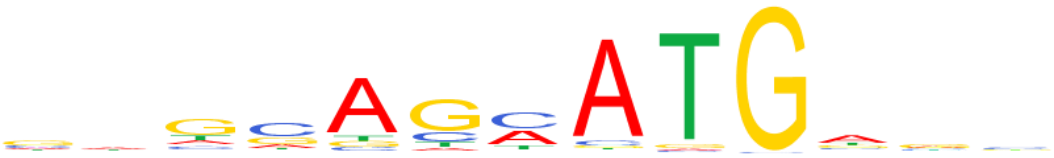
startCodon_1_iteration_15



startCodon_2_iteration_15



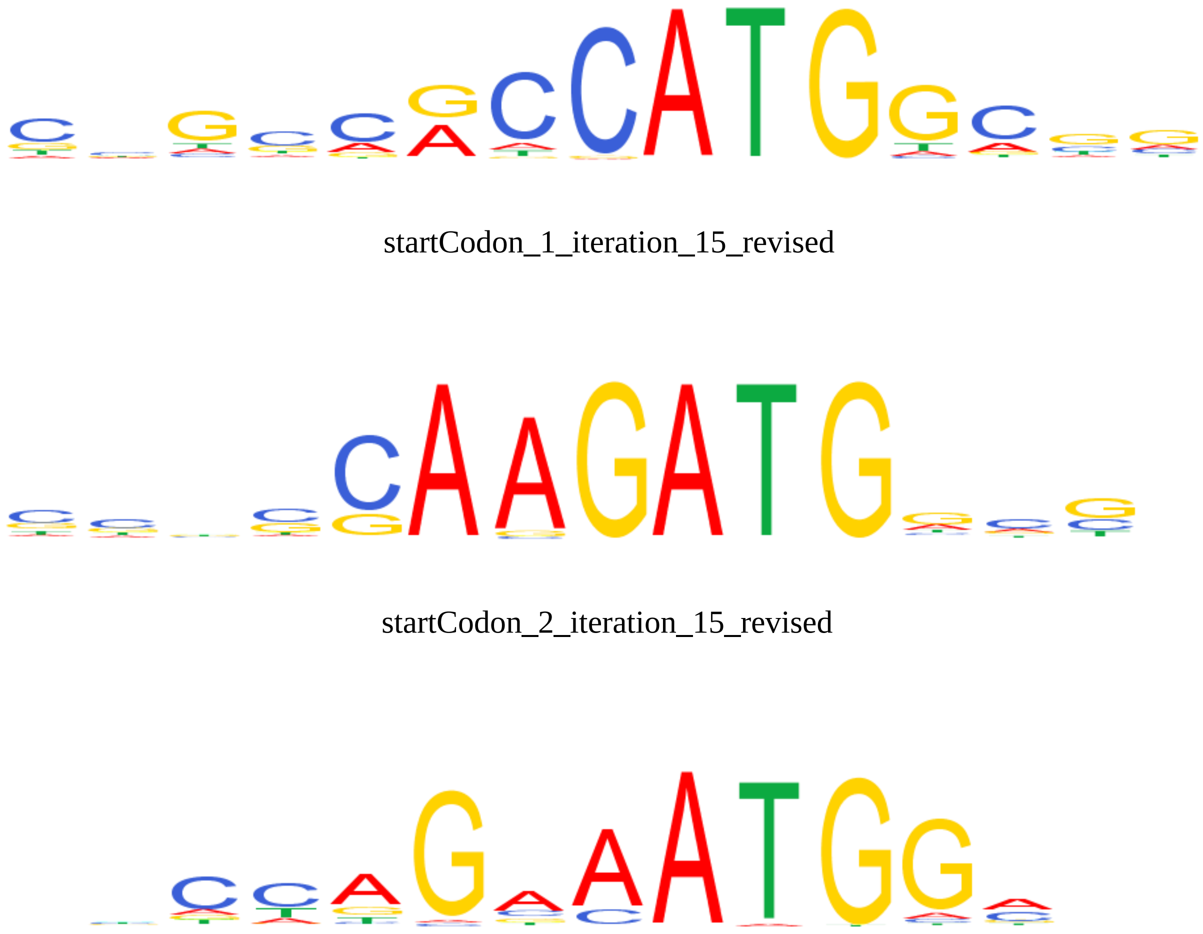
startCodon_3_iteration_15



startCodon_4_iteration_15

Fig. 1. Start codon consensus obtained by using RiboSeq-based weight matrices, algorithm 1

