Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

Role of thymine in protein coding frames of mRNA sequences

Perumal Anandagopu[1], Sivakumar Suhanya[2], Veerasamy Jayaraj[1] and Ekambaram Rajasekaran[1,*]

[1]Department of Biotechnology, Periyar Maniammai University, Thanjavur, Tamil Nadu-613403, India; [2]Bharathidasan University, Tiruchirapalli, Tamil Nadu-620024, India; Ekambaram Rajasekaran – E-mail: ersekaran@gmail.com; * Corresponding author

received December 31, 2007; revised March 15, 2008; accepted April 07, 2008; published April 19, 2008

Abstract:
Distribution of thymine in protein coding mRNA sequences has been studied here. Our study suggest that thymine in protein coding sequences are not randomly distributed but with probability. Frame1 prefers to have definite amount of thymine. It is observed that the thymine content of frame 4 is also involved in protein coding. Frame 3 prefers to have least amount of thymine. However, frame 2 and frame 6 shows a variable degree of thymine content. The mRNA sequences of heterosexual animals, particularly, the human show a different distribution profile (less thymine in frame 1) compared to that of yeast and plants.

Keywords: frame analysis; thymine distribution; sequence analysis; protein stability

Background:
Though all living systems share a common platform of DNA-RNA-Protein relationship; it differs in many for various life related activities. The DNA and RNA use only four types of bases to represent the living system. There are attempts to understand the distribution of these bases in protein coding nucleic acids sequences. Kolaskar and Reddy used content and signals methods to identify protein-coding regions in prokaryotes **[1]**. Raghava and colleagues developed a web-based tool for analyzing nucleotides sequences in identifying protein-coding sequences **[2]**. Fickett reported a new algorithm called TESTCODE to distinguish coding sequences from non-coding sequences **[3]**. Wong and colleagues developed a statistical method (mutagenesis assistant program) to equip protein engineers with a tool to develop promising directed evolution strategies **[4]**.

Thymine is the only residue replaced to uracil in the base upon transcription. This leaves the mRNA less hydrophobic compared than its DNA counterpart. Hydrophobic interaction is the dominant force that drives biological reaction in living systems. The distribution of thymine in protein coding mRNA sequences and its relationship with protein stability and function are the main focus of this work, assuming that the understandings from these analyses will pave the way for solving the long-standing issue of diseases control. The conclusions drawn from this study are general in nature for mRNA sequences, which are drawn at species level. Similar calculations on individual mRNA sequences are expected to give the same results.

Methodology

Dataset

The complete sets of protein coding mRNA sequences of *H. sapiens, B. taurus, C. familiaris, G. gallus, D. melanogaster, A. mellifera, D. rerio, C. elegans, A. thaliana, S. cerevisiae and V. cholera* are taken from NCBI genome set **[5]** in FASTA format. Total number of mRNA sequences in the given species is given in Table 1 (see supplementary material).

Definitions, assignments and analysis

The frame 1, 2 and 3 are representing the 5'-3' of mRNA sequences and frame 4, 5 and 6 are the reverse complementary strand. Each mRNA sequences are read in different frames and counted XTX (X=A, T, G or C) in each frames as illustrated in Figure 1.

The thymine in the second position of the codons is considered as frame1. Nucleotides 1 and 3 of the codons are considered as frames 3 and 2 respectively. The complementary nucleotide of 1, 2 and 3 nucleotides of codons are considered as frame 6, 4 and 5, respectively. The number and fraction of thymine in all six frames in each sequence of every species are calculated using a program written in 'C'. This is the number of XTX in all six frames and the total number of base pairs are counted in each mRNA sequences and tabulated. The fraction of XTX in each frame (i.e., ratio between the XTX of individual frames and total) is also computed and tabulated. These fractional quantities are then grouped based on the thymine content and plotted as shown in

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

figure 2. We then calculated thymine content at which maximum frequency is identified for each frame.

Strand1: 5'-ATGCGGTTGAATTCGTATGCT…-3' (mRNA sequence)
Strand2: 3'-TACGCCAACTTAAGCATACGA…-5' (complementary strand)

Frame No.	Codons	No. of XTX
1	5'-ATG CGG TTG AAT TCG TAT GCT…-3'	2
2	5'-TGC GGT TGA ATT CGT ATG… -3'	2
3	5'-GCG GTT GAA TTC GTA TGC…-3'	3
4	3'- TAC GCC AAC TTA AGC ATA CGA…-5'	2
5	3'-CGC CAA CTT AAG CAT ACG…-5'	1
6	3'-ACG CCA ACT TAA GCA TAC…-5'	None

Figure 1: Reading of mRNA sequences in different frames.

