Bioinformation
www.bioinformation.net

open access
Hypothesis

On finding poorly translated codons based on their usage frequency

Lalit Ponnala

Computational Biology Service Unit, Cornell University, Ithaca NY 14853, USA; Lalit Ponnala – E-mail: lp257@cornell.edu; *Corresponding author

Received June 16, 2009; Accepted July 18, 2009; Published September 05, 2009

Abstract:
Long stretches of “rare” codons are known to severely inhibit the efficiency of translation. Understanding the distribution of such rare codons is of critical importance in improving the efficiency of heterologous gene expression systems. Accurate estimates of codon usage take the abundance of each protein into consideration. In this paper, we analyze the correlation between approximate measures of codon usage and the availability of tRNA at various growth rates in *E. coli*. We show that the computationally derived estimates of tRNA isoacceptor concentration enable the finding of poorly translated codons.

Keywords: codons; codon usage; tRNA; frequency; *E. coli*

Background:

The translation of codons to amino acids takes place via charged tRNA molecules that can recognize and bind to each codon while carrying the corresponding amino acid **[1]**. Owing to wobble pairing between codon and anticodon, there are many tRNA isoacceptors that recognize more than one codon **[2-5]**. The tRNA-codon recognition pattern is known to vary across eubacterial species and even across different strains of the same species (note that Table2 in **[6]**, which is based on E.coli K-12 strain W3110, differs from Table2 in **[7]** which pertains to strain W1485 of E.coli K-12). It is cumbersome to experimentally identify these codon recognition “rules” and to measure the concentration of each individual isoacceptor accurately **[8]**. The presence of many modified versions of tRNA isoacceptors further complicates matters **[4, 9]**.

The rate at which the ribosome translates codons in a eubacterial mRNA sequence is shown to depend on the availability or abundance of the corresponding tRNA isoacceptors **[10]**. Genes that are highly expressed generally tend to have codons recognized by relatively abundant tRNAs **[7]**. The frequency of codon usage has been shown to be roughly proportional to the tRNA isoacceptor concentration **[3, 6]**, leading, alongwith other evidence, to the *Translational Efficiency Hypothesis*, which states that natural selection favors codons that increase the rate of peptide elongation **[11]**. It has also been proposed that tRNA abundance and codon usage can co-evolve to favorable states **[2, 12]**. Extensive studies have been conducted to understand the distribution of poorly-translated codons, i.e. codons whose tRNA isoacceptors are in low concentration in the cell, and to investigate if they serve any functional roles **[13]**. Such investigations have led to quite a few interesting findings, such as the effect of rare codons on modulating gene expression **[14]** and in attenuating viruses **[15]**. Most recently, it has been shown that the presence of rare codons serves to pause the ribosome and gives protein domains time to fold **[16]**.

In heterologous gene expression systems, it has been shown that eliminating such pausing signals leads to high level expression of protein **[17, 18]**. Substituting rare codons with synonymous codons that are recognized by abundant tRNA isoacceptors reduces the total time of translating an mRNA sequence, thereby increasing efficiency **[19]**. In order to identify locations or regions of such poorly-translated codons using statistical methods, a clear way of quantifying the “rareness” of a codon is needed. In order to make such a quantification method widely applicable to any species/strain, it should not rely on the exact pattern of isoacceptor-codon recognition and measurements of isocacceptor concentration. The frequency of use of codons is known to correlate reasonably well with the concentration of tRNA isoacceptors, so it seems logical to rely on codon frequency to help identify codons that most-inhibit translation efficiency. It also seems worthy to investigate how the estimates of codon frequency vary across different gene categories. Our objective in this paper is to compare the sensitivity of computational (sequence-based) approaches to finding poorly translated codons. In particular, we use a crude measure of codon frequency based only on the gene sequence without taking the corresponding protein abundance into consideration. We will also investigate the effect of cellular growth rate on our proposed methods.

Methodology

We use experimental data on tRNA isoacceptor concentration measured at different growth rates, as published in **[7]**. These measurements have been made using the Ecoli K-12 strain W1485, for which, unfortunately, the whole genome sequence is not available. But, W1485 is very closely related taxonomically to the well-documented Ecoli K-12 strain MG1655 **[20]**, and we assume that codon frequencies calculated using gene-sequences from the MG1655 strain can be meaningfully used as close substitutes for the W1485 strain.

