Gene 630 (2017) 13–20

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Research paper

Factors affecting the codon usage bias of SRY gene across mammals

Supriyo Chakraborty*, Arif Uddin*, Monisha Nath Choudhury

Assam University, Biotechnology, Dargakona, 788011 Silchar, Assam, India

ARTICLE INFO

Keywords:
Codon usage
Natural selection
Mutation pressure
SRY gene

ABSTRACT

Codon usage bias (CUB) is extensively found in a wide variety genomes and it is mostly affected by mutation pressure and natural selection. Analysis of CUB helps in studying the evolutionary features of a genome. The SRY gene plays an important role in male reproductive organ and a good candidate to study the evolutionary forces, since little work was reported earlier on this gene. We used bioinformatic methods to analyze the protein-coding sequences of SRY gene in 172 different mammalian species to understand the patterns of codon usage and the evolutionary forces acting on it. We found that the codon bias of SRY gene varies widely across mammals. Relative synonymous codon usage (RSCU) value revealed that the codons such as TCG, CCG, CAT, ATT, ACT, GCT, GTT, GCG, GGG and GGT were over-represented. Correspondence analysis indicated that the distribution of codons was more close to the axes indicating that compositional constraints might correlate to codon bias. *Z*-score analysis on RSCU values of codons identified a set of 11 codons *viz.* TCT, TTT, CTA, CTC, TAT, CAG, CGT, ATA, ACC, AAT and GTA which differed significantly at $p < 0.01$ between 5% high and low gene expression datasets. Further, it was evident from the neutrality plot that GC12 was influenced by both mutation pressure and natural selection. From the study we concluded that natural selection played a dominant role, but mutational pressure played a minor role in the codon usage pattern of SRY gene across mammals.

1. Introduction

In standard genetic code it is found that 61 sense codons encode 20 standard amino acids. Except methionine and tryptophan, each amino acid is encoded by more than one codon and these are called synonymous codons for that amino acid. The synonymous codons are used non-randomly *i.e.* some codons are used more frequently than others for translation of mRNA transcript to protein and this unequal usage of synonymous codons is known as codon usage bias (CUB) (Bulmer, 1988). Generally, codon bias is found in genes which are highly expressed in some organisms. The difference of preferred codons among genes provides differential efficiency along with accuracy in the translation of genes (Rocha, 2004). CUB is mainly influenced by mutational pressure and natural selection due to translation. Mutational biases occur when there is very high or low frequency of G or C nucleotide in the third position of open reading frames (ORF) while natural selection occurs due to translation when genes choose codon usage for translation of mRNA to protein so as to facilitate rapid translation. Natural selection under translational process also influences codon usage bias. Apart from mutational pressure and natural selection, other factors such as expression level of gene, compositional constraints, hydrophobicity and aromaticity of encoded protein, recombination rates and RNA stability, gene length, are known to influence codon bias (Kumwenda, 2013) (Duret and Mouchiroud, 1999; Fullerton et al., 2001). Previous investigations suggested that pattern of codon usage bias differ within and between genomes (Shields and Sharp, 1987).

The study of codon usage bias gained the attention of scientists with the beginning of whole genome sequencing of different organisms and with the availability of their DNA/RNA sequence data in National Centre for Biotechnology Information, Bethesda, USA (Gardner et al., 2002). Gene wise investigations of codon bias patterns, their causes and identification of evolutionary forces that influence their evolution are of major importance to studies of molecular biology, genetics and the findings from such studies pave the way to understand evolution at a finer scale.

CUB has many important applications in determining the origins of species (Kanaya et al., 1999), design of primers, heterologous gene expression and in the prediction of gene expression level (Yang et al., 2014).

However, earlier reports on CUB have focused on model organisms like *Caenorhabditis* (Duret and Mouchiroud, 1999), *Drosophila* (Powell and Moriyama, 1997), many microorganisms such as *Entamoeba histolytica* (Ghosh et al., 2000), *Borrelia burgdorferi* (McInerney, 1998), insects (Behura and Severson, 2012), platyhelminths (Mazumder et al.,

Abbreviations: ENC, effective number of codons; RSCU, relative synonymous codon usage; CBI, Codon bias Index; CUB, Codon usage bias

* Corresponding authors.
E-mail addresses: supriyoch_2008@rediffmail.com (S. Chakraborty), arif.uddin29@gmail.com (A. Uddin).

http://dx.doi.org/10.1016/j.gene.2017.08.003
Received 7 June 2017; Received in revised form 20 July 2017; Accepted 3 August 2017
Available online 04 August 2017
0378-1119/

2015), nemathelminths (Mazumder et al., 2016) and mitochondrial genes in pisces, aves and mammals (Uddin and Chakraborty, 2016). In contrast, a few studies were carried out on specific genes like TP53 gene (Mazumder and Chakraborty, 2015), CYB gene (Uddin et al., 2016), ND1 gene (Uddin et al., 2016), ND2 gene (Uddin et al., 2015), SPANX gene (Choudhury and Chakraborty, 2015), testis specific gene (Choudhury and Chakraborty, 2016) but little or no work was reported on codon usage pattern for SRY gene which plays a very important role in male reproductive organs (Choudhury et al., 2017).

