Innovare Academic Sciences

International Journal of Pharmacy and Pharmaceutical Sciences

ISSN- 0975-1491 Vol 7, Issue 3, 2015

Original Article

IMPACT OF tAI IN TRANSLATIONAL DYNAMICS OF *HOMO SAPIENS* GENES IN *ESCHERICHIA COLI* GENOME

BINATA HALDER[1], SUPRIYO CHAKRABORTY*[2]

[1, 2]**Department of Biotechnology, Assam University, Silchar 788011, Assam, India.**
Email: supriyoch_2008@rediffmail.com

Received: 08 Jan 2015 Revised and Accepted: 31 Jan 2015

ABSTRACT

Objective: Translation of mRNA to protein is a central biological process, and its regulation is vital for cell growth, development and differentiation. It is known that the tRNA molecules play a role of nearly 40% in the process of translation. The objective of this study was to find out the major DNA and protein determinants that play major role in heterologous expression of human genes in a prokaryote (*E. coli* K12 strain) for pharmaceutical applications.

Methods: In this article, we have analyzed the expression of 40 randomly chosen genes of *Homo sapiens* in *E. coli* K12 strain with tAI (tRNA adaptation index) as an expression measure in the background of E. coli tRNA gene pool using bioinformatic tools. We have studied how three major local features of a gene's coding sequence (namely coding sequence adaptation to the tRNA pool, nuleobases at three positions in codons and codon encoded-amino acid) affect the translation elongation process.

Results: Correlation analysis revealed T3 (r = 0.52**), and T1 (r= 0.51**) are positively correlated with tAI. The anticodon of tRNA genes refers the nucleobase T in the mRNA for translation efficiency. Moreover, AT1 (r= 0.28) and GC2 (r= 0.18) contents showed positive correlation with tAI indicating the role of AT composition at first position and the GC composition at third position of codon in human transgene expression in a bacterium. Several amino acids namely, Asp, Ile, Thr and Tyr showed highly significant positive correlation with tAI.

Conclusion: Our results suggest that nucleobase compositional dynamics is a determinant of transcriptional dynamics in heterologous gene expression.

Keywords: tAI- tRNA Adaptation Index, Translational activity, Synonymous codons, Gene expression.

INTRODUCTION

Expression of human genes in a bacterium like *E. coli* has a tremendous potential in commercial production of recombinant protein therapeutics. Large amount of these proteins could be produced over a short time period to save human lives as evident from insulin production in a bacterium. Recombinant proteins are widely utilized as tools in cellular and molecular biology. However, the production of recombinant proteins experiences many challenges such as loss of expression, post-translational processing, protein transport and localization. For the production of recombinant proteins, *E. coli* are widely used because of its fast growth rate and many other well-known genetic uniqueness [1]. Heterologous protein production of human may be diminished by biased codon usage. In such cases, if the codon bias of the genes to be expressed differs significantly from that used by *E. coli*, the newly synthesized protein using host cell rare codons will face many translational errors such as amino acid substitution, stalling, termination, and possibly frame-shifting [2]. Two codons, *i. e.* AGA and AGG encode the amino acid arginine. These are rarely used in *E. coli* and so is their cognate tRNA. These codons have been shown to cause mis-translational errors as well as lower levels of protein expression [3]. Mis-incorporation of rare CGG codons has also been observed [4].

Study on human genes is a major challenge to understand how post-translational events affect the activities and functions of the proteins in relation to health and disease. To overcome such challenges during gene translation, a number of statistical algorithms have been developed by the computational biologists to unravel the conundrums. Gene translation is one of the fundamental biological processes in all living organisms by which an mRNA sequence is decoded by the ribosome to synthesize a specific protein. Efficient translation of proteins, either in their natural biological context or in heterologous expression systems, amounts to maximizing protein production and at the same time minimizing the costs of the process. During the elongation stage of gene translation, each codon is iteratively translated by the ribosome to an amino acid. Translation elongation is known to be conserved in all living organisms [5]. Thus the understanding of this process and other related determinants have important ramifications for human health [6-8], biotechnology [9] and evolution [8, 10]. There are various factors that influence the translational activity and efficiency during the process of protein production in a cell. These include gene expression level [11-14], gene length [15-16], gene translation initiation signal [17], protein amino acid composition [18], protein structure [19], tRNA abundance [20-21], mutation frequency and patterns [22], and GC composition [23]. Out of these, tRNA plays a vital role in transporting the anticodons to pair with its respective codons. In order to investigate the role of tRNA in protein expression, we used tAI as a prime measure to gauge the expression levels in numerical values.

