# MAGNITUDE OF THYMINE IN DIFFERENT FRAMES OF MESSENGER RNAs

## RAJASEKARAN E.[1]*, ASHA JACOB[1] AND KLAUS HEESE[2]

[1]School of Biotechnology and Health Sciences, Karunya University, Karunya Nagar, Coimbatore-641114, TN, India.
[2]School of Biological Sciences, Nanyang Technological University, Singapore- 637551, Singapore.
*Corresponding Author: Email- sekaran@karunya.edu

**Abstract-** Thymine is the one and only base transcribed into uracil during production of proteins. Thymine in DNA and uracil in mRNA plays a major role in producing proteins with appropriate carbon content for stability and activity. Thymine distribution is different frames of coding nucleic acids are investigated statistically. The results confirm that frame 1 supposed to have definite thymine content. Frame 3 prefers to have least thymine content. Frames 4 & 5 maintain some degree of thymine while 2 & 6 have a variable fraction of thymine.

**Keywords-** Comparative genomics, thymine distribution, thymine content, coding frame, CDS, mRNA.

**Citation:** Rajasekaran E.,  Asha Jacob and Klaus Heese (2012) Magnitude of Thymine in Different Frames of Messenger RNAs. International Journal of Bioinformatics Research, ISSN: 0975-3087 & E-ISSN: 0975-9115, Volume 4, Issue 3, pp.-273-275.

## Introduction

Thymine the one and only base transcribed into different base (uracil) during transcription. Though oxygen is added in every nucleotide on transcription, the uracil loses one methyl group which is present in thymine. The transcribed RNA becomes hydrophilic than the corresponding DNA. Thymine becomes important both in terms of translation and intermolecular interaction in biology. How important in translating mRNA into protein? Suppose a codon containing thymine at the centre, it always code for large hydrophobic residues. The synthesized proteins contain mix of these 20 naturally occurring amino acids. The mixing is done in such way that it maintains 31.45% of carbon in its structure for stability [10]. This is not only in global but at local as well. The thymine plays a major role in doing so. So it is necessary to quantify the role of thymine in producing proteins with adequate carbon distribution. Reduction of thymine in mRNA sequences [8] is a concern in human proteins during evolution [9].

There is record of evidences state that specific mutations within the AT-rich region of replication origin affect either origin opening or helicase loading [5]. This is both in prokaryotic and eukaryotic replicons. Sequence variation causes polymorphism. Seven polymorphic sites were identified in human leptin receptor gene in lean and obese Indians [3]. Two such polymorphisms are in coding region, one polymorphism is a silent one and four occur in non-coding regions. Four of these sites are in linkage disequilibrium with one another. Nucleotides at three noncoding polymorphic sites were found exclusively in obese Pima Indians. This demonstrates an association between variation at the leptin receptor gene and obesity in humans. After analyzing several introns and exons of protein coding genes [6], it is reported that (1) in most exons, adenine is increased over the thymine. In other words, adenine and thymine are distributed in an asymmetric way between the exon and the complementary strand, and the coding sequence is mostly located in the adenine-rich strand. (2) Thymine dominates over adenine not only in the strand complementary to the exon but also in introns. (3) A general bias is further revealed in the distribution of adenine and thymine among the three codon positions in the exons, where adenine dominates over thymine in the second and mainly the first codon position while the reverse holds in the third codon position. One of the A-T rich genome containing species, Plasmodium falciparum has 82% of AT content [11]. The coding regions contain 69% and non coding region contain 86%. Within the coding sequences, the A/T ratio was 1.68 in the mRNA sense strand, and overall A + T content in the three codon positions increased in the order 1st-2nd-3rd position. Codons with T or especially A in the third position were strongly preferred. Codon usage among individual parasite genes was very similar compared to genes from other species. Dinucleotide frequencies for the parasite DNA were close to those expected for a random sequence with the known base composition, except that the CpG frequency in the coding sequences was low. Codon usage in selected AT-rich bacteria [12] suggests that use U or A in the first and third positions of the codon when possible. Comparisons of codon usage between the two organisms reveal that preferential use of A- and U-rich codons [7]. More than 90% of the third positions and 57% of the first positions is either A or U while in other species it is 51% and 36% respectively. The biased choice of the A- and U-rich codons has been

observed in the codon replacements for conservative amino acid substitutions. A frequency analysis of codon usage in different frames deduce that the RNY model (R = purine = A or G, Y = pyrimidine = C or T, N = R or Y) [2].

