© Indian Academy of Sciences

RESEARCH COMMENTARY

Cotranslational protein folding reveals the selective use of synonymous codons along the coding sequence of a low expression gene

SUVENDRA KUMAR RAY[1]*, VISHWA JYOTI BARUAH[1], SIDDHARTHA SANKAR SATAPATHY[2] and RAJAT BANERJEE[3]

[1]*Department of Molecular Biology and Biotechnology, and* [2]*Department of Computer Science and Engineering, Tezpur University, Napaam, Tezpur 784 028, India*
[3]*Department of Biotechnology and Dr. B. C. Guha Centre for Genetic Engineering and Biotechnology, Ballygunge Science College, University of Calcutta, 35 Ballygunge Circular Road, Kolkata 700 019, India*

[Ray S. K., Baruah V. J., Satapathy S. S. and Banerjee R. 2014 Cotranslational protein folding reveals the selective use of synonymous codons along the coding sequence of a low expression gene. *J. Genet.* **93**, 613–617]

Due to the degeneracy in the genetic code, many amino acids are encoded by more than one codon, called synonymous codons. The usage of synonymous codons in a coding sequence is not random, a phenomenon known as codon usage bias (CUB) that occurs in all genomes. It is believed that synonymous codons are not equivalent with respect to their coding efficiency during translation. This phenomenon has been demonstrated by the difference between the high expression genes (HEG) and low expression genes (LEG) within a genome with respect to the compositional abundance values of different synonymous codons (Ermolaeva 2001; Hershberg and Petrov 2008; Plotkin and Kudla 2011). Optimal codons, whose frequency is higher in HEGs than in LEGs or whole genome, are believed to be translated more rapidly and accurately than the other synonymous codons or nonoptimal codons (Sharp *et al.* 2005; Ran and Higgs 2010; Satapathy *et al.* 2014). As greater numbers of proteins are synthesized by the HEGs (Ghaemmaghami *et al.* 2003; dos Reis *et al.* 2003; Ishihama *et al.* 2008; Hiraoka *et al.* 2009), selection pressure on the coding sequence of these genes is believed to be higher for efficient translation to occur. Therefore, synonymous codons that are more efficient in translation are found in higher frequency in these genes in comparison to the LEGs (Satapathy *et al.* 2012). Thus, the translational selection pressure is considered to be the major determining factor for CUB in HEGs. In contrast, translational selection pressure is believed to be weaker in LEGs as a fewer number of protein molecules are synthesized by these genes. Therefore, CUB in LEGs are predominantly determined by mutation pressures such as genome G+C composition (Muto and Osawa 1987; Palidwor *et al.* 2010) and strand compositional bias in DNA (Lobry and Sueoka 2002; Powdel *et al.* 2010; Paul *et al.* 2013).

In a genome, why are HEGs not composed solely of optimal codons and similarly, why are not LEGs composed of either nonoptimal codons or codons favoured by mutational bias? The selection mutation drift (SMD) theory proposed by Bulmer (1991) suggests that the use of synonymous codons in a gene is a combined result of selection, mutation and drift. According to this theory, CUB in HEGs is determined by selection pressure whereas the same in the LEGs is determined by mutation pressure. The SMD theory supports the theory of unidirectional selection on codons given by Sharp and Li (1986), from their analysis of CUB in *Escherichia coli* genes. The theory of unidirectional selection suggests that there can be selection on CUB only for the HEGs but not for the LEGs. However, in the early 1980s, selectionists had proposed the expression regulation (ER) theory to explain CUB in organisms (Gouy and Gautier 1982; Bulmer 1991). According to the ER theory, translational selection is operative both in HEGs and LEGs. Similar to HEGs, where optimal codons are favoured for greater abundance of the encoded protein, in LEGs nonoptimal codons are also favoured for lower abundance of the encoded protein. The observations of rare codons in different LEGs in *E. coli* were described in support of the ER theory (Konigsberg and Godson 1983). In summary, while SMD theory favours both the role of selection theory of evolution and neutral theory of evolution, the ER theory favours only the selection theory of evolution.