Most of the research works involving gene expression studies were carried out on several organisms like yeast, *Saccharomyces, E. coli, C. perfringens*. However, in case of *Homo sapiens* very few studies have been carried out, and these were mostly computational or experimental with small data sets. Particularly little attention was given to genes with low gene expression levels. We have taken a few therapeutic proteins for the study of expression level and performed various types of correlation and regression analysis centered on a total of 40 human proteins. The prime aim of this study is to predict the expression levels of these 40 genes with a prerequisite for expression in *E. coli* K12 genome using statistical parameters. These sequences can later be utilized to yield therapeutic proteins through recombinant DNA technology for saving human life.

MATERIALS AND METHODS

Materials

Quite a lot of studies and experiments on gene expression have been performed in order to produce the economically important proteins, but many of these did not yield promising results. Gene expression study has got tremendous potential in the production of therapeutic proteins. Forty coding sequences of human genes were retrieved from the NCBI (http://www.ncbi.nlm.nih.gov) database randomly. The serial numbers (SN), accession numbers, and other information are presented in supplementary sheet.

Methods

Basic data manipulation and statistical analysis were performed using Microsoft Excel. A PERL program developed by SC (author) was used to estimate the general properties of the nucleotides with respect to their composition at third position. The program also estimates the compositions of GC and AT at the three codon sites. Moreover, it also perceives the statistically relevant conserved rare codon clusters and the frequently used codons which are the causative reasons for the protein level to rise or drop during gene expression. Then we engrossed on the estimation of the gene expression pattern using tAI [24]. tAI is measured relative to the supply of the tRNAs that are required for translation of codons to amino acids. Since there are fewer tRNA molecules available than the number of codons in a cell, the cds sequence will be more efficient in translation by using not all the synonymous codons for an amino acid but by using a few restricted synonymous codons for the same amino acid. The tRNA availability is a driving force for translational selection. The tRNA adaptation index estimates the extent of adaptation of a gene (cds) to its genomic tRNA pool. It is a measure for predicting gene expression.

$$tAI_g = \left(\prod_{k=1}^{l_g} w_{i_k} \right)^{1/l_g}$$

Where l_g is the length of the gene in codons, and W_{ik} is the relative adaptiveness value of the codon defined by the k^{th} triplet in the gene.

RESULTS AND DISCUSSION

Our rationale is that if gene sequences are selected to maximize expression efficiency, we would expect a significant correlation between gene transcriptions, represented by tRNA levels, and other parameters related to translational activity. We selected a number of parameters representing the characteristics of the gene sequences that could be influential in the expression variation: coding sequence length, nucleotides and its composition at the different codon positions, translation efficiency measured by tAI and amino acid frequencies. The results of the detailed study on these human gene expressions and its parameters are as follows:

Analysis of the nucleotide bases

The randomly chosen coding sequences were analyzed thoroughly for their nucleotide composition. The gene sequences starting with the initiator codon (ATG), having the length as an exact multiple of three bases and devoid of N (any unknown base) were used in the study. Individual nucleotides as well as GC and AT content at three synonymous codon positions were calculated. The nucleotide composition in the analyzed genes is summarized in table I