We have reported earlier that frame 1 of coding sequences maintain about 27% of thymine [1]. This has been investigated again and compared in human, yeast and viral genome. That is the magnitude of thymine in different frames of coding regions is investigated.

## Methodology

The mRNA sequences of human, yeast and Influenza A virus are retrieved from NCBI. The thymine content in each frames of different sequence is computed using XTX tool available online. It tabulates the number of thymine in all six frames plus the total number of nucleotides. From this the fraction of thymine (thymine in the concerned frame divided by total number of bases) in each frame is computed. It is computed for all sequences of different species. A total of 100 sequences are taken from each species and computed thymine fraction. The modified version of XTX tool called DNAFRAME is used to compute the thymine fraction in one goes for all sequence. A table of number of sequences (frequency) with different fractions of thymine is created for each frame. One can make a graph of thymine fraction versus frequency. From the graph or table, a frequent thymine fraction is considered as probable thymine content in the particular frame. But actually the mean of the distribution is taken as the preferable value here. This value in each frame is calculated and tabled like this.

*Table 1- The Mean Thymine Fraction in Different Frames*

| Frame | Mean thymine fraction |
|-------|----------------------|
| 1 | 0.084 |
| 2 | 0.066 |
| 3 | 0.054 |
| 4 | 0.1 |
| 5 | 0.081 |
| 6 | 0.056 |

A plot of thymine fraction versus frame number is plotted for comparison and discussed. A standard deviation of the distribution is computed side by side and made another table and plot for comparison. The standard deviation is to see how significant the thymine fraction in different frames.

## Results and Discussion

The earlier works on thymine distribution conclude that frame 1 should have definite amount thymine for translating mRNA into proteins with adequate large hydrophobic residues for defined carbon content [1,4]. It is confirmed the same in different species studied here. Beyond that what happens in other frames and species? This is addressed here.

### The Mean Thymine Fraction in Different Frames

Figures 1-3 show the mean thymine fraction and standard deviation in different frames of mRNA sequences of human, yeast and Influenza virus A respectively. The frame 1 in all the species remain same (~0.09) in thymine fraction, though there is varying degrees in total. Similarly the frame 3 contains least amount of thymine. The varying number of thymine in frame 2 reveals that

extra thymine is tolerated as it does not make a difference in translation. Frame 4 tries to have higher thymine than in frame 1 which gives mostly hydrophilic residues during translation. Frame 5 maintains lower thymine than in frame 4. This is exceptional in virus sequence. Probably this makes the virus different from normal species. The viral genome is different from normal at the strand 2 as well. That is frame 4-6 are differing. Frame 6, the complementary frame to 2, is varying degree of thymine. Here again thymine is adjusted.

The human mRNAs generally contain less thymine than yeast. During evolution the thymine got reduced [9]. The reduction is mostly in frame 2 and 6. Similarly the high thymine containing yeast, the excess thymine is adjusted in frame 2 and 6.
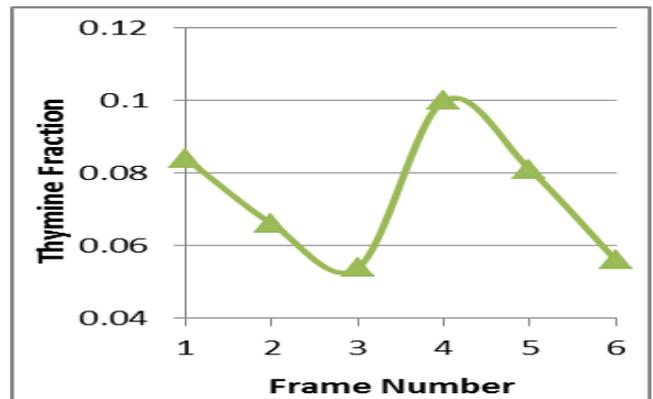