If selection is equally operative on HEGs and LEGs then mutation should also be equally effective on both the classes of genes. Study of mutation rates by the analysis of orthologous gene sequences in organisms suggested that mutation

*For correspondence. E-mail: suven@tezu.ernet.in.

Keywords. codon degeneracy; codon usage bias; gene expression; selection–mutation drift theory; cotranslational protein folding.

rate was higher in LEGs than HEGs, which was in favour of the SMD theory rather than the ER theory (Sharp and Li 1987). In fact, the earlier finding has recently been further supported with a publication where the genome sequences of 34 *E. coli* strains were compared for studying mutation rate in different genes (Martincorena *et al.* 2012). But there was no experimental demonstration to prove the ER theory is wrong. The ideal approach to prove ER theory was to replace the nonoptimal codons by the optimal synonymous codons (efficient in translation) in a LEG followed by studying its function in the organism. If the protein synthesized from the modified coding sequence will still be equally functional then it will support no selection for the nonoptimal codons. Interestingly, a recent discovery published in *Nature* (Zhou *et al.* 2013) has indeed proved the selection on nonoptimal codons along the coding sequence of an LEG.

Neurospora crassa is a filamentous fungus which is known to exhibit strong CUB in the HEG. The FREQUENCY (FRQ) is a low expression protein with little CUB in its coding sequence and is important for circadian oscillator in *N. crassa*. Unlike FRQ, the FRH, the other protein is also important for circadian oscillator in this fungus, is a high expression protein with strong CUB in its coding sequence. The FRQ–FRH complex is important for circadian oscillation regulation. Recently, Zhou *et al.* (2013) studied the role of nonoptimal codons in *frq* gene. To elucidate whether nonoptimal codon in the low expression FRQ is under any selection, the *frq* gene was expressed with optimal codons instead of nonoptimal codons. However, they observed that the FRQ protein failed to complement the *frq* mutant. The FRQ protein abundance value was higher which suggested that the optimal codon was in fact getting translated efficiently as was expected and the changed mRNA sequence was not affecting its stability in *N. crassa*. The inability of FRQ to complement was surprising. It was experimentally demonstrated that this was due to the changed protein structure with identical amino acid sequence. This experiment by Zhou *et al.* (2013) was somewhat similar to the earlier work done by Kimchi-Sarfaty *et al.* (2007), and the conclusion drawn regarding the affect of cotranslational protein folding on protein structure was similar. But the recent work carried out by replacing the nonoptimal codons with optimal codons in a low expression protein and the finding has indeed challenged the more than two decade long notion that codon usage in LEGs are under low selection. The experiments of Zhou *et al.* (2013) proved that in the *frq* gene the nonoptimal codons are actually under Darwinian selection and the selection force is the cotranslational folding.

Several studies in the late 1990s revealed that the protein folding is a cotranslational process (Komar 2009) (figure 1). Similar discovery by Zhou *et al.* (2013) is immensely significant both at fundamental as well as in applied research of protein folding study. The cotranslational protein folding event provided a clue to scientists that if two coding sequences differ only at their synonymous sites, even though they generate two identical polypeptide sequences, the produced proteins may attain different tertiary structures due to the difference in their translational kinetics as synonymous codons are different with respect to their coding efficiencies. Thus, one cannot ignore synonymous mutations in the context of mutant phenotype. In this aspect, the seminal work on a mammalian protein demonstrated that the change of an optimal codon to rare codon had resulted in a protein whose structure was different from the original proteins leading to altered substrate specificity (Kimchi-Sarfaty *et al.* 2007). This observation has challenged the long standing Anfinsen hypothesis which had suggested that the primary structure of protein determines its 3-D structure (Anfinsen 1972) (figure 2). However, several studies now reveal that the presence of nonoptimal codons at specific parts of the coding regions of HEGs are important to allow the protein to fold properly by decelerating the translation rate at these sites. Thus the cotranslational protein folding is believed to be true for many proteins and is an important selection factor for the selective codon usage to optimize proper gene expression and function (Komar 2009). A web server CS and S has been created by scientists that predicts protein 3-D structure not only from the amino acid sequence encoded by a coding sequence but also takes into account the presence of the synonymous codons along the coding sequence (Saunders and Deane 2010). In a recent study it has been shown that the expression of 342 variants of an antibody coding sequence in *E. coli* only differing at synonymous codons results in protein with significant differences in protein solubility and functionality, while retaining the identical amino acid sequence (Hu *et al.* 2013).