Table I: Nucleotide composition in the genes of *Homo sapiens*

Cds No.	Accession No.	Gene Length	Total no. of codons	No. of amino acids	%A	%T	%G	%C
1	*BT019584.1*	873	279	290	30.01	22.57	26.46	20.96
2	*BT007060.1*	1689	542	562	22.38	19.24	29.90	28.48
3	*BT007413.1*	1260	405	419	28.65	22.46	20.63	28.25
4	*BT019496.1*	762	233	253	22.44	19.29	32.02	26.25
5	*BT006694.1*	378	121	125	38.36	16.93	26.72	17.99
6	*BT007395.1*	333	106	110	31.23	21.02	23.12	24.62
7	*BT007411.1*	819	270	272	22.83	21.73	28.69	26.74
8	*BT007345.1*	996	324	331	26.41	27.11	22.89	23.59
9	*BT007298.1*	669	205	222	27.35	17.79	31.99	22.87
10	*BT007058.1*	963	307	320	26.27	20.35	27.00	26.38
11	*BT007309.1*	1278	406	425	29.34	25.74	23.63	21.28
12	*BT007280.1*	1602	516	533	20.04	19.04	30.27	30.65
13	*BT007421.1*	2562	825	853	20.26	16.32	32.28	31.15
14	*BT007379.1*	774	253	257	21.19	17.83	32.30	28.68
15	*BT007446.1*	1458	463	485	32.30	21.67	26.54	19.48
16	*BT007314.1*	789	248	262	29.78	23.83	25.48	20.91
17	*BT007415.1*	1485	490	494	30.24	15.49	31.85	22.42
18	*BT019673.1*	972	311	323	16.05	14.20	36.83	32.92
19	*BT007433.1*	813	257	270	29.27	18.82	26.94	24.97
20	*BT007357.1*	777	250	258	29.21	15.44	27.54	27.80
21	*BT020171.1*	1638	527	545	26.86	16.30	30.04	26.80
22	*BT006833.1*	480	157	159	17.29	22.71	29.79	30.21
23	*BT007287.1*	930	299	309	26.24	19.57	26.99	27.20
24	*BT019491.1*	687	226	228	27.80	26.49	25.62	20.09
25	*BT007430.1*	609	193	202	19.05	28.24	28.90	23.81
26	*BT007428.1*	1203	391	400	29.18	24.52	21.03	25.27
27	*BT019517.1*	633	205	210	25.91	14.38	33.65	26.07
28	*BT007354.1*	327	103	108	39.45	18.65	26.91	14.98
29	*BT007211.1*	582	188	193	16.49	20.10	29.73	33.68
30	*BT007420.1*	1968	636	655	24.34	20.02	26.02	29.62
31	*BT019622.1*	1182	376	393	23.43	19.80	25.97	30.80
32	*BT007255.1*	972	316	323	17.39	15.23	32.92	34.47
33	*BT007416.1*	513	162	170	30.80	26.32	23.20	19.69
34	*BT019618.1*	729	230	242	33.33	24.97	20.99	20.71
35	*BT019518.1*	822	259	273	34.06	21.05	27.37	17.52
36	*BT007417.1*	1041	339	346	18.54	18.25	30.74	32.47
37	*BT007427.1*	1617	525	538	28.08	19.91	29.25	22.76
38	*L13470.1*	1248	397	415	27.96	27.16	21.31	23.56
39	*M14091.1*	1248	397	415	27.96	27.08	21.31	23.64
40	*JN849371.1*	582	186	193	19.42	20.27	29.21	31.10

Our analysis revealed that these human genes were abundant in guanine (average = 27.51%) and adenine (average = 26.10%) followed by cytosine (average = 25.63%) and thymine (average = 20.76%). Similarly the average percentages of GC (53.14%) were found to be higher than that of AT (46.86%). The overall percentages of GC3 (avg. = 60.09%) and GC1 (avg. = 58.21%) showed clear evidence of these genes being rich in GC content.

The GC content levies constraints on the codon usage and thus may indirectly affect the translation process. The %GC and %AT contents at their three respective codon positions and overall %GC and %AT content of cds are depicted in fig. [1].