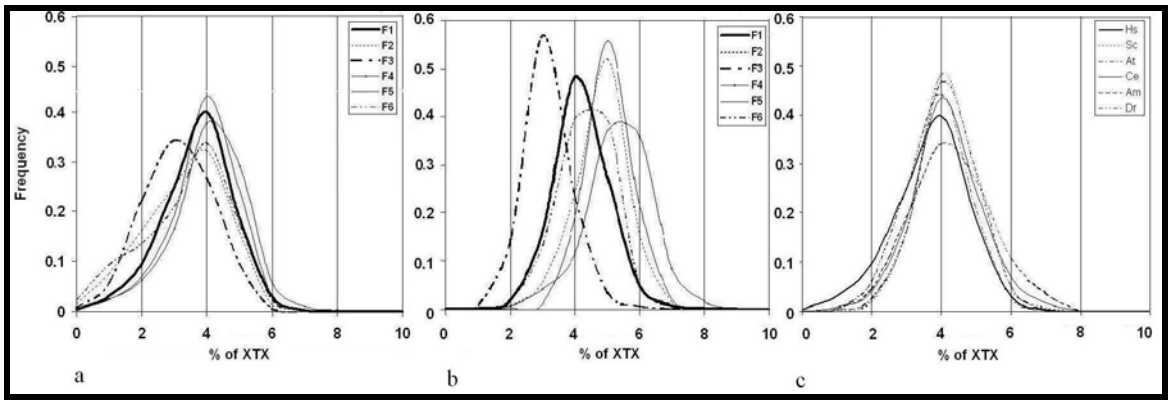

Figure 2: Distribution of mRNA sequences based on XTX in all six frames of human 2(a), yeast 2(b) and frame 1 for different species 2(c). Note that there is no variation in distribution curve in model organisms. (Hs=*H. sapiens*, Am=*A. mellifera*, Dr=*D. rerio*, Ce=*C. elegans*, At=*A. thaliana and* Sc=*S. cerevisiae*).

Discussion:

Distribution of mRNA sequences based on XTX in different frames of human and yeast .3

The distribution of mRNA sequences based on fraction of XTX in all 6 frames of human and yeast are plotted in Figure 2. The labels F1, F2, F3, F4, F5 and F6 indicate the corresponding frames. The frequency in the Y - axis means that fraction of mRNA sequences in the given percentage of XTX. The human and yeast results are shown in Figure 2 because of the two extreme cases of thymine content (data not shown for the distribution in other species). The distributions are normal. All frames show a similar distribution curve with human. The frame 3 shows up a least amount of XTX while frame 4 with highest always. The protein coding frame 1 is coming between these frames. It shows a narrow distribution while frame 4 shows a broad distribution. Frame 1 prefers a fixed amount of XTX while frame 4 has varied amount of XTX. Human mRNA sequences lack thymine in all 6 frames and the distribution profile also same except frames 3 and 5. The human mRNA shows a variation of XTX in the complementary strand while it is so much with fungus and plant. It is observed that increase in thymine fraction in frame 1 proportionally decreases in frame 4. The thymine fraction in frames 1 and 4 is not varying, which reveals that these frames involved in protein coding.

The comparison of distribution profile of different chromosomes of a same species does not show much difference. The total amount of thymine in human mRNA sequences is generally low compared to that of yeast. This is reflected in frames 2 and 6. That is any alteration in thymine content is adjusted in these frames and not in frames 1 and 3.

The variation in thymine content in different frames of mRNA sequences of a given species is calculated. That is the standard deviation is computed for each frames of different species. The variation is high in human compared to fungus or plant. This is evident from the figure 1 that the distribution curve is broad for human and narrow for yeast. The frame 3 of human shows more variation than frame 2 and it is reverse in fungi and plant. Variation observed in frames 1, 2 and 4 also considerable. Heterosexual animals show more variation

in all frames compared to plants and fungi. Frame 1 and 4 prefers to be in definite variation.

Frame 2 and 6 shows a varying amount of thymine in different sequences and different species, as it is third nucleotide of the codons in strand 1 and 2. So during evolution any alteration in amount of thymine is tolerated by adjusting at these frames.

The total amount of thymine in human mRNA sequences is less in general when compared to fungi and plants. But fraction of the total thymine in frame 1 is high in human. These things strongly suggest that the involvement of thymine for synthesis of large hydrophobic residues that stabilize and keep functioning of the proteins. In conclusion the amount of thymine is reduced during evolution that leads to production of different proteins (diseased proteins) that unable to function normally.

As can be seen here in Figure 2, the distribution of mRNA sequences based on thymine amount in frame1 is same in different species though some difference is observed in animals. This suggest that the bases particularly thymine in protein coding mRNA sequences are not randomly distributed but with some probability. The mRNA sequences of animals, in particular the human show a different distribution profile compared to that of a fungus or plant. The distribution based on thymine fraction reveals that the variation is high in heterosexual human compared to fungus and plant. The frame 1 prefers to have 27 XTX out of 100 codons.

The probable amount of XTX in different frames and different species

The probable amount of XTX is calculated in different frames of different species. That is maximum number of mRNA sequences posses a fixed amount of XTX is computed and it is again observed that there is no difference in frame1 values. The frame 4 has higher amount of XTX than frame 1 always. The frame2 shows a variable amount of XTX in different species.

