Ab initio prediction of dual coding regions in mammalian mRNAs

Kseniya Petrova

Department of Biological and Medical Physics, Moscow Institute of Physics and Technology,

9 Institutskiy per., Dolgoprudny, Moscow Region, 141701, Russian Federation

kseniya.petrova@phystech.edu

Ivan Antonov

Institute of Bioengineering, 60 let Oktjabrja pr-t, 7, bld. 1, 117312, Moscow, Russian Federation

ivan.antonov@gatech.edu

It is well known that prokaryotic organisms produce polycistronic mRNAs containing a number of coding sequences (CDSs). Several proteins are translated from such mRNAs. In contrast, vast majority of the mammalian mRNAs contain only one CDS encoding a single protein. Surprisingly, recent studies have demonstrated that more than one region of some mature mammalian mRNAs can be translated. In addition to the annotated CDS region (also known as the "reference open reading frame" or "refORF") these mRNAs contain additional alternative ORFs (altORFs) which are located in the 5' or 3' untranslated regions (UTRs) or even inside the CDS (in this case the altORF is translated in one of the two alternative reading frames)[1]. Investigation of such cases is important because it can significantly expand our knowledge of the mammalian proteomes and reveal new patterns of expression regulation. Moreover, new genes can originate from the emergence of altORFs[2].

Several approaches have been suggested to find altORFs both computationally and experimentally. Experimental ways include direct detection of peptides by mass-spectrometry methods (MS) [9] and indirect detection from the Ribo-Seq data [10]. Such approaches are the most precise, but are limited to the availability of the experimental data.

Known computational approaches exploit unexpected dN/dS relation [11], search for long overlapping ORFs together with "BLASTing" the protein database [12,13] and substitution rate measuring together with altORF length distribution model [14].

Here we focus on identification of the altORFs which are either completely located inside the corresponding refORFs or have a significant overlap with them [3,4,5,6,7]. Interestingly, it has been shown that alternative proteins correlate with intrinsic structural disorder [8]. Overlapping CDSs are of special interest since such regions are under double selection. This restriction influences the codon frequencies and provides an opportunity for an ab initio prediction of the dual coding regions (i.e. using the nucleotide sequences alone).

Our computational approach of ab initio altORFs detection is based on the analysis of the CP (coding potential) and altORFs length distribution. The CP is a measure of probability of a nucleotide sequence to code the peptide in either one particular reading frame or in two reading frames at the same time. The CP is calculated from the pentaplet usage frequencies. Additionally, ORF length distribution gives the probability of an altORF of a given length to occur simply by chance. We are not aware of any other ab initio method that has been developed for prediction dual coding regions.

To test our tool we apply it to 356 human mRNAs containing experimentally validated dual coding regions [15] and 356 randomly selected human mRNAs, for which the expression of the alternative protein has never been detected. Classification accuracy of our model measured by the area under the curve (AUC) is 68.8%. Thus, we demonstrate that ab initio approaches are able to detect dual coding regions. Since our method only requires nucleotide sequences as input it can be applied to a variety of species and the produced predictions can be used to study the evolution of the dual coding genes in mammals.

1. H.Mouilleron et al. (2016) Death of a dogma: eukaryotic mRNAs can code for more than one protein, Nucleic Acids Res. 44(1):14-23

2. C.Rancurel et al. (2009) Overlapping Genes Produce Proteins with Unusual Sequence Properties and Offer Insight into De Novo Protein Creation, Journal of virology, vol. 83, No. 20, 10719–10736

3. M.Hameed et al. (2003) Expression of IGF-I splice variants in young and old human skeletal muscle after high resistance exercise, J Physiol., 547(Pt 1): 247–254

4. H.Yoshida et al. (2001) XBP1 mRNA Is Induced by ATF6 and Spliced by IRE1 in Response to ER Stress to Produce a Highly Active Transcription Factor, Cell, vol. 107, Issue 7, 881–891

5. D.Bergeron et al. (2013) An Out-of-frame Overlapping Reading Frame in the Ataxin-1 Coding Sequence Encodes a Novel Ataxin-1 Interacting Protein, J Biol Chem., 288(30): 21824–21835

6. R.-F.Wang et al. (1998) A Breast and Melanoma-Shared Tumor Antigen: T Cell Responses to Antigenic Peptides Translated from Different Open Reading Frames, J Immunol., 161 (7) 3596-3606

7. N.E.Sharpless, R.A.DePinho (1999) The INK4A/ARF locus and its two gene products, Curr. Opin. Genet. Dev., vol. 9, Issue 1, 22–30

8. E.Kovacs et al. (2010) Dual coding in alternative reading frames correlates with intrinsic protein disorder, PNAS, vol. 107, no. 12, 5429–5434

9. B.Vanderperre et al. (2013) Direct detection of alternative open reading frames translation products in human significantly expands the proteome, PLoS ONE, 8, e70698

10. A.M.Michel et al. (2012) Observation of dually decoded regions of the human genome using ribosome profiling data, Genome Res., 22:2219–2229

11. F.L.Michael (2011) Locating protein-coding sequences under selection for additional, overlapping functions in 29 mammalian genomes, Genome Res., 21: 1916-1928

12. H.Liang, L.F. Landweber (2006) A genome-wide study of dual coding regions in human alternatively spliced genes, Genome Res., 16: 190-196

13. S.Ribrioux et al. (2008) Bioinformatics prediction of overlapping frameshifted translation products in mammalian transcripts, BMC Genomics., 9: 122

14. W.Y.Chung (2007) A first look at ARFome: Dual-coding genes in mammalian genomes, Plos Comput Bio, vol. 3, Issue 5, e91, 855-861

15. B.Vanderperre et al. (2012) HAltORF: a database of predicted out-of-frame alternative open reading frames in human, Database (Oxford), vol. 2012, Article ID bas025