2. Then we used algorithm 2 to obtain more accurate versions of these matrices. As input parameters for the algorithm, we used these four matrices, resulting in three matrices (Fig. 2). The fourth matrix disappeared, i. e. it merged with the first of the three ones left. The newly obtained matrices have become more conservative.



startCodon_1_iteration_15_revised

startCodon_2_iteration_15_revised

startCodon_3_iteration_15_revised

Fig. 2. Start codon consensus obtained by using RiboSeq-based weight matrices, algorithm 2

3. We performed a comparative analysis (Matrix Comparison, BioUML) to evaluate the accuracy of the newly obtained matrices by construction of ROC curves for each individual matrix and for the three matrices together (Fig. 3).

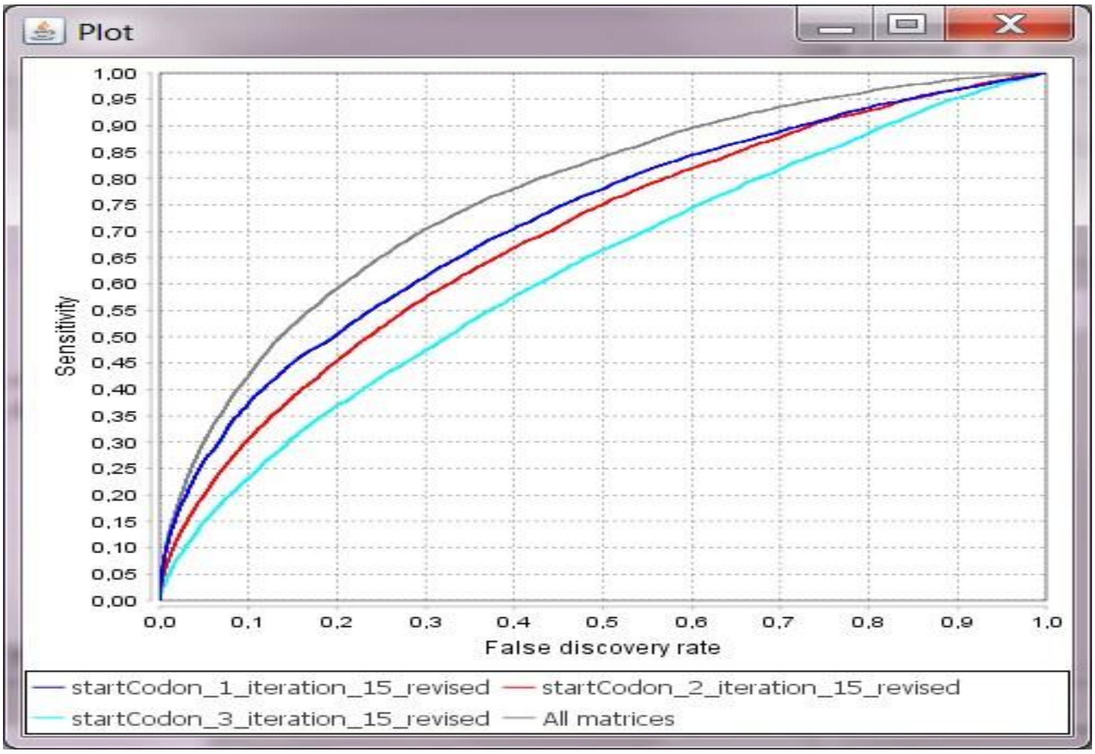


Рис. 3. ROC curves for each individual matrix and for the three matrices together.

It is obvious that the first matrix is better than the second one, and the second one is better than the third one. We can also conclude that simultaneous use of the three matrices is better than individual. In other words, the constructed ROC curves completely confirm the initial assumption of heterogeneity of the full start-codon set.

It is also interesting to note a connection between the third matrix and the secondary structure. It is obvious that this matrix is consistent with the following hairpin: CCA xxxx TGG, where xxx is a loop.

4. We performed some additional analysis of this matrix and compared its scores on two mRNA subsets, viz. those with high and low values of translational efficiency (TE) (Volkova at el, 2016). We found that the matrix scores for the mRNA subset with low TE values were 1.37 times higher than those for the mRNA subset with high TE values. In other words, this matrix is more specific for start codons in mRNA with low TE values.

The matrices obtained allow theoretical prediction of start codons in newly sequenced eukaryotic mRNA with subsequent prediction of TE using the Regression Analysis of BioUML.

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