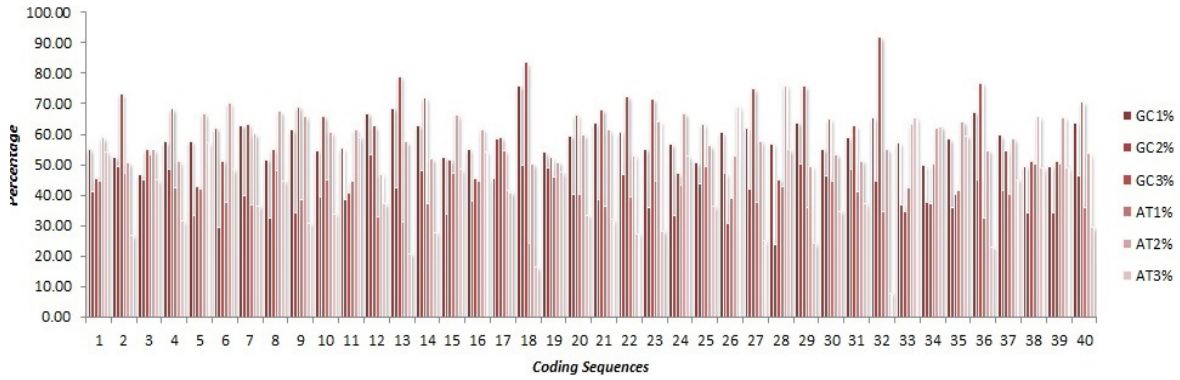

Fig. 1: Distribution of the %GC and %AT at three different codon positions

To predict the expression level of the cds, %GC, %AT and the presence of the four nucleotides were correlated with tAI. Our analysis reveals association between gene (cds) sequence features and gene expressivity measure (tAI). The correlation values of the sequence features with tAI indicate that selection for gene sequence characteristics towards expression efficiency in human genes may be relevant. We perceived that correlations between gene expression as measured by tAI and %GC-AT contents at any codon site are very weak (rGC1=-0.30, rGC2=0.18 and rGC3=-0.04). But in paradox with others, %GC content at the third codon position comes out to be a very poor predictor of gene expression [25-26]. The weak correlation between GC at any codon site and gene expression reveals variability of gene expression and suggests that there might be the existence of other evolutionary forces affecting human gene expression in *E. coli* genome. Correspondingly the %AT frequency at each codon site was also analyzed, and it was found that the AT content at second codon position showed a moderate negative correlation (rAT2=-0.19) with tAI. However, AT1 and AT3 showed positive correlation (rAT1= 0.28, rAT3= 0.06) with tAI, respectively as shown in fig. [2]. This reveals that AT content of cds might affect gene expressivity in *E. coli*.

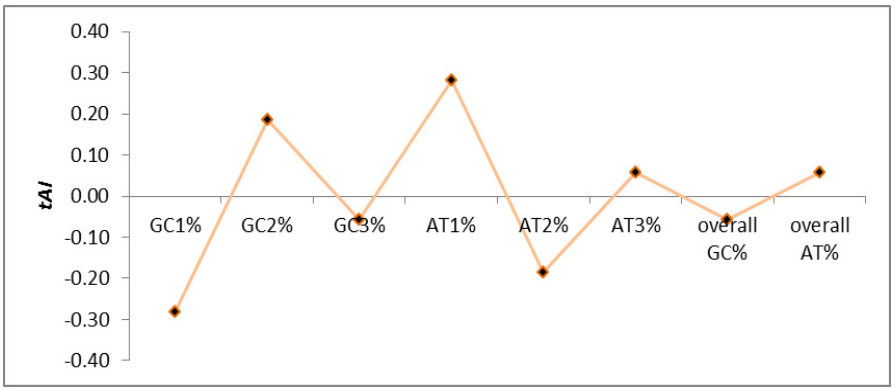

Fig. 2: Correlation analyses of %GC and %AT at three different codon positions with tAI

We have also confirmed the presence of G3 and C3 in much higher frequency than A1 and A3, which exhibited lower values as presented in fig. [3].

