International Journal of Bioinformatics Research, ISSN: 0975–3087, Volume 1, Issue 2, 2009, pp-65-69

Codon context analysis of *Methanocaldococcus jannaschii*

Verma D.K.* and Gokhale M.S.**

*[1]Department of Biotechnology and Bioinformatics, Padmashree Dr. D Y Patil University, Navi Mumbai, 400614, India, * Both authors contributed equally to this work

Abstract- Evolution has caused major changes as well as minor changes in the mRNA sequences. Changes in the nucleotides can be result of some selectional pressure leading changes in the codon or due to influences of neighboring codon on selected possible synonymous codon. This research generally analyze the concept of evolutionary constraints on the biased selection of a particular codon-pair present in mRNA to be influenced by the 5'-3' tRNA anti-codon and its neighboring triplet explaining reason for the lesser amount of tRNA in a particular genome as compared to that of the standard genetic code available. The occurrence of one codon next to another has co-evolved with the structure of tRNA and lead to accommodate maximum amount of genetic code to be translated by limited number of tRNA.

Keywords: Codon Context, Evolution, tRNA, mRNA, Methanocaldococcus jannaschii, amino acids

Introduction

An innumerable amount of evolutionary strength shapes the sequences of coding components in mRNA of genomes. Synonymous codon usage and codon-pair are expected under selective pressure since they can primarily influences change in the translational efficiency of the protein to be formed [1-6]. In addition to the translational efficiency the codon pair also influences the accuracy of the process, leading to suppressions (missense as well as nonsense) [7-12] and frame shift errors [13-15]. Synonymous codons are used with bias as proposed by Fiers et al. , Air et al., Grantham et al. We assume that anti-codon and the neighboring codon in the 3' direction of the tRNA can be used to understand about the synonymous pair usage in the mRNA. Peptide bond formation takes place when the mRNA is placed on the A and P sites of the ribosome assembly [16], and the selectively charged aminoacyl t-RNA binds to the mRNA with anticodon leading to the formation of the long amino acid chains (proteins). This binding of the t-RNA through anticodon along with adjacent bases next to anticodon is dependent upon codon pair present in the mRNA which can be of 64*61 combinations (61 since three stop codons cannot be present at previous position). These combinations would require the same amount of the tRNA, which is not seen in organisms.

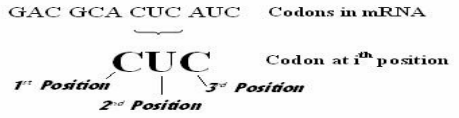

Fig.(1) The adjacent diagram explains the concept of Anticodon-Codon interaction with adjacent Codon pair of mRNA (ith position and i-1th position) to (ith position and i+1th position) of tRNA.

This suggests that some dual codon pairs (Figure 1) combinations of mRNA-tRNA would have advantage in terms of translation efficiency and the driving force for the evolution of codon usage leading to the over expression of same. This leads to large occurrence of a particular dual codon pair of tRNA-mRNA. Statistically this means the data from the genome would be deviating in both directions for the over expressed as well as under expressed codon pair. This dual codon pair is not only biased in the expression but also biased with respect to the position of bases in the codon (Figure 2).

Methods

Genomic sequences of *Methanocaldococcus jannaschii* DSM 2661 were downloaded from GenBank release as *.gb file. The Refseq accession number of the genome downloaded is NC_000909.The entire mRNA sequences i.e. a total of more than 1700 sequences and the tRNA sequences of the genome were extracted through a program written in PERL.

Extraction of data from mRNA:

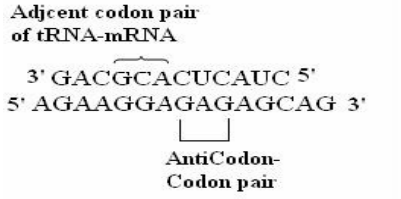

Fig. (2) Above figure explain different position of bases in the Codon at position i.

