BIOINFORMATICS APPLICATIONS NOTE

Vol. 28 no. 4 2012, pages 591–592
doi:10.1093/bioinformatics/btr706

Sequence analysis

Advance Access publication December 22, 2011

LaTcOm: a web server for visualizing rare codon clusters in coding sequences

Athina Theodosiou, and Vasilis J. Promponas*

Bioinformatics Research Laboratory, Department of Biological Sciences, University of Cyprus, PO Box 20537, CY 1678, Nicosia, Cyprus

Associate Editor: Martin Bishop

ABSTRACT

Summary: We present LaTcOm, a new web tool, which offers several alternative methods for 'rare codon cluster' (RCC) identification from a single and simple graphical user interface. In the current version, three RCC detection schemes are implemented: the recently described %MinMax algorithm and a simplified sliding window approach, along with a novel modification of a linear-time algorithm for the detection of maximally scoring subsequences tailored to the RCC detection problem. Among a number of user tunable parameters, several codon-based scales relevant for RCC detection are available, including tRNA abundance values from *Escherichia coli* and several codon usage tables from a selection of genomes. Furthermore, useful scale transformations may be performed upon user request (e.g. linear, sigmoid). Users may choose to visualize RCC positions within the submitted sequences either with graphical representations or in textual form for further processing.

Availability: LaTcOm is freely available online at the URL http://troodos.biol.ucy.ac.cy/latcom.html.

Contact: vprobon@ucy.ac.cy

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on September 17, 2011; revised on December 13, 2011; accepted on December 16, 2011

1 INTRODUCTION

tRNA abundance is asymmetric (non-uniform) in cells of different organisms and can cause variation in the translation rate for each codon. It has been suggested that rare codons and their clusters are associated with translational pausing (Guisez *et al.*, 1993; Komar and Jaenicke, 1995). The translational rate can be maximized at mRNA regions with codons read by high cellular levels of cognate tRNA species and minimized at sites corresponding to rare tRNAs (RCCs) (Lavner and Kotlar, 2005). A critical issue is that the concentration of isoaccepting tRNAs for a set of synonymous codons varies among organisms, tissues and stages of differentiation (Ikemura, 1985). Moreover, different organisms show specific preferences for codons encoding the same amino acid (codon bias), reflected in the frequency of occurrence of synonymous codons in genomic DNA. In *Escherichia coli*, it has been shown that the non-random choice of codons is mostly attributable to the availability of transfer RNA within a cell (Ikemura, 1981); however, more recent work by Dong *et al.* (1996) has shown that the correlation of codon usage and tRNA level is not perfect.

In principle, the RCC detection process is the identification of codon clusters corresponding to rare tRNA species along mRNAs, as quantified using scales of experimental cellular tRNA levels. A complete dataset of this type is only available for *E.coli* (Dong *et al.*, 1996); therefore, approaches based on codon usage scales are alternatively used. Recently, two different methods have been developed for identifying RCCs in coding sequences: the %MinMax algorithm (Clarke and Clark (2008); http://www.codons.org) and RiboTempo [(Zhang *et al.* (2009); Zhang and Ignatova (2009); http://hxapp.hexun.com/RiboTempo/Default.aspx)].

%MinMax utilizes codon usage scales, and RiboTempo relies on scales based on tRNA-abundance data to quantify translational elongation rates (see Supplementary Material for further details). Both web servers offer different features, e.g. scale options and output formats, and they use a simple sliding window approach, with *fixed* window size w equal to 18 and 19, respectively. These w values correspond to the optimal window size (18) proposed for the problem of identifying translational pause sites by Makhoul and Trifonov (2002) (for a discussion see Supplementary Material). An obvious limitation of window-based methods is the detection of RCCs with length at least equal to the applied window size.