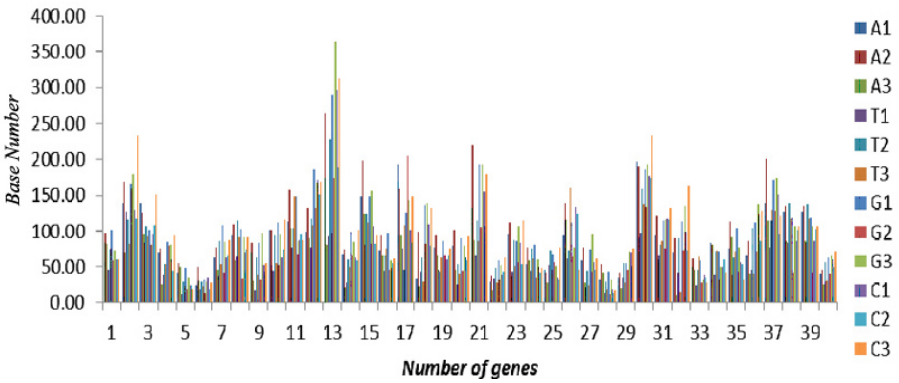

Fig. 3: Distribution of bases at three codon positions in 40 cds of *Homo sapiens*

To detect possible relationship between the base composition at different synonymous codon positions and tAI, the estimated values of the four nucleotides adenine, thymine, guanine and cytosine were compared with tAI value of each coding sequence. The correlation coefficient (r) of the positional attributes of the nucleotide bases with tAI were moderately higher for T3 (r=0.52**) followed by T1 (r=0.51**), and showed trifling lower values in the other positions of nucleotides in the synonymous codons. G3 (r=0.209) had the least influence of the codon usage bias, which implied that the preference of nucleotide composition on the synonymous codon usage might have a role in altering the degree of gene expression. The base compositions were most likely influenced by thymine in first and third positions of synonymous codons. Fig. [4] shows the correlation coefficient values of tAI with the nucleotide bases at three codon positions in the human genes.

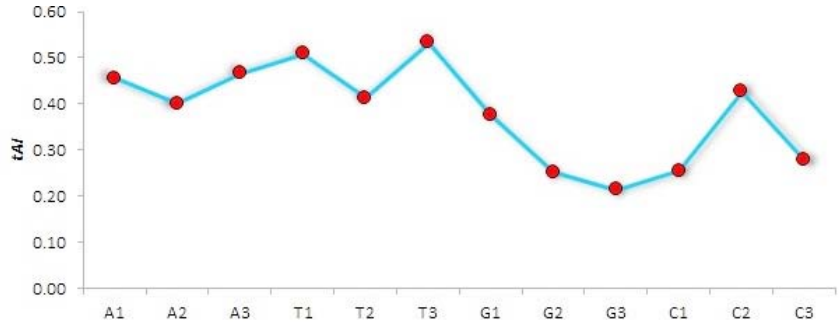

Fig. 4: Correlation of nucleobases at three different codon positions with tAI

Codon Analysis

The number of each codon in the coding sequences was analyzed as depicted in fig. [5]. The average of all the codons was calculated and it was found that the codons GAA (*avg.* =14.74), CTG (*avg.* =13.19), CAG (*avg.* =12.61) and AAG (*avg.* =12.60) have higher average values than the rest of the codons. It reveals that these specific codons more frequently used.