It is worth discussing the evolutionary selection mechanism on codon usage in genes in the context of protein folding. Proper folding of a protein is important for its function. If folding is dependent upon translational elongation which, in turn, depends upon the use of proper synonymous codons, it can be argued that protein folding is indeed a selection mechanism of synonymous codons use in organisms (figure 3). The ultimate goal of gene expression is to produce a functional protein, i.e. a properly folded protein. In case of a HEG, where a large amount of protein is required inside the cell, the protein is needed to fold fast not only for its function but also to avoid unnecessary aggregation of the nascent polypeptides that are synthesized at high rate. For fast folding, a polypeptide is likely to be translated fast for which optimal codons are needed in the coding sequence. This might be a reason why HEGs contain more number of optimal codons in its coding sequence. But if a protein is needed to be translated slowly for its proper folding, even a HEG will have a higher proportion of nonoptimal codons. This might be a reason for the presence of a higher proportion of nonoptimal codons in *ompT*, a HEG in *E. coli* (dos Reis *et al.* 2003; Ishihama *et al*. 2008). The above hypothesis of protein folding can also explain CUB in LEGs. As expression of these proteins is low inside the cell, fast folding

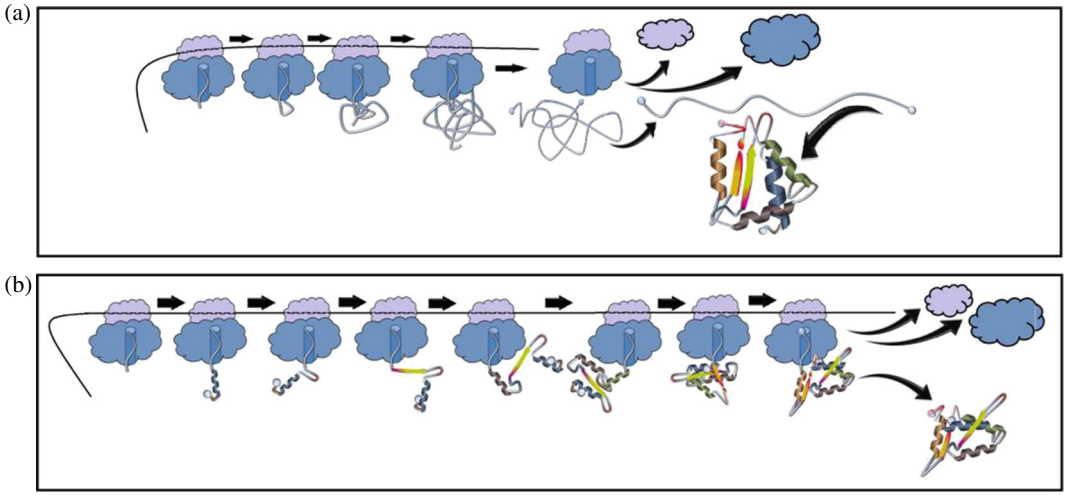

Figure 1. Schematic representation of posttranslational and cotranslational protein folding. The two panel figure illustrates two models of protein folding inside the cell. (a) Folding of the polypeptide occurs after it synthesized completely. This is referred to as posttranslational folding. (b) Folding of a polypeptide occurs along with its synthesis. This is referred to as cotranslational folding. This particular cotranslational folding is important from the biophysical point of view as the same synonymous change in the coding sequence of the mRNA result in same sequence of the protein, the final native structure of the protein is altered.