Calculation of codon frequency

We downloaded the complete genome sequence of Escherichia coli K-12 strain MG1655 from Genbank **[21]**, and extracted sequences for the following sets of genes: (1) all annotated genes from Genbank; (2) verified genes listed in the EcoGene database **[22];** (3) genes classified into functional categories based on their cellular role **[23]**

We calculated codon frequencies using each of the above gene sets separately. Ignoring genes that contain a frameshift, we calculate the total number of times each codon occurs in the selected set of genes, and divide it by the total number of codons in the gene-set to get its usage frequency. This is not an accurate way of estimating codon frequency, since we ignore the abundance of each protein. Accurate measures of codon usage that take protein levels into consideration are presented in **[7]**. Our method, on the other hand, relies only on the gene sequences and is very straightforward and simple.

Calculation of tRNA availability

Using the experimental measurements of tRNA isoacceptor concentration (see Table5 of **[7]**) and the pattern of tRNA-codon recognition (see Table2 of **[7]**), the tRNA “availability” for each codon is calculated as described in **[16]**: (1) if the codon has only one tRNA isoacceptor and vice versa, the concentration of that tRNA isoacceptor is assigned to the codon; (2) for an isoacceptor that recognizes more than one codon, its concentration is distributed among those codons in the ratio of their usage frequency

Discussion:

We calculated the Pearson correlation coefficient between the estimated tRNA availability and codon frequency for each set of selected genes **(Table 1 in supplementary material)**. We see that the correlation is quite strong (approximately 0.8), regardless of which gene set is used. We also find a similar correlation across growth-rates, consistent with what was previously thought to be the case **[6]**. We then compared the codons having lowest usage frequency with those having lowest tRNA concentration by creating two lists: one list (C) containing 10 of the codons having lowest usage frequency and another list (T) containing 10 codons of lowest tRNA availability. Note that the list T needs to be re-calculated for each growth rate, since the concentration of tRNA isoacceptors varies across growth rates (see Table5 of **[7]**). For each set of genes, at each growth rate, we count the number of codons that are present in both lists, i.e. how many codons from T are present in C, as a measure of how accurate our codon usage frequencies are in identifying slowly-translated codons. We found that about 5-6 of the codons from T are are usually present in C, regardless of which gene set is used to prepare the two lists **(Table 1 in supplementary material)**. This indicates a fairly good accuracy in finding poorly translated codons based on our simple measure of codon usage frequency.

Based on their own calculations as well as some experimental evidence from others, Zhang *et al* **[16]** have prepared a set of 10 codons that they consider to have slowest rate of translation. This list (K) contains the following codons: (CUA, UCC, UCA, CCU, CCC, CCA, ACA, AGG, UUA, GUC). On an average, we found that about 5-6 of the codons from C are present in K (data not shown), indicating that our measure of codon frequency alone is not a very good estimate of translation rate. Most of the previous studies evaluating the correlation between codon frequency and tRNA concentration use old data **[6]** and have been published almost three decades ago. It is worthwhile to re-evaluate these results using more recent data (as published in **[7]**). Kanaya *et al* **[24]** attempt to do almost the same things that we do in this paper, but they use an unconventional way of measuring codon frequency. Ikemura **[6]** addresses similar issues, but our focus in this paper is not on the strength of correlation between tRNA isoacceptor concentration and codon frequency as a whole, rather our aim is to find codons having poorest translation rate.

We have begun to analyze the distribution of poorly-translated codons in relation to ribosome-mRNA hybridization. Recent papers **[16, 25]** use sliding-window methods for finding clusters of rare codons, but such methods are not statistically rigorous. In order to refine our models to capture the coding regions that are genuinely “problematic”, we need a set of codons that are known to have low tRNA isoacceptor concentration under the desired experimental conditions for the specified species strain. This information is usually hard to attain when the exact number of tRNA isoacceptors and their codon-recognition rules are not completely known. True measures of codon frequency need to consider protein abundances, which are not easily accessible for every species and strain. What we have shown in this paper is that quick-and-easy ways to estimate translation rate (based only on simple codon usage frequency in a closely-related strain, without relying on exact codon-tRNA mapping) do a reasonably good job of identifying poorly-translated codons, irrespective of cellular growth conditions. To the best of our knowledge, this is the first time that codon frequencies calculated without considering protein abundance have been compared against experimentally measured tRNA concentrations. We hope the results in this paper clarify some of the issues relating tRNA availability and codon usage.