The frequency of occurrence of a specific codon ($codon_{i-1}|codon_i$) is calculated and a contingency matrix of 61*64 is formed. This calculated contingency matrix is then analyzed for the normality test using SPSS package, suggesting data not to be normally distributed. Also the standard deviation calculated is above mean, suggesting irregular $codon_{i-1}|codon_i$ data distribution. To

extract the data for over expressed and under expressed $codon_{i-1}|codon_i$ pair each case were considered separately.
To calculate frequency of each codon pair following equation was used

$$ki = \sum_{j=1}^{64} Cji \quad i = 1,2,3 \ldots .61 \qquad Equation\ 1$$

Where ki is the frequency of the occurrence of a codon at position i and Cji is the frequency of the occurrence of a codon j given i has occurred.
To calculate sum total of all codon pairs is Equation 2 is used

$$K = \sum_{i=1}^{61} ki \qquad Equation\ 2$$

Where K is total sum of ki
The over-expressed data was extracted using Equation 3

$$X = \{ki | ki \geq 4\%\ of\ K\} \qquad Equation3$$

Where X is the set of codons having a frequency greater than or equal to four percent of K
From X, over-expressed codons at i-1 position was extracted using Equation 4

$$A = \{Cji | Cji \geq 4\%\ of\ ki\ where\ ki \in X\}\ Equation4$$

where A is the set of codons at i-1 position, having a frequency greater than or equal to 4% of ki belonging to X
Here under expressed codons are not selected because value 0 is associated with them which are of no use.

Treatment for the tRNA:
Genomic sequences of *Methanocaldococcus jannaschii* DSM 2661 contain 37 tRNA sequences having 27 in the Positive strand and 10 in the complementary strand. Anticodon position as well as i+1 position in the tRNA sequence was extracted. Contingency tables giving the information regarding the 20 X 61 amino acid to codon were formed. The data of the tRNA was obtained in 3'-5' direction.

Comparison between mRNA and tRNA
The set of the codon from equation 5 is taken as $codon_{i-1}|codon_i$ position pair in the mRNA. These are the sequences which are expressed to a maximum extent in this genome, hence can give the information regarding the highly expressed tRNA. The redundant $codon_{i-1}$ of the mRNA is taken to find the complementary $codon_{i+1}$ position of the tRNA sequences. The compared pair leads to tRNA which is expressed in the larger amount in the genome. Similarly equation 6 is used to extract sequences which are under expressed in the genome. The compared pair leads to tRNA which is under expressed in the genome. After removing the codon pair from the mRNA the normality test was performed again, and observed data had a normal distribution. The standard deviation obtained was also less than that of average.

Results
The extracted data set contains GAU, UUA, GUU, AAA, AUA, AUU, GAA codon that are highly expressed in the mRNA. These codons contribute to approximately 36% in the entire length of the mRNA, suggesting the expression of aspartic acid, Leucine, Valine, Lysine, Isoleucine, glutamic acid in a larger amount. These codons were kept constant and their neighboring codon (codon context) was studied to find the most conserved pair represented in (table 1). Position based chart analysis was also done as shown in the (figure 3).

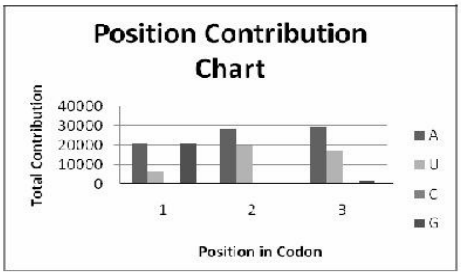

Fig. (3) Position wise contribution chart in over expressed

Position wise analysis of the conserved $codon_{i-1}|codon_i$ showed that complete absence of the cytosine in all the three position of the codon in the $codon_{i-1}$ of the highly expressed codons as derived from equation 5. Adenine contributes to a maximum in the all the position, were as contribution of the guanine in the second position of the codon is none as compared to the first position suggesting the presence of the smaller weight amino acids like alanine and valine to be present in the mRNA. Comparison of the data obtained from the genome of Methanocaldococcus jannaschii for the tRNA and mRNA shown in (table 2) leads to a conclusion that codon position next to the codon-anticodon pair is highly wobbled for the highly expressed pairs. This wobble nature of the tRNA with respect to the mRNA leads to changes in the expression level of the particular codon pair in translation process. Presence of one tRNA to the highly occurring pair in the mRNA also states that it is one of the rate limiting steps