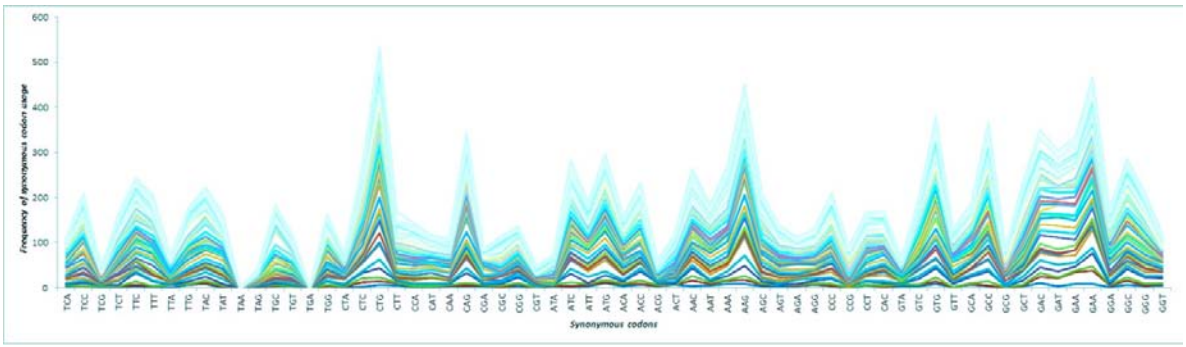

Fig. 5: Frequency of codon usage in the cds of *Homo sapiens*

Codon usage pattern of the synonymous codons for the genes was analyzed**.** From our analyses it was evident that the codons were inconsistent in distribution and were not equally synonymously used. Several early studies focused on the correlation between tRNA content and codon usage within and/or in between species [27-28]. The complementary approach, *i. e.* understanding the distribution of tRNA gene number and anti-codon type has been much less developed in the framework of comparative genomics. The total number of tRNA genes in any organism remains the same for all the coding sequences. In order to better understand the process of protein prediction *in silico*, we correlated the usage of 61 codons with tAI value in *E. coli* K12 strain. The graphical representation of the correlation analysis is given in fig. [6].

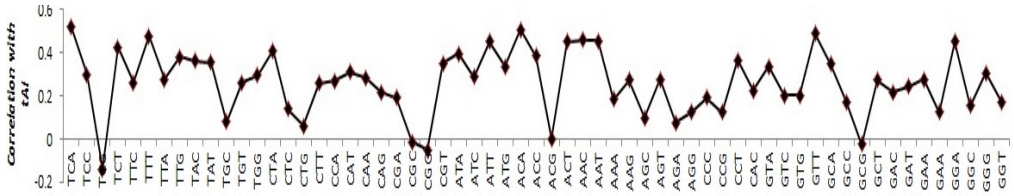

Fig. 6: Correlation of 61 codons with tAI

The plot revealed that the codons TCA (*r*=0.52**), ACA(*r*=0.51**), GTT(*r*=0.49**) and TTT (*r*=0.48**), ATT (*r*=0.45**), TCT (*r*=0.43**) showed high positive correlation with tAI while the rest of the codons showed low correlation. This analysis provides an indication that the strength of association between codon bias and translational efficiency is dependent on the levels of codon usage. This is in accordance with the previous study reported by Tuller *et al.,* 2010 that a significant association between codon bias and translation efficiency across all endogenous genes in *E. coli* and *S. cerevisiae* was observed [29]. It demonstrates the role of codon bias

as an important determinant of translation efficiency. From this we can assume that the codon – tRNA adaptation may serve as a code in determining the translational speed and efficiency.

Distribution of amino-acids and their association with gene expressivity measure

We computed the number of different anticodons (tRNA genes) present in the tRNA gene sets, decoding standard 20 amino acids in *E. coli* K12, and their distribution in the coding sequences. The plot revealed that amongst all the amino-acids, leucine (*avg.* = 31.90), serine (*avg.* =26.45), glutamic acid (*avg.* =25.80) and alanine (*avg.* =22.35) were present in higher amounts**.**

Interestingly the correlation analysis between tAI and different amino acids revealed that the amino acids Asp (r=0.57***), Ile (r=0.56***), Thr (r=0.50**) and Tyr (r=0.47**) showed highly significant positive correlation with tAI. This suggests that the codons which showed positive correlation with tAI could influence tAI value through the corresponding amino acid fig. [7].

