DNA Research **15**, 347–356, (2008) doi:10.1093/dnares/dsn023

Differential Selective Constraints Shaping Codon Usage Pattern of Housekeeping and Tissue-specific Homologous Genes of Rice and Arabidopsis

Pamela Mukhopadhyay[1], Surajit Basak[2], and Tapash Chandra Ghosh[1,*]

Bioinformatics Centre, Bose Institute, P 1/12, C.I.T. Scheme VII M, Kolkata 700 054, India[1] and Biomedical Informatics Centre, National Institute of Cholera and Enteric Diseases, P- 33, CIT Scheme XM, Kolkata 700 010, India[2]

(Received 17 March 2008; accepted 2 September 2008; published online 30 September 2008)

Abstract

Intra-genomic variation between housekeeping and tissue-specific genes has always been a study of interest in higher eukaryotes. To-date, however, no such investigation has been done in plants. Availability of whole genome expression data for both rice and Arabidopsis has made it possible to examine the evolutionary forces in shaping codon usage pattern in both housekeeping and tissue-specific genes in plants. In the present work, we have taken 4065 rice–Arabidopsis homologous gene pairs to study evolutionary forces responsible for codon usage divergence between housekeeping and tissue-specific genes. In both rice and Arabidopsis, it is mutational bias that regulates error minimization in highly expressed genes of both housekeeping and tissue-specific genes. Our results show that, in comparison to tissue-specific genes, housekeeping genes are under strong selective constraint in plants. However, in tissue-specific genes, lowly expressed genes are under stronger selective constraint compared with highly expressed genes. We demonstrated that constraint acting on mRNA secondary structure is responsible for modulating codon usage variations in rice tissue-specific genes. Thus, different evolutionary forces must underline the evolution of synonymous codon usage of highly expressed genes of housekeeping and tissue-specific genes in rice and Arabidopsis.

Key words: error minimization; housekeeping; mRNA folding energy; synonymous rates; tissue specific; tRNA copy number

1. Introduction

The completed genome sequences of rice[1] (*Oryza sativa*) and Arabidopsis[2] (*Arabidopsis thaliana*) constitute a valuable resource for comparative genomic analysis, as they are representatives of the two major evolutionary lineages within the angiosperms: the monocotyledons and the dicotyledons. The divergence in codon usage patterns between rice and Arabidopsis genes has occurred since the evolutionary divergence of the dicots and monocots ~200 million years (My) ago, with increment in GC content of some rice genes.[3,4] The large scale variation in DNA base composition due to increment of GC revealed two gene classes, namely GC-rich and GC-poor in monocots, but not in dicots.[5–8] It is estimated that codon usage variation in monocots is mainly determined by spatial arrangement of genomic G + C-content, i.e. the isochore structure similar to mammals.[9] The biased gene distribution in the rice genome raised a question about the distribution of tissue-specific and widely expressed genes according to the GC level of the isochores. Several studies indicated that the distribution of widely expressed genes in human is not correlated with GC levels of isochores.[10–13]

Edited by Kenta Nakai

* To whom correspondence should be addressed. Fax. +91 33-2355-3886. E-mail: tapash@boseinst.ernet.in

However, Lercher et al.[14] reported that there is a strong correlation between gene expression breadth and GC content in human, suggesting that there might be selective pressure favoring the concentration of housekeeping genes in GC-rich isochors. Evolutionary studies on housekeeping and tissue-specific genes in mammalian genome have recently gained much more interest.[15–18] Working on codon usage of tissue-specific genes in human, interestingly, Plotkin et al.[19] reported that there is a significant difference in synonymous codon usage between genes specifically expressed in different human tissues. The results suggest that selective constraint acts on synonymous codon usage to optimize translation by adapting to the pool of tRNAs available in each tissue for tissue-specific genes in human.[19] However, Semon et al.[20] by analyzing 2126 human tissue-specific genes expressed in 18 libraries demonstrated that there is no evidence for tissue-specific adaptation of synonymous codon usage in human.

Conversely, all the previous studies on housekeeping and tissue-specific genes have been done on human genome. Rice which is heterogeneous in base composition similar to human has not been explored till date. Rice–Arabidopsis pair is a well-known model to study codon usage divergence in plants.[4,21] Availability of whole genome expression data for both rice and Arabidopsis has made it possible to examine the pattern of evolutionary forces shaping codon usage in housekeeping and tissue-specific genes of these two plants. In the present study, we have traced the pattern of evolutionary forces shaping codon usage in both housekeeping and tissue-specific genes of rice and Arabidopsis and discussed the presence of contrasting selective constraint affecting the evolution of these sets of genes.

2. Materials and methods

2.1. Sequence data

The genomes of rice and Arabidopsis were downloaded, respectively, from RiceGAAS Rice Genome Automated Annotation System ftp://ftp.dna.affrc.go.jp/pub/RiceGAAS/current/ and Arabidopsis Information Resource (TAIR) http://www.arabidopsis.org/. All sequences having <100 codons were ignored from our data set. Also, genes containing internal stop codons were removed and thus data set comprising a total of 18 658 rice genes was taken for further analysis.

Homologous genes between rice and Arabidopsis genomes were identified using gapped BLASTP searches using cut-off expects of 10.0×10^{-6}.[22] Pairs of coding sequences which have at least 30% amino acids positives and overlaps over at least 80% of their length were retained for the analysis. The maximum gap size allowed between a pair of sequence is 5%. Owing to presence of much multi-copy genes both in Arabidopsis and rice, some sequences from one species showed high levels of sequence similarity with more than one sequence from the other species. In those cases, the sequence pairs that produced higher degree of sequence similarity were retained.[23] We also eliminated pseudo genes and mitochondrial protein from the homologous gene set. Finally, our data set consists of 4065 homologous gene pairs (Supplementary Table S1 contains rice–Arabidopsis homologous genes pairs).

2.2. Expression profile

The public domain MPSS (massively parallel signature sequencing) expression data for rice[24] (http://mpss.udel.edu/rice/) and Arabidopsis[25] (http://mpss.udel.edu/at/) present more accurate estimation of gene transcript levels and are easily accessible.[25] The expression level of a gene expressed in a single library is estimated by counting the number of individual 17-base signature sequences representing each gene.[26] It should be noted that current MPSS data set for rice is based on the TIGR rice genome annotation. We retrieved expression level of individual rice genes with RiceGAAS ID using Rice MPSS: Query by Sequence tool that basically extract all possible tags from the sequence and compare them against their database. The