We present LaTcOm, a novel flexible web server, aiming to address shortcomings of the existing methods and to also provide new tools and features for RCC detection. LaTcOm offers users the option to use the %MinMax and RiboTempo methods, along with a new window-less RCC detection approach, based on the linear time Maximal Scoring Subsequences algorithm [MSS; (Ruzzo and Tompa, 1999)]. In addition, both tRNA-abundance and codon usage scales are supported, including the option for users to enter their own scales, and a selection of novel transformations that may prove useful for RCC analyses. Moreover, the ability to choose different values of w for window-based RCC detection schemes, the explicit report of RCC coordinates and simulation-based P-values as a measure for statistical RCC validation, enable more sophisticated analyses of RCCs.

2 METHODS

Implementation: we developed in Perl a generic sliding window algorithm, for implementing the %MinMax and RiboTempo methods. Contrary to the initial implementations, the sliding window procedure offered by LaTcOm is purposedly of decreasing size when approaching sequence extremeties for enabling computing biologically relevant values near the 5′ and 3′ termini.

In addition, LaTcOm offers access to a novel RCC identification scheme, based on tailoring the Maximal Scoring Subsequences algorithm [MSS;

*To whom correspondence should be addressed.

Ruzzo and Tompa (1999)]. Briefly, MSS takes as input an array of numerical values and is guaranteed to find in linear time all maximal segments with an aggregate score exceeding some predefined threshold.

Scales for RCC identification integrated in LaTcOm include (see also Supplementary Material and Supplementary Figure S1 for details):

- *Escherichia coli* tRNA abundance values, based on data from Dong *et al.* (1996), as calculated in Zhang *et al.* (2009), and an in-house calculated variant;
- (weighted) codon usage from a subset of highly expressed *E.coli* genes, taken from Table 4 of Dong *et al.* (1996);
- codon usage tables available from http://www.kazusa.or.jp/codon/ (Nakamura *et al.*, 2000); and
- user defined scales in GCG format.

Tunable parameters include the window size or the least RCC length and the genetic code. A number of scale transformations is available (see below) and their utility is discussed in Supplementary Material and Supplementary Figure S2:

- 'Linear shift' ($x \rightarrow \hat{x} - x$), where the scale-average ($\hat{x}$) is subtracted from each given scale value x, followed by reversing the sign.
- 'Multiplicative inverse' ($x \rightarrow \frac{1}{x}$), where each scale value x is substituted by $\frac{1}{x}$.
- 'Sigmoid' applied to a linearly shifted scale ($x \rightarrow \frac{2}{1+e^{-(\hat{x}-x)}} - 1$).
- A combination of the multiplicative inverse and linear transformations ($x \rightarrow \frac{\hat{1}}{x} - \frac{1}{x}$).

Workflow: users may enter or upload FASTA formatted coding DNA/mRNA sequences, and choose the desired RCC identification algorithm, parameters and output type (Supplementary Figure S3). LaTcOm validates user input (see Supplementary Material) and transforms the sequence into a linear array of tRNA abundance or codon usage values based on the selected scale (and possibly transformation). This numerical array is fed as input to the selected RCC identification algorithm. Results are presented as PNG images (graphical output) or as the array of values calculated by each algorithm (text output), and returned to the user within an HTML page, along with the set of chosen parameters for reference (Supplementary Figure S4). Text output enables users to visualize the RCC detection results using third-party tools, and displays the analytically computed expected score value and a simulation-based *P*-value for each cluster (see Supplementary Material). All output can be visualized online or downloaded as a compressed archive.

3 DISCUSSION

Clusters of unusual codon composition have been the focus of several recent research efforts, especially as indicators of an extra level of cell regulation (Tuller *et al.*, 2011), as well as possible mediators of correct protein folding (Saunders and Deane, 2010). We developed LaTcOm, which is to the best of our knowledge the first flexible web application offering alternative methods for detecting RCCs, and we introduce a new window-less RCC identification algorithm. It is worth mentioning that when the LaTcOm web server was being developed, another window-less RCC-detection approach, based on the spatial scan method [introduced by Huang *et al.* (2007)] was published (Ponnala, 2010). A detailed comparison of the features offered by different RCC-detection algorithms is available as Supplementary Material (Supplementary Table S1).