**Fig. 1A-** Thymine fraction in different frames of human mRNAs.
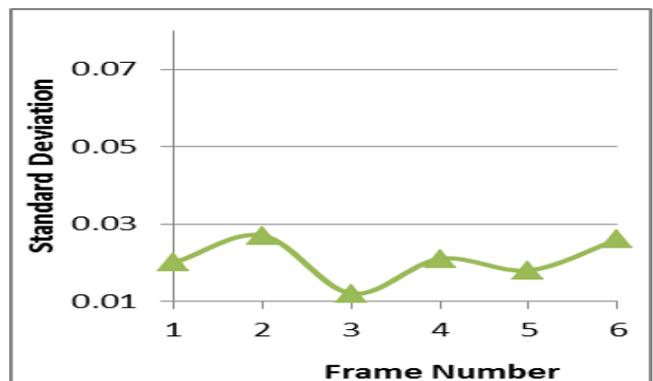


**Fig. 1B-** Standard deviation of ditribution in different frames of human mRNAs.
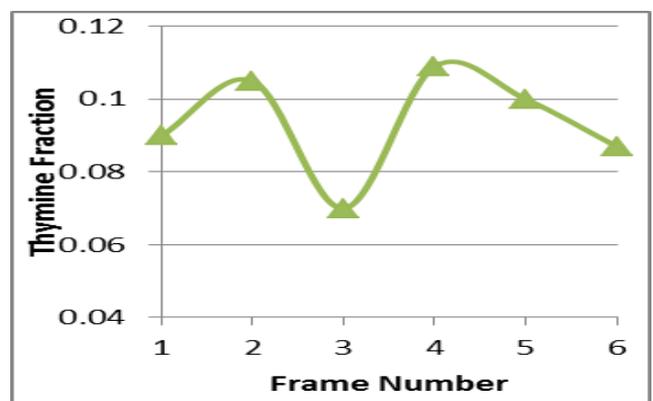


**Fig. 2A-** Thymine fraction in different frames of yeast mRNAs.
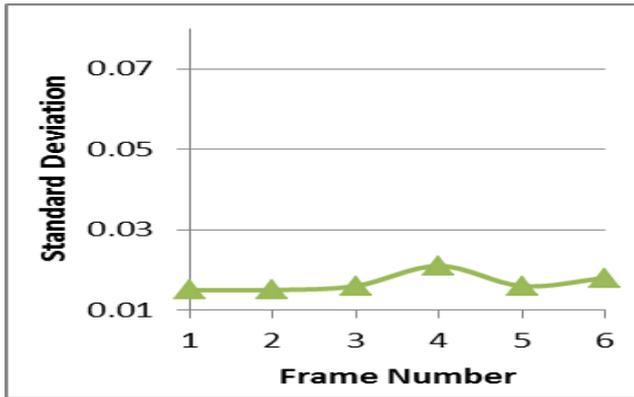
**Fig. 2B-** Standard deviation of ditribution in different frames of yeast mRNAs.
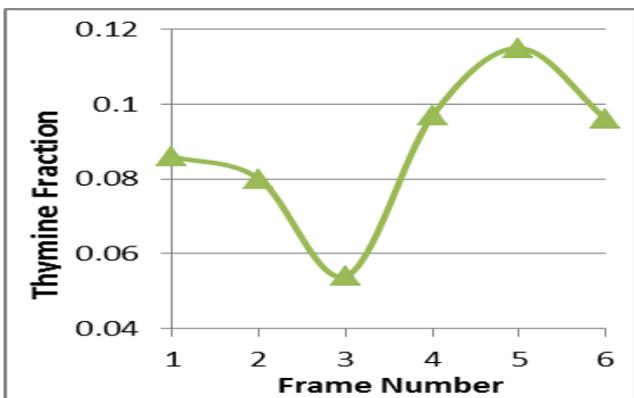