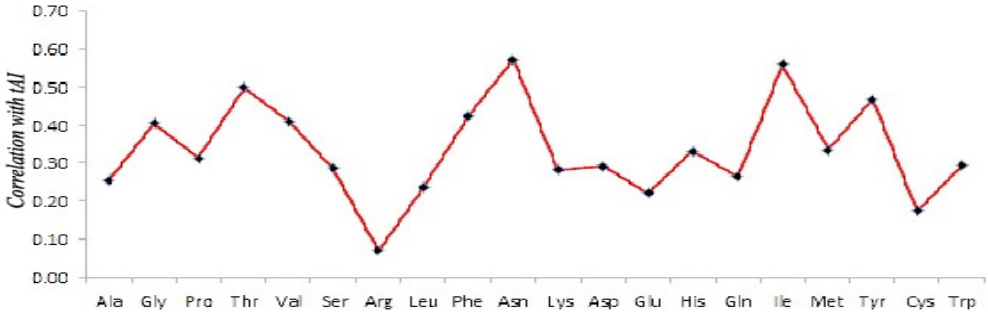

Fig. 7: Correlation of amino acids with gene expressivity measure tAI for cds of Homo sapiens

Another possible explanation for these results is that high demands for translation efficiency may occur at the expense of the demand of tRNA genes carrying amino-acids and the supply of codons on the chain of mRNA, which may have played a relevant role on the evolution of the human gene sequences. These correlations between amino acids and tAI indicate that sequence characteristics that modulate transcription and translation processes co-evolve in order to optimize tRNA usage based on the codon bias. The correlations suggest that selection for gene sequence characteristics could influence expression efficiency of human genes in a heterologous system in *E. coli* K12 genome.

CONCLUSION

In order to understand the translational dynamics of human genes in a prokaryotic system, we have analyzed tAI as a measure of gene expression for forty randomly chosen human coding sequences in the background of *E. coli* K12 tRNA gene pool. Our analysis revealed that tAI is positively correlated with T3 (r = 0.52**) and T1 (r =0.51**) nucleobases, respectively. The composition of AT bases at first position and that of GC at second position of codons influenced human gene expression as evident from the positive correlation of AT 1% and GC 2% each with tAI. This suggested the role of nucleotide composition in the heterologous expression of human proteins. Moreover, a few amino acids namely Asp, Ile, Thr and Tyr showed highly significant positive correlation with tAI. It could be attributed to high positive correlation of tAI with one or more codons encoding the specific amino acid. Our results suggest that nucleobase compositional dynamics is a determinant of transcriptional dynamics in heterologous gene expression.

ACKNOWLEDGEMENT

We are thankful to Assam University for providing the necessary facilities to undertake the research work. The first author is thankful to UGC for providing the Rajiv Gandhi National Fellowship for the financial assistance to carry out the research work.

ABBREVIATION

E. coli-Escherichia coli, *C. perfringens-Clostridium perfringens*, *S. cerevisiae-Saccharomyces cerevisiae*, tAI-tRNA adaptation index, A-Adenine, TThymine, G-Guanine, C-Cytosine, cds-Coding sequence, Asp -Aspartic acid, lle -Isoleucine, Thr -Threonine, Tyr-Tyrosine.

