Bioinformatics, 33(6), 2017, 923–925
doi: 10.1093/bioinformatics/btw736
Advance Access Publication Date: 23 December 2016
Applications Note

Sequence analysis

AltORFev facilitates the prediction of alternative open reading frames in eukaryotic mRNAs

Alex V. Kochetov[1,2,*], Jens Allmer[3], Alexandra I. Klimenko[1], Bulat S. Zuraev[1,2], Yury G. Matushkin[1] and Sergey A. Lashin[1,2]

[1]Institute of Cytology & Genetics, SB RAS, Novosibirsk, Russia, [2]Novosibirsk State University, Novosibirsk, Russia and [3]Molecular Biology and Genetics, Izmir Institute of Technologies, Izmir, Turkey

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on June 8, 2016; revised on October 31, 2016; editorial decision on November 15, 2016; accepted on November 16, 2016

Abstract

Motivation: Protein synthesis is not a straight forward process and one gene locus can produce many isoforms, for example, by starting mRNA translation from alternative start sites. altORF evaluator (altORFev) predicts alternative open reading frames within eukaryotic mRNA translated by a linear scanning mechanism and its modifications (leaky scanning and reinitiation). The program reveals the efficiently translated altORFs recognized by the majority of 40S ribosomal subunits landing on the 5′-end of an mRNA. This information aids to reveal the functions of eukaryotic genes connected to synthesis of either unknown isoforms of annotated proteins or new unrelated polypeptides.

Availability and Implementation: altORFev is available at http://www.bionet.nsc.ru/AUGWeb/and has been developed in Java 1.8 using the BioJava library; and the Vaadin framework to produce the web service.

Contact: ak@bionet.nsc.ru

1 Introduction

The dogma of one gene one product has long been abandoned and it is clear that a gene locus can lead to multiple transcripts which in turn lead to alternative proteins. On the translation level, the usage of alternative open reading frames (altORFs) adds further options for creating various alternative proteins from one genomic locus, which has been shown for eukaryotic transcriptomes using ribo-seq and proteomics techniques (Ingolia, 2016; Mouilleron *et al.*, 2016). Some bioinformatics resources have been proposed that allow the detection of altORFs within mRNAs from large scale data analysis such as Ribotools (Legendre *et al.*, 2015), RiboGalaxy (Michel *et al.*, 2016), Proteoformer (Crappé *et al.*, 2015) and RFPdb (Xie *et al.*, 2016). Knowledge of the full set of polypeptides encoded by a eukaryotic gene is an essential pre-requisite for comprehensive investigation of its functions. However, reproducibility of published ribo-seq data is still poor (Diament and Tuller, 2016) and conventional nucleotide sequence databanks do not contain annotated altORFs. In addition, the individual genetic variants may cause changes in mRNA translation rate and coding potential (Cenik *et al.*, 2015; Schafer *et al.*, 2015). If the nucleotide sequence of an mRNA is not identical to the available ribo-seq-checked reference sequence, the positions of altORF(s) and their relative translation rates may differ. However, due to various parameters influencing recognition and translation efficiency of altORFs their accurate and comprehensive *ab inito* prediction is highly convoluted and no tool performing this task is currently available. Here we present altORFev which fills this gap and performs *ab initio* altORFs prediction based on the linear scanning model and its extensions and also provides estimates of altORFs relative translation efficiency.

2 Description and usage of altORFev

2.1 Objective

altORFev is intended to answer the question which ORFs within a eukaryotic mRNA are likely to be efficiently translated. For this it employs a linear scanning mechanism. The usefulness of this information is based on the following reasons:

1. Currently, linear scanning is considered the default translation mechanism of most eukaryotic mRNAs under normal physiological conditions while alternative translation mechanisms such as internal ribosome entry sites (IRES) may need the presence of special signals (Jackson *et al.*, 2010).
2. Conventional nucleotide sequence databanks (e.g. GenBank and EMBL) do not provide information on altORFs even for reference sequences. These annotation tools were designed to predict only one 'genuine' protein-coding sequence (CDS) per mRNA and, therefore, skip other potentially translated ORFs as 'nonfunctional'. However, ribo-seq and proteomics data clearly demonstrated that this approach needs to be reconsidered (Andrews and Rothnagel, 2014).

2.2 Algorithm

The altORFev program predicts ORFs translated by the majority of ribosomes landing onto eukaryotic mRNA. It is based on the linear scanning model (Jackson *et al.*, 2010; Kozak, 2005) and takes into account the leaky scanning and reinitiation mechanisms. In brief, 40S ribosomal subunits bind to the 5′-end of an mRNA and move linearly in 3′-direction scanning mRNAs for AUG start codon(s). The probability of AUG recognition depends on its nucleotide context. Start codons in the optimal context are recognized by the majority of 40S ribosomal subunits. Thus, if the start codon is located in the optimal context and its ORF is larger than 30 codons, this ORF is defined as 'terminal' since the majority of incoming 40S ribosomal subunits cannot pass through it. If the start codon is located in a suboptimal context, some 40S ribosomal subunits will recognize it and initiate translation, whereas others skip it and may initiate translation downstream (leaky scanning). Finally, if the start codon is located in the optimal context but the ORF size is small (less than 30 codons), the reinitiation process is possible (Kozak, 2005). In this case, some 40S ribosomal subunits, after termination of translation of the small ORF, remain connected to the mRNA and may continue movement in 3′-direction. During scanning they restore their initiation competence by acquiring eIFs and met-tRNAi and may initiate translation further downstream. Since the algorithm is targeted to find ORFs recognized by the majority of ribosomes, some limits on the number of predicted altORFs were applied.