References:

[1] T Ikemura, *Mol. Biol. Evol.* **2**:13 (1985) [PMID: 3916708].
[2] PG Higgs *et al., Mol. Biol. Evol.* **25**:2279 (2008) [PMID: 18687657].
[3] J Elf *et al., Science* **300**:1718 (2003) [PMID: 12805541].
[4] H Grosjean *et al., Biochimie.* **77**:139-41 (1995) [PMID: 7541252].
[5] F Jhling *et al., Nucleic. Acids. Res.* **37**:D159 (2009) [PMID: 18957446].
[6] T Ikemura, *J. Mol. Biol.* **146**:1 (1981) [PMID: 6167728].
[7] H Dong *et al., J. Mol. Biol.* **260**:649 (1996) [PMID: 8709146].
[8] Buvoli *et al., RNA* **6**:912 (2000) [PMID: 10864048].
[9] M Withers *et al., RNA* **12**:933 (2006) [PMID: 16618964].
[10] S Varenne *et al., J. Mol. Biol.* **180**:549 (1984) [PMID: 6084718].
[11] X Xia, *Genetics* **149**:37 (1998) [PMID: 9584084].
[12] T Ikemura, *J. Mol. Biol.* **151**:389 (1981) [PMID: 6175758].
[13] M Marin, *Biotechnol. J.* **3**:1047 (2008) [PMID: 18624343].
[14] JF Kane, *Curr. Opin. Biotechnol.* **6**:494 (1995) [PMID: 7579660].
[15] JR Coleman *et al., Science* **320**:1784 (2008) [PMID: 18583614].
[16] G Zhang *et al., Nat. Struct. Mol. Biol.* **16**:274 (2009) [PMID: 19198590].
[17] S Jana, JK Deb, *Appl. Microbiol. Biotechnol.* **67**:289 (2005) [PMID: 15635462].
[18] GW Hatfield, DA Roth, *Biotechnol. Annu. Rev.* **13**:27 (2007) [PMID: 17875472].
[19] M Widmann *et al., BMC Genomics* **9**:207 (2008) [PMID: 18457591].
[20] K Hayashi *et al., Mol. Syst. Biol.* **2**:7 (2006) [PMID: 16738553].
[21] http://www.ncbi.nlm.nih.gov/nuccore/NC 000913
[22] KE Rudd, *Nucleic. Acids. Res.* **28**:60 (2000) [PMID: 10592181].
[23] MH Serres *et al., Nucleic. Acids. Res.* **32**:D300 (2004) [PMID: 14681418].
[24] S Kanaya *et al., Gene* **238**:143 (1999) [PMID: 10570992].
[25] TF Clarke IV, PL Clark, *PLoS ONE* **3**:e3412 (2008) [PMID: 18923675].

Edited by P. Kangueane
Citation: Ponnala, Bioinformation 4(2): 63-65 (2009)

Supplementary Material:

Table 1: Correlation between codon frequency and tRNA concentration calculated using selected gene-sets at various growth rates, in brackets is shown the number of codons from T that are present in C

	Growth rate				
Gene set	0.4	0.7	1.07	1.6	2.5
All genes	0.780(5)	0.788(5)	0.800(6)	0.787(5)	0.812(5)
Verified genes	0.772(6)	0.778(6)	0.786(6)	0.779(6)	0.805(6)
Enzyme	0.783(7)	0.787(6)	0.797(6)	0.789(6)	0.813(6)
Metabolism	0.786(7)	0.794(5)	0.806(6)	0.792(6)	0.816(6)
Cell structure	0.773(5)	0.787(4)	0.804(5)	0.777(4)	0.793(4)
Location of gene products	0.795(6)	0.804(5)	0.817(6)	0.801(5)	0.821(5)
Information transfer	0.792(6)	0.795(5)	0.803(6)	0.801(5)	0.823(6)
Regulation	0.776(5)	0.781(5)	0.789(5)	0.788(4)	0.812(4)
Transport	0.778(5)	0.793(4)	0.811(5)	0.782(4)	0.796(4)