The SRY gene encodes the SRY protein, which is basically a DNA binding protein. It is also known as testis determining factor (TDF) that is responsible for commencement of male sex determination (Berta et al., 1990). It is an intron less sex determining gene located in Y chromosome of placental mammals and marsupials (Wallis et al., 2008). The mutations of SRY gene affect an individual's phenotypic and genotypic traits related to sex (Wagner et al., 1994). The TDF is a member of the SOX (SRY like box) gene family, which when complexes with the SF1 protein, acts as a transcription factor (TF). This TF can unregulate SOX9, another transcription factor (Kashimada and Koopman, 2010). The expression of this transcription factor leads to the development of primary sex cords, which afterwards develops into seminiferous tubules. These cords form the undifferentiated gonad, which turns it into a testis. The Leydig cells of the testis then start secreting testosterone, but the Sertoli cells produce antiMüllerian hormone (Finucane et al., 2000). Normally the function of SRY gene takes place 6–8 weeks after foetus formation which stops the female anatomical structural growth in males (O'Day, 2012).

The objective of this study is to perform a analysis of compositional dynamics, codon usage bias and evolutionary forces using various codon usage bias indices such as ENC and RSCU of SRY gene across 172 different mammalian species. This study elucidates novel insight into the pattern of codon usage bias of SRY gene that would help for better exploration of molecular biology, genetics, mechanisms and distribution of codons as well as help in understanding the evolutionary biology of the SRY gene among the mammalian species. This also sheds light on the evolutionary forces and their magnitude which influences the pattern of codon usage bias SRY gene.

2. Materials and Methodology

2.1. Sequence data retrieval

The coding sequences (cds) of SRY gene from 172 different mammalian species were retrieved from the National Centre for Biotechnology Information (http://www.ncbi.nlm.nih.gov/Genbank/). The S1 contains different species with accession number and gene length. The coding sequences were analyzed using a perl script developed by corresponding author (SC).

2.2. Compositional properties

The overall nucleotide composition (A%, C%, T% and G%) and nucleotide composition at the third codon position (A3%, C3%, T3% and G3%) of SRY gene in each species were analyzed using the perl script. The GC, GC1, GC2 and GC3 contents in the analysis refer to the overall G + C content in the gene sequence, GC content at the 1st, 2nd and 3rd position of synonymous codons (excluding codons for *met, trp* and termination codons).

2.3. Effective number of codon (ENC)

The effective number of codons (ENC) of a cds is a measure that quantifies the degree of codon usage bias. It is an excellent index to measure the codon usage bias in genes. The lower limit of ENC value is 20, indicating high codon usage bias, while the higher limit of ENC is 61, signifying equally likely usage of all synonymous codons (Wright, 1990). Numerical value of ENC has an inverse relationship with the degree of codon bias.

ENC was estimated as:

$$ENC = 2 + \frac{9}{F_2} + \frac{1}{F_3} + \frac{5}{F_4} + \frac{3}{F_6}$$

where, F_k (k = 2, 3, 4 or 6) is the average of the F_k values for k-fold degenerate amino acids. The F value denotes the probability that two randomly chosen codons for an amino acid with two codons are identical.

2.4. Relative synonymous codon usage (RSCU)

It refers to the ratio of the usage frequency of one specific codon to the usage frequency of its corresponding synonymous codons for the same amino acid. If the RSCU value of a codon is 1, it means there is no codon bias and the codon is used equally with other synonymous codons. Codons with an RSCU value > 1 display strong codon bias while codons with an RSCU value < 1 show negative codon bias which are used less frequently than other codons. If the RSCU value > 1.6, it indicates that the codon is over-represented whereas if the RSCU value < 0.6, it signifies that the codon is under-represented in the coding sequence (Shields and Sharp, 1987).

The RSCU was calculated as:

$$RSCU = \frac{g_{ab}}{\sum_b^{na} g_{ab}} n_a$$

where, g_{ab} is the observed number of the ath codon for the bth amino acid which has n_a kinds of synonymous codons.

2.5. Correspondence analysis (COA)

Correspondence analysis (COA) is widely used to investigate the variation in the pattern of synonymous codon usage among genes. COA is a multivariate statistical method in which codon usage data of 59 codons were plotted in a multidimensional space of 59 axes. This plot is used to reveal the axes that correspond to the most important factors influencing the codon usage variation among genes (Greenacre, 1984).

2.6. PR2-bias plot analysis

The Parity Rule 2 (PR2) bias is analyzed by the value of AT-bias [A/(A + T)] as the ordinate and GC-bias [G/(G + C)] as the abscissa [31]. In this plot, the centre of the plot is 0.5, where A = T and G = C (PR2), with no biases between the two complementary strands of DNA. The PR2 bias plots are mostly informative when [A/(A + T) and [G/(G + C] values are plotted for 3rd codon position of 4-codon amino acids (Sueoka, 1988).

2.7. Neutrality plot

The neutrality plot analysis *i.e.* regression of GC12 on GC3 is used to explore the degree to which mutational pressure influences codon usage bias in comparison to natural selection. This analysis was done using GC3 as abscissa and GC12 as the ordinate for SRY gene across mammals. If the regression line falls near the diagonal (*i.e.* slope close to 1) it indicates weak external selection pressures influencing the codon usage bias and the dominant role of mutation pressure while regression line deviating from the diagonal (*i.e.* slope tends to 0) signifies the dominant role of natural selection on the codon usage bias (Sueoka, 1988).

2.8. High and low gene set for optimal codon

The SCUO (synonymous codon usage order) is an information-based measure for the non-random usage of synonymous codons of an amino acid and its value ranges from 0 (no bias) to 1 (highly biased). It is also