of these proteins is not essential. Therefore, translational kinetics in general might not be a selection factor for the evolution of CUB in these genes, but CUB is determined by mutational bias in LEGs. But in some LEGs, like *frq*, slow translational kinetics may be essential for proper folding of the protein where nonoptimal codons are under selection. Similarly, it should not be a surprise to observe some LEGs with higher number of optimal codons might be required to fold fast for its function.

Future research will prove if the above hypothesis is true or not. If 'folding directed codon usage bias' turns out to be true, then it will further challenge the theory of 'expression directed selection of synonymous codon' (Hershberg and Petrov 2008).

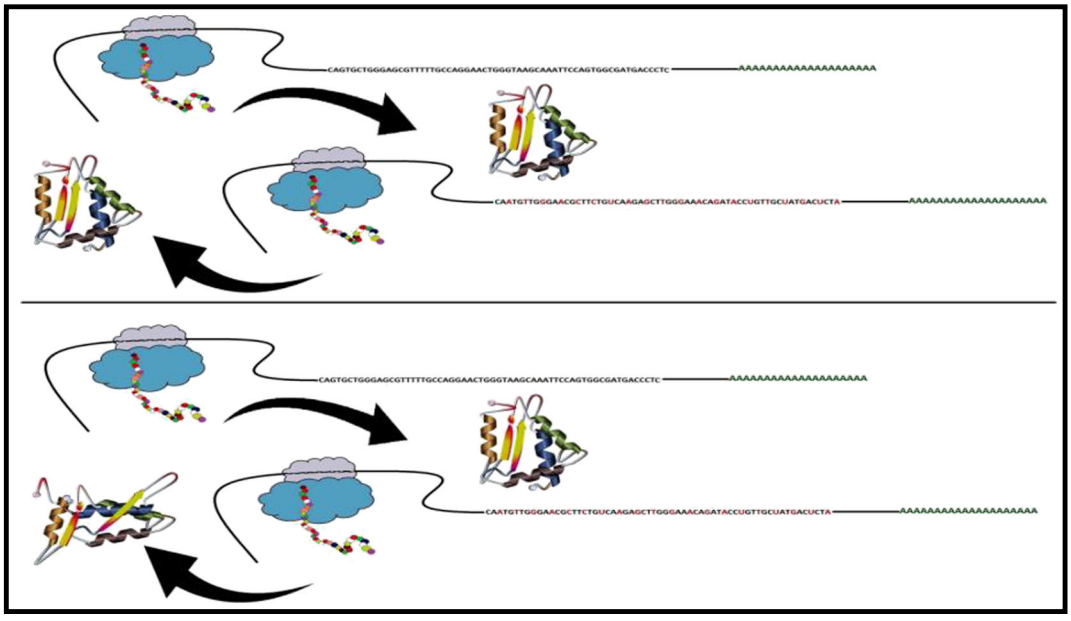

Figure 2. Consequence of translation of synonymous site in mRNA on protein folding. The two panels present a set of paired diagram to explain translation occurring in two mRNA only differing at synonymous site (change in nucleotide is shown in red colour). The upper panel depicts the conventional concept of the translational outcome. Here, the two proteins formed consequently after translation of mRNAs differing only at synonymous site have no change in their overall structure as their respective amino acid sequence remains the same. In the lower panel, the consequence of the different elongation rate of the growing polypeptide chain vis-à-vis translation of the mRNAs with synonymous mutation is reflected in the synthesis of proteins with modulated structure with the plausibility of altered structure.