available for the organism in relation to the translation mechanism. This analysis is done for the mRNA to be taken as continuous stretch. So no information associated with respect to particular gene expression could be deduced. The list of the amino acids that are under control during the process of translation in the organism is given in (table 3). Similar analysis of the low expressing codon position were done determining arginine, serine and proline to the amino acids that are least expressed but in terms of the codon pairs but as seen from (table 4) the tRNA pairs associated for the least expressed amino acids are more. There is a striking difference between the under expressed and the over expressed. The number of the tRNA coding for the under expressed is more in number as compared to that of the over expressed, Stating that these amino acids do not require control while being synthesis of the protein.

Discussion

Evolution plays a vital role in the conservation of data which it likes in genome of organism to be present. mRNA genome analysis of Methanocaldococus janaschii remarks the presence of similar group of amino acids like leucine, isoleucine, valine in one group and aspartic acid, glutamic acid in other group. Group expressions in a genome are one of the factors associated with the evolution. Bulkier amino acids like tryptophan, phenylalanine, histidine, arginine are generally avoided, leading to protein being formed in economic rate. The analysis of the genome also suggests, translation as one of the rate limiting steps available during protein synthesis, due to the limited availability of the tRNA recognizing similar type of codon wobble. Highly expressed codons tend to occur in pair which is having some correlation between base composition of the previous codon of mRNA and triplet of bases next to anticodon in the tRNA. Anticodon based conservation is not found in the genome, but the conservation is found within two bases next to anticodon in the tRNA where a preference for adenine is observed. Analysis of the position wise contribution of the previous codons suggests the dominance of adenine and uracil in the second and the third position, giving guanine is the least preferred in the evolution for overall contribution in codon pair evolution within mRNA. It may be possible that dominance of uracil at second and third position of previous codon and high occurrence of adenine within two positions next to anticodon of tRNA is causing some weak association thereby helping more efficient base pairing between mRNA-tRNA and thereby checking the rate of translation of highly expressed codon pairs. This check might not be done in case of least expressed codons since the expression is less as the tRNA anticodon available for least expressing codon are more.

Acknowledgements

We gratefully acknowledge Department of Biotechnology and Bioinformatics, Dr. D Y Patil University, Navi Mumbai for their support.

References

[1] Irwin B., Heck J.D., Hatfield G.W. (1995) *J. Biol.Chem.*, 270, 22801–22806.

[2] Berg O.G., Silva P.J. (1997) *Nucleic Acids* 25(7), 1397-1404.

[3] Fedorov A., Saxonov S., Gilbert W. (2002) *Nucleic Acids* 30(5), 1192-1197.

[4] Boycheva S., Chkodrov G., Ivanov I. (2003) *Bioinformatics* 19(8), 987-998.

[5] Moura G., Pinheiro M., Silva R., Miranda I., Afreixo V., Dias G., Freitas A.,Oliveira J.L., Santos M.A. (2005) *Genome Biol* 6(3), R28.

[6] Moura G., Pinheiro M., Arrais J., Gomes A.C., Carreto L., Freitas A.,Oliveira J.L., Santos M.A. (2007) *PLoS ONE* 2(9), 847.

[7] Murgola E.J., Pagel F.T., Hijazi K.A. (1984) *J Mol Biol*, 175(1), 19-27.

[8] Bossi L., Ruth J.R. (1980) *Nature*, 286(5769), 123-127.

[9] Miller J.H., Albertini A.M. (1983) *J Mol Biol*, 164(1), 59-71.