CONFLICT OF INTERESTS

Declared None

REFERENCES

1. Makrides SC. Strategies for achieving high-level expression of genes in *Escherichia coli*. Microbiol Rev 1996;60(3):512-38.
2. Kane JF. Effects of rare codon clusters on high-level expression of heterologous proteinsin *Escherichia coli*. Curr Opin Biotechnol 1995;6(5):494-500.
3. Calderone TL, Stevens RD, Oas TG. High-level mis-incorporation of lysine for arginine at AGA codons in a fusion protein expressed in *Escherichia coli*. J Mol Biol 1996;262(4):407-12.
4. McNulty DE, Claffee BA, Huddleston MJ, Porter ML, Cavnar KM, KaneJF. Mistranslational errors associated with the rare arginine codon CGG in *Escherichia coli*. Proteinexpression Purification 2003;27(2):365-74.
5. Kapp LD, Lorsch JR. The molecular mechanics of eukaryotic translation. Annu Rev Biochem 2004;73:657-704.
6. Kimchi-Sarfaty C, Oh JM, Kim IW, Sauna ZE, Calcagno AM, Ambudkar SV, *et al*. A "silent" polymorphism in the MDR1 gene changes substrate specificity. Sci 2007;315(5811):525-8.
7. Bahir I, Fromer M, Prat Y, Linial M. Viral adaptation to host: a proteome-based analysis of codon usage and amino acid preferences. Mol Syst Biol 2009;5:311.
8. Drummond DA, Wilke CO. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 2008;134(2):341-52.
9. Kudla G, Murray AW, Tollervey D, Plotkin JB. Coding-sequence determinants of gene expression in *Escherichia coli*. Sci 2009;324(5924):255-8.
10. Warnecke T, Hurst LD. GroEL dependency affects codon usage--support for a critical role of misfolding in gene evolution. Mol Syst Biol 2010;6:340.
11. Gouy M, Gautier C. Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res 1982;10(22):7055-74.
12. Sharp PM, Li WH. An evolutionary perspective on synonymous codon usage in unicellular organisms. J Mol Evol 1986;24(1-2):28-38.
13. Sharp PM, Li WH. The codon Adaptation Index--a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 1987;15(3):1281-95.
14. Chakraborty S, Paul P. GC2 biology dictates gene expressivity in *camellia sinensis*. Comput Mol Biol 2014;4:4.
15. Bains W. Codon distribution in vertebrate genes may be used to predict gene length. J Mol Biol 1987;197(3):379-88.
16. Eyre-Walker A. Synonymous codon bias is related to gene length in *Escherichia coli*: selection for translational accuracy? Mol Biol Evol 1996;13(6):864-72.

17. Ma J, Campbell A, Karlin S. Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures. J Bacteriol 2002;184(20):5733-45.
18. Lobry JR, Gautier C. Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes. Nucleic Acids Res 1994;22(15):3174-80.
19. D'Onofrio G, Ghosh TC, Bernardi G. The base composition of the genes is correlated with the secondary structures of the encoded proteins. Gene 2002;300(1-2):179-87.
20. Ikemura T. 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 1981;151(3):389-409.
21. Ikemura T. Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and *Escherichia coli* with reference to the abundance of isoaccepting transfer RNAs. J Mol Biol 1982;158(4):573-97.
22. Sueoka N. Two aspects of DNA base composition: G+C content and translation-coupled deviation from intra-strand rule of A = T and G = C. J Mol Evol 1999;49(1):49-62.
23. Sueoka N, Kawanishi Y. DNA G+C content of the third codon position and codon usage biases of human genes. Gene 2000;261(1):53-62.
24. Dos Reis M, Savva R, Wernisch L. Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res 2004;32(17):5036-44.
25. Sharp PM, Lloyd AT. Regional base composition variation along yeast chromosome III: evolution of chromosome primary structure. Nucleic Acids Res 1993;21(2):179-83.
26. Marin A, Gallardo M, Kato Y, Shirahige K, Gutierrez G, Ohta K, Aguilera A. Relationship between G+C content, ORF-length and mRNA concentration in *Saccharomyces cerevisiae*. Yeast 2003;20(8):703-11.
27. Ikemura T. Codon usage and tRNA content in unicellular and multicellular organisms. Mol Biol Evol 1985;2(1):13-34.
28. Duret L. Evolution of synonymous codon usage in metazoans. Curr Opin Genet Dev 2002;12(6):640-9.
29. Tuller T, Carmi A, Vestsigian K, Navon S, Dorfan Y, Zaborske J, *et al*. An evolutionarily conserved mechanism for controlling the efficiency of protein translation. Cell 2010;141(2):344-54.