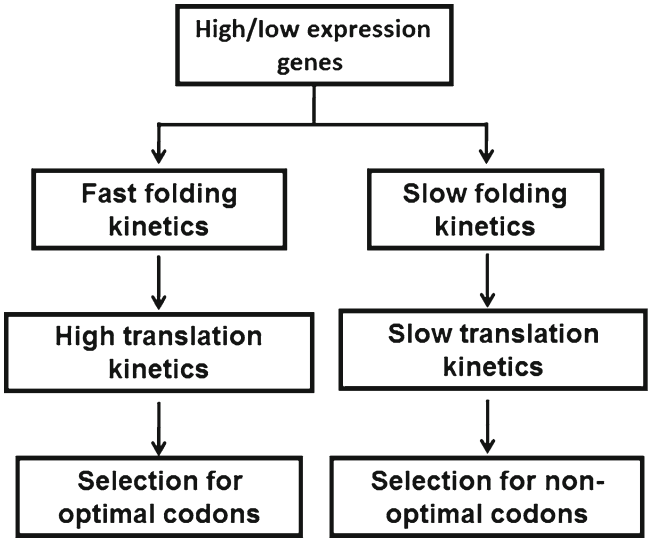

Figure 3. A schematic flow diagram demonstrating the hypothesis of protein folding kinetics influencing CUB in genes. Expression of a gene can be low or high. For a high expression gene, fast protein folding is required for its quick function and also to avoid protein aggregation due to large number of unfolded protein inside the cell. A fast protein folding kinetic is needed to have fast translational kinetics for which its coding sequence should have more optimal codons. But if a high expression protein is needed to be folded slow for its function then the coding sequence will select nonoptimal codon more for its slower translational kinetics. The same is also true for a low expression protein. As fast folding kinetics may not be essential for majority of low expression proteins, so selection of optimal codons is low in these genes. Similarly slow folding kinetics may not be essential for majority of low expression proteins for which the presence of nonoptimal codons might not be under purifying selection but due to mutational bias in these genes.

Acknowledgements

The authors are extremely grateful to the anonymous reviewers whose critical comments on the manuscript helped to compose the manuscript in a more focussed way and also significantly changed the discussion and conclusion. Authors also thank Dr Rocktotpal Konwarh, Tezpur University and Dr A. Bachawat, IISER, Mohali, for their comments on the manuscript. VJB thankfully acknowledges the receipt of his Research Associateship from the Department of Biotechnology (DBT), Govt. of India funded Bioinformatics Infrastructure Facility (BIF), Tezpur University. SKR and SSS are thankful to DBT, Govt. of India for the twinning project grant on codon usage bias under the research area Bioinformatics.

References

Anfinsen C. B. 1972 The formation and stabilization of protein structure. *Biochem. J.* **128**, 737–749.

Bulmer M. 1991 The selection-mutation-drift theory of synonymous codon usage. *Genetics* **129**, 897–907.

dos Reis M., Wernisch L. and Savva R. 2003 Unexpected correlations between gene expression and codon usage bias from microarray data for the whole *Escherichia coli* K-12 genome. *Nucleic Acids Res.* **31**, 6976–6985.

Ermolaeva M. D. 2001 Synonymous codon usage in bacteria. *Curr. Issues. Mol. Biol.* **3**, 91–97.

Ghaemmaghami S., Huh W. K., Bower K., Howson R. W., Belle A., Dephoure N., O'Shea E. K. and Weissman J. S. 2003 Global analysis of protein expression in yeast. *Nature* **425**, 737–741.

Gouy M. and Gautier C. 1982 Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Res.* **10**, 7055–7074.

Hershberg R. and Petrov D. A. 2008 Selection on codon bias. *Annu. Rev. Genet.* **42**, 287–299.

Hiraoka Y., Kawamata K., Haraguchi T. and Chikashige Y. 2009 Codon usage bias is correlated with gene expression levels in the fission yeast Schizosaccharomyces pombe. *Genes Cells* **14**, 499–509.