**Fig. 3A-** Thymine fraction in different frames of Influenza virus.
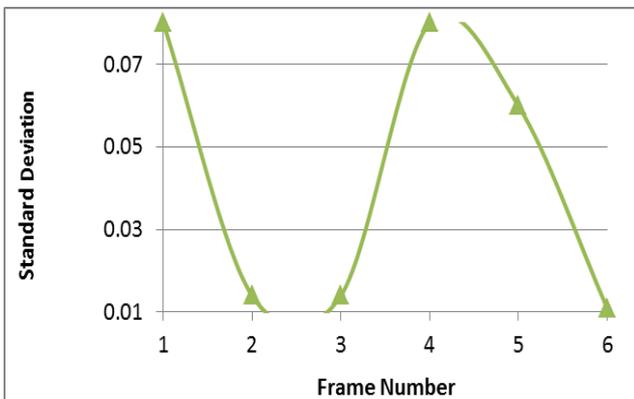


**Fig. 3B-** Standard deviation of ditribution in different frames of Influenza virus

## Standard Deviation of Thymine Distribution in Different Frames

The standard deviation graph for human, yeast and Influenza virus A is shown in figures 1(B), 2(B) and 3(B). When the distribution is broad or not normal then the standard deviation becomes high and significant. Compared to Influenza virus, the human and yeast show less standard deviation in all frames, stating that variation in thymine content among the sequences are less. In human, the frame 2 is having higher standard deviation than the other frames. This means that a varying number of thymine in frames 2 in differ-

ent sequence observed. Frame 6 observed to be the same. Frame 3 shows a least values, stating that variation is not high. This means that the frame 3 should have a small but defined number of thymine. Frames 1, 4 & 5 have a defined value of standard deviation. In yeast, all frames are having less and equal value of standard deviation suggest that well defined order of thymine in all frames of yeast sequences. There is slight difference in frame 4. The complementary strand has variation in thymine distribution. The virus sequences give a reverse trend of having high values for frames 1, 4 & 5 which is unusual. Again the frames 2 & 6 have least value stating the no variation in thymine content in these frames. This is the reason why the viral sequences are different from normal one.

## Conclusion

Thymine in protein coding sequences are analysed in different frames which conclude that the presence of thymine is not random but with certainty. Frame 1 maintains to have a definite amount of thymine. Frame 3 should have a least amount of thymine. Variation in thymine content in frame 6 can be tolerated. Viral genome is found to have different in frame 5 compared to others.

## References

[1] Anandagopu P., Suhanya S., Jayaraj V. and Rajasekaran E. (2008) *Bioinformation,* 2, 304-307.

[2] Arquès D.G. and Michel C.J. (1996) *Theor. Biol.,* 182(1), 45-58.

[3] Bruce Thompson D., Eric Ravussin, Peter H. Bennett and Clifton Bogardus (1997) *Human Molecular Genetics*, 6(5) 675–679.

[4] Jayaraj V., Suhanya S., Vijayasarathy M., Anandagopu P. and Rajasekaran E. (2009) *Bioinformation,* 3, 409-412.

[5] Magdalena Rajewska, Lukasz Kowalczyk, Grazyna Konopa and Igor Konieczny (2008) *Proc. Natl. Acad. Sci.,* 105(32), 11134-11139.

[6] Mrázek J. and Kypr J. (1994) *Mol. Evol.*, 39(5), 439-47.

[7] Muto A., Kawauchi Y., Yamao F. and Osawa S. (1984) *Nucleic Acids Res.*, 12(21), 8209-17.

[8] Rajasekaran E. and Anandagopu P. (2010) *J. Adv. Biotech*, 9, 9-10.

[9] Rajasekaran E. Rajadurai M., Vinobha C.S.V. and Senthil R. (2008) *Comput. Intelli. Bioinfo.,* 1, 109-114.

[10] Rajasekaran E., Vinobha C.S., Vijayasarathy M., Senthil R. and Shankarganesh P. (2009) *IACSIT-SC,* 5, 452- 453.

[11] Weber J.L. (1987) *Gene.*, 52(1), 103-9.

[12] Winkler H.H. and Wood D.O. (1988) *Biochimie*. 70(8), 977-86.