The animals, particularly *H. sapiens* shows a similar amount of XTX in almost all frames. The plant (*A. thaliana*) and fungus (Yeast) show a similar XTX amount in all frames but they differ from animals. The probable values are comparable to the average values only in frame1 but not in other frames. This shows the significance of amount of XTX in frame1. The virus *P. falciparum* shows totally a different amount of XTX in all frames including the frame 1. *V. cholerae* shows an equal amount of XTX in strand 1 and strand 2.

The deviation of thymine values in different frames of different species is not much in frame1 and 3 while frames 2 and 6 showing higher values. This indicates that the number of thymine is adjusted in frames 2 and 6 and not in frames 1 and 3. Frame 4 and frame 5 shows up intermediate variation in standard deviation values. The average and probable amount of XTX in frame1 is not varying which is significant. The XTX variation is significant in frame2 while frame1 and 3 are not. The standard deviation is high in heterosexual human compared to fungus and plant.

It is observed in all cases that frame 3 is having least amount of thymine. The amount of thymine frame1 is fixed and in frame 3 to some extent as compared in all species. The frame 2 shows up an altered amount of thymine confirm the wobble hypothesis. Both average and probable results on amount of thymine in all six frames show almost same result except in frames 1 and 3 of Yeast and *A. thaliana*.

The frame 1 has almost same amount of XTX in all species shown here except virus. Both animals shown here has same amount of XTX in all frames. Also the plant and fungus show similar values in all frames. The virus has different amount of XTX in all frames. In all frames it is observed that the amount of sequences for a given fraction of XTX is less in human. Frames 1 and 2 are showing a similar distribution in all species. The other frames show a different distribution for different species. Frame 2 shows a shift in the fraction of XTX. The shift in the distribution curve of frame 1 shows the shift in the opposite direction in frame 2.

Generally, the total amount of thymine in human protein coding DNA sequences is less. It observed from ratio of XTX in frame 1 and total that the human mRNA sequences utilizes more thymine in the frame 1 from the given total thymine. But it is less utilized in fungi and plants. This proves that frame 1 prefers to have definite amount of XTX for production of stable and active proteins by incorporating large hydrophobic residues at appropriate places. The point is that whether the mRNA sequence has high or less amount of thymine but with definite amount in frame 1. This is essential for production of protein with adequate hydrophobic residues. The distribution of mRNA sequences based on other bases A, G and C does not produce the same results except that the base A is complementary to T and it shows a reverse distribution profile that of T.

Conclusion:

Distribution analysis on the role of thymine in protein coding mRNA sequences of some of the organisms such as *H. sapiens, B. taurus, C. familiaris, G. gallus, D. melanogaster, A. mellifera, D. rerio, C. elegans, A. thaliana, S. cerevisiae and V. cholera* have been studied here. It is concluded that thymine in protein coding sequences are not randomly distributed but with probability. Frame 1 prefers to have a definite amount of thymine. The thymine fraction in frame 4 is also not

Bioinformation by Biomedical Informatics Publishing Group *open access*

www.bioinformation.net **Hypothesis**

varying significantly suggesting the involvement in protein coding. In frames 2 and 6 (third nucleotide of the coding frame in strand 1 and 2 respectively) a variable amount of thymine is observed suggesting that any alteration in amount of thymine is tolerated by adjusting in these frames. Frame 3 prefers to have least amount of thymine. Frame 4 has higher amount of thymine than frame1. The animal, in particular the human mRNA sequences show a different distribution profile compared to that of a fungi or plants. In human the probable amount of thymine in frame 1 is less compared to fungus and plant but fraction of the total thymine in frame 1 is high while it is reverse in plant and fungus. This strongly suggests that the fraction of thymine is essential for production of stable and functional protein with definite amount of hydrophobic elements. The variation in fraction of thymine in DNA sequences of human is very high compared to fungi and plant. The final conclusion is that the amount of thymine is reduced during evolution that leads to production of diseased proteins that unable to function normally.

References:

[01] S. Kolaskar and B. V. B. Reddy, *Nucl. Acid Res.*, 13: 185 (1985) [PMID: 3839071]

[02] B. Issac *et al.*, *Bioinformatics*, 18: 196 (2002) [PMID: 11836230]

[03] J. W. Fickett, *Nucl. Acid Res.*, 10: 5303 (1982) [PMID: 7145702]

[04] T. S. Wong, *et al*., *J Mol Biol*., 355: 858 (2006) [PMID: 16325201]

[05] ftp://ftp.ncbi.nlm.nih.gov/genomes/

Edited by P. Kangueane

Citation: Anandagopu *et al.*, Bioinformation 2(7): 304-307 (2008)

Supplementary material

S. no	Species	mRNA sequences (number)
1	*H. sapiens*	19131
2	*B. taurus*	26501
3	*C. familiaris*	33644
4	*G. gallus*	39218
5	*D. melanogaster*	19545
6	*A. mellifera*	9487
7	*D. rerio*	35695
8	*C. elegans*	10205
9	*A. thaliana*	31711
10	*S. cerevisiae*	5880
11	*V. cholera*	3835

Table 1: Number of mRNA sequences in each species.