Hu S., Wang M., Cai G. and He M. 2013 Genetic code guided protein synthesis and folding in *E. coli*. *J. Biol. Chem.* **288**, 30855–30861.

Ishihama Y., Schmidt T., Rappsilber J., Mann M., Hartl F. U., Kerner M. J. and Frishman D. 2008 Protein abundance profiling of the *Escherichia coli* cytosol. *BMC Genomics* **9**, 102.

Kimchi-Sarfaty C., Oh J. M., Kim I. W., Sauna Z. E., Calcagno A. M., Ambudkar S. V. and Gottesman M. M. 2007 A "silent" polymorphism in the MDR1 gene changes substrate specificity. *Science* **315**, 525–528.

Komar A. A. 2009 A pause for thought along the co-translational folding pathway. *Trends Biochem. Sci.* **34**, 16–24.

Konigsberg W. and Godson G. N. 1983 Evidence for use of rare codons in the DnaG gene and other regulatory genes of Escherichia coli. *Proc. Natl. Acad. Sci. USA* **80**, 687–691.

Lobry J. R. and Sueoka N. 2002 Asymmetric directional mutation pressures in bacteria. *Genome Biol.* **3**, RESEARCH0058.

Martincorena I., Seshasayee A. S. N. and Luscombe N. M. 2012 Evidence of non-random mutation rates suggests an evolutionary risk management strategy. *Nature* **485**, 95–98.

Muto A. and Osawa S. 1987 The guanine and cytosine content of genomic DNA and bacterial evolution. *Proc. Natl. Acad. Sci. USA* **84**, 166–169.

Plotkin J. B. and Kudla G. 2011 Synonymous but not the same: the causes and consequences of codon bias. *Nat. Rev. Genet.* **12**, 32–42.

Palidwor G. A., Perkins T. J. and Xia X. 2010 A general model of codon bias due to GC mutational bias. *PLoS One* **5**, e13431.

Paul S., Million Weaver S., Chattopadhyay S., Sokurenko E. and Merrikh H. 2013 Accelerated gene evolution through replication– transcription conflicts. *Nature* **495**, 512–516.

Powdel B. R., Borah M. and Ray S. K. 2010 Strand-specific mutational bias influences codon usage of weakly expressed genes in Escherichia coli. *Genes Cells* **15**, 773–782.

Ran W. and Higgs P. G. 2010 The influence of anticodon-codon interactions and modified bases on codon usage bias in bacteria. *Mol. Biol. Evol.* **27**, 2129–2140.

Satapathy S. S., Dutta M., Buragohain A. K. and Ray S. K. 2012 Transfer RNA gene numbers may not be completely responsible for the codon usage bias in asparagine, isoleucine, phenylalanine and tyrosine in the high expression genes in bacteria. *J. Mol. Evol.* **75**, 34–42.

Satapathy S. S., Powdel B. R., Dutta M., Buragohain A. K. and Ray S. K. 2014 Selection on GGU and CGU codons in the high expression genes in bacteria. *J. Mol. Evol.* **78**, 13–23.

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

Sharp P. M. and Li W. H. 1986 Codon usage in regulatory genes in Escherichia coli does not reflect selection for 'rare' codons. *Nucleic Acids Res.* **14**, 7737–7749.

Sharp P. M. and Li W. H. 1987 The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. *Mol. Biol. Evol.* **4**, 222–230.

Sharp P. M., Bailes E., Grocock R. J., Peden J. F. and Sockett R. E. 2005 Variation in the strength of selected codon usage bias among bacteria. *Nucleic Acids Res.* **33**, 1141–1153.

Zhou M., Guo J., Cha J., Chae M., Chen S. and Barral J. M 2013 Non-optimal codon usage affects expression, structure and function of clock protein FRQ. *Nature* **495**, 111–115.

Received 17 March 2014, in final revised form 26 May 2014; accepted 30 May 2014
Unedited version published online: 24 June 2014
Final version published online: 14 October 2014