[10] Kopelowitz J., Hampe C., Goldman R., Reches M., Engelberg-Kulka (1992) *J Mol Biol*, 225(2), 261-269.

[11] Stormo G.D., Schneider T.D., Gold L. (1986) *Nucleic Acids Res*, 14(16), 6661-6679.

[12] Curran J.F., Poole E.S., Tate W.P., Gross B.L. (1995) *Nucleic Acids Res*, 23(20), 4104-4108.

[13] Shah A.A., Giddings M.C., Gesteland R.F., Atkins J.F., Ivanov I.P. (2002) *Bioinformatics* 18, 1046–1053.

[14] Murgola E.J., Pagel F.T., Hijazi K.A. (1984) *J Mol Biol* 175, 19–27.

[15] Tork S., Hatin I., Rousset J.P., Fabret C. (2004) *Nucleic Acids Res* 32, 415–421.

[16] Rheinberger H.J., Sternbach H., Nierhaus K.H. (1981) *Proc Natl Acad Sci U S A.* 1981 Sep; 78(9), 5310-5314.

[17] Yi Lu and Stephen Freeland (2006) *Genome Biology* 7, 102

[18] Margaret E. Saks and John S. Conery. (2007) *RNA* 13, 651-660

Table 1: Most expressed on the basis of 4% of the total codon$_i$ position contribution					
Most expressed codon$_{i-1}$ on the basis of 4% of total					
UUA|GAU	GAU|UUA	GAU|GUU	UUA|AAA	GAU|AUU	GAU|GAA
GUU|GAU	GUU|UUA	GUU|GUU	AAA|AAA	UUA|AUU	UUA|GAA
AAA|GAU	AAA|UUA	AAA|GUU	AUA|AAA	AAA|AUU	GUU|GAA
AUU|GAU	AUU|UUA	AUU|GUU	GAA|AAA	GAG|AUU	AAA|GAA
GAA|GAU		GAA|GUU		GAA|AUU	AUU|GAA
			AAA|AUA		GAA|GAA

Table 2: Observed and Extracted data comparison			
Codon from tRNA	**Codon from mRNA**	**Codon from tRNA**	**Codon from mRNA**
UCA**CUG**	UUA|GAU	UAG**AAU**	GAU|UUA
	GUU|GAU		GUU|UUA
	AAA|GAU		AAA|UUA
	AUU|GAU		AUU|UUA
	GAA|GAU		
Codon from tRNA	**Codon from mRNA**	**Codon from tRNA**	**Codon from mRNA**
CAA**UUU**	UUA|AAA	CAA**UAG**	GAU|AUU
	AAA|AAA		UUA|AUU
	AUA|AAA		AAA|AUU
	GAA|AAA		GAG|AUU
			GAA|AUU
Codon from tRNA	**Codon from mRNA**		
CAA**UAG**	AAA|AUA		
Codon from tRNA	**Codon from mRNA**	**Codon from tRNA**	**Codon from mRNA**
	GAU|GUU	GAG**CUU**	GAU|GAA
CGG**CAC**	GUU|GUU		UUA|GAA
GAA**CAC**	AAA|GUU		GUU|GAA
GAA**CAU**	AUU|GUU		AAA|GAA
	GAA|GUU		AUU|GAA
			GAA|GAA

Table 3 : Amino acid to codon table for over expressed

Amino acid	Codon in tRNA
Aspartic Acid	UCA**CUG**
Leucine	UAG**AAU**
Valine	CGG**CAC**
	GAA**CAC**
	GAA**CAU**
Lysine	CAA**UUU**
Isoluecine	CAA**UAG**
Isoleucine	CAA**UAG**
Glutamic Acid	GAG**CUU**

Table 4: Amino acid to codon table for under expressed

Amino acid	Arginine	Serine	Proline
AntiCodon in tRNA With triplet at 3' end	GAG**GCG**	UAG**AGU**	UAG**GGU**
	UAG**GCU**	UAG**AGG**	GGG**GGG**
		CCU**AGA**	