We anticipate that the availability of a versatile online tool for RCC identification will enable a number of analyses to be performed. For example, when optimizing coding sequences for heterologous gene expression experiments, LaTcOm results could be indicative of codon choices that may interfere with proper folding of the polypeptide chain. More specifically, RCCs according to the host organism's tRNA abundance/codon usage may have to be preserved for expressing functional proteins. In addition, LaTcOm may be used to study patterns of translational rate within diverged protein families, or the correlation of translational rate with protein structural and functional features, such as protein disorder, aggregation and co-translational folding. Such applications may trigger extensions of the current methods, as for example for the analysis of multiple sequence alignments (Widmann *et al.*, 2008) and the study of the mechanics of translation (Tuller *et al.*, 2011).

ACKNOWLEDGEMENT

The authors wish to thank the anonymous referees for invaluable comments on the manuscript and the LaTcOm functionality. We also thank Professor Walter Ruzzo (University of Washington) and Shane Neph (University of Washington) for help with the MSS source code. We also thank Professor Zoya Ignatova (University of Potsdam) and Dr Gong Zhang (University of Potsdam) for providing their tRNA scale, Professor Konstantinos Fokianos (University of Cyprus) for useful discussions on the RCC validation procedure and Ioanna Kalvari (University of Cyprus) for helping with interfacing MSS with the LaTcOm modules.

Funding: A.G. Leventis Foundation (PhD Scholarship to A.T.); University of Cyprus.

Conflict of interest: none declared.

REFERENCES

Clarke,T.F. IV and Clark,P.L. (2008) Rare codons cluster. *PLoS One*, **3**, e3412.

Dong,H. *et al*. (1996) Co-variation of tRNA abundance and codon usage in *Escherichia coli* at different growth rates. *J. Mol. Biol.*, **260**, 649–663.

Guisez,Y. *et al*. (1993) Folding of the MS2 coat protein in *Escherichia coli* is modulated by translational pauses resulting from mRNA secondary structure and codon usage: a hypothesis. *J. Theor. Biol.*, **162**, 243–252.

Huang,L. *et al*. (2007). A spatial scan statistic for survival data. *Biometrics*, **63**, 109–118.

Ikemura,T. (1981) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes: a proposal for a synonymous codon choice that is optimal for the *E. coli* translational system. *J. Mol. Biol.*, **151**, 389–409.

Ikemura,T. (1985) Codon usage and tRNA content in unicellular and multicellular organisms. *Mol. Biol. Evol.*, **2**, 13–34.

Komar,A.A. and Jaenicke,R. (1995) Kinetics of translation of gamma B crystallin and its circularly permutated variant in an in vitro cell-free system: possible relations to codon distribution and protein folding. *FEBS Lett.*, **376**, 195–198.

Lavner,Y. and Kotlar,D. (2005) Codon bias as a factor in regulating expression via translation rate in the human genome. *Gene*, **345**, 127–138.

Makhoul,C.H. and Trifonov,E.N. (2002) Distribution of rare triplets along mRNA and their relation to protein folding. *J. Biomol. Struct. Dyn.*, **20**, 413–420.

Nakamura,Y. *et al*. (2000) Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic Acids Res.*, **28**, 292.

Ponnala,L. (2010) Detecting slow-translating regions in *E. coli. Int. J. Bioinform. Res. Appl.*, **6**, 522–530.

Ruzzo,W.L. and Tompa,M. (1999) A linear time algorithm for finding all maximal scoring subsequences. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 234–241.

Saunders,R. and Deane,C.M. (2010) Synonymous codon usage influences the local protein structure observed. *Nucleic Acids Res*, **38**, 6719–6728.

Tuller,T. *et al*. (2011) Composite effects of gene determinants on the translation speed and density of ribosomes. *Genome Biol.*, **12**, R110.

Widmann,M. *et al*. (2008) Analysis of the distribution of functionally relevant rare codons. *BMC Genomics*, **9**, 207.

Zhang,G. and Ignatova,Z. (2009) Generic algorithm to predict the speed of translational elongation: implications for protein biogenesis. *PLoS One*, **4**, e5036.

Zhang,G. *et al*. (2009) Transient ribosomal attenuation coordinates protein synthesis and co-translational folding. *Nat. Struct. Mol. Biol.*, **16**, 274–280.