2.3 Restrictions and limitations

altORFev is a sophisticated tool for the *ab initio* prediction of alternative ORFs within eukaryotic mRNAs produced by the majority of scanning ribosomes. However, it is not comprehensive since more experimental information on altORF recognition is needed to increase prediction accuracy. The absolute synthesis rates of altORFs-encoded polypeptides depend on a number of factors (mRNA quantity, translation efficiency and polypeptide stability) and may not be predicted by this program alone. altORFev is limited to finding only AUG initiated altORFs and cannot predict the IRES-governed translation start sites or cases of specific translational control (trans-factors binding sites, enhancers, shunting sites, etc.). Instead, it provides a list of alternative ORFs potentially judged to be efficiently translated from eukaryotic mRNA. This information enables the detection of functions of eukaryotic genes connected to synthesis of either new isoforms of annotated proteins or new unrelated polypeptides.

2.4 Interface

altORFev was implemented in Java 1.8 (standalone version available upon request), using the BioJava library (Prlić *et al.*, 2012) while the web application uses the Vaadin framework.

The program features a basic and an advanced mode. The former uses default parameters and is suitable for prediction of efficiently translated alternative ORFs within mRNAs of vertebrate animals or higher plants under normal conditions. The advanced mode provides the opportunity to change several prediction parameters and is useful in specific cases: (i) if there is a need to change the definition of optimal/moderate/weak AUG contexts (defaults: High = AnnAUGn, GnnAUGG; Moderate = GnnAUGH, YnnAUGG; Weak = YnnAUGH (Volkova and Kochetov, 2010)); (ii) if the mRNA is translated under stress conditions implying eiF2ά phosphorylation (unfolded protein response, virus infection, etc. (Ventoso *et al.*, 2012)). altORFev provides a graphical representation of the predicted altORFs within the mRNA as well as information in text format like sequence, positions, start codon contexts, predicted translation rates and amino acid sequences of the corresponding polypeptides.

3 Conclusion

AltORFev may be used to get additional information on eukaryotic genes taking into consideration alternative coding abilities of their mRNAs. It uses an advanced linear scanning model with leaky scanning and reinitiation modules and thereby extends predictions provided by other tools which are commonly based on ribo-seq data analysis, evolutionary conservation of ORFs, or the basic linear scanning algorithm.

Funding

The work was supported by the Scientific and Technological Research Council of Turkey [grant number: 114Z177] to JA (algorithm development), RAS Program "Biodiversity" to AVK (algorithm development) and RSF [grant number: 142400123] to AIK, BZ, SL and YM (software development).

Conflict of Interest: none declared.

References

Andrews,S.J. and Rothnagel,J.A. (2014) Emerging evidence for functional peptides encoded by short open reading frames. *Nat. Rev. Genet.*, **15**, 193–204.

Cenik,C. *et al.* (2015) Integrative analysis of RNA, translation, and protein levels reveals distinct regulatory variation across humans. *Genome Res.*, **25**, 1610–1621.

Crappé,J. *et al.* (2015) PROTEOFORMER: deep proteome coverage through ribosome profiling and MS integration. *Nucleic Acids Res.*, **43**, e29.

Diament,A. and Tuller,T. (2016) Estimation of tibosome profiling performance and reproducibility at various levels of resolution. *Biol. Direct.*, **11**, 24.

Ingolia,N.T. (2016) Ribosome footprint profiling of translation throughout the genome. *Cell*, **24**, 22–33.

Jackson,R.J. *et al.* (2010) The mechanism of eukaryotic translation initiation and principles of its regulation. *Nat. Rev. Mol. Cell Biol.*, **11**, 113–127.

Kozak,M. (2005) Regulation of translation via mRNA structure in prokaryotes and eukaryotes. *Gene*, **361**, 13–37.

Legendre,R. *et al*. (2015) RiboTools: a Galaxy toolbox for qualitative ribosome profiling analysis. *Bioinformatics*, **31**, 2586–2588.

Michel,A.M. *et al*. (2016) RiboGalaxy: a browser based platform for the alignment, analysis and visualization of ribosome profiling data. *RNA Biol*., **13**, 316–319.

Mouilleron,H. *et al*. (2016) Death of a dogma: eukaryotic mRNAs can code for more than one protein. *Nucleic Acids Res*., **44**, 14–22.

Prlić,A. *et al*. (2012) BioJava: an open-source framework for bioinformatics in 2012. *Bioinformatics*, **28**, 2693–2695.

Schafer,S. *et al*. (2015) Translational regulation shapes the molecular landscape of complex disease phenotypes. *Nat. Commun*., **6**, 7200.

Ventoso,I. *et al*. (2012) Extensive translatome remodeling during ER stress response in mammalian cells. *PLoS One*, **7**, e35915.

Volkova,O.A. and Kochetov,A.V. (2010) Interrelations between the nucleotide context of human start AUG codon, N-end amino acids of the encoded protein and initiation of translation. *J. Biomol. Struct. Dyn*., **27**, 611–618.

Xie,S.Q. *et al*. (2016) RPFdb: a database for genome wide information of translated mRNA generated from ribosome profiling. *Nucleic Acids Res*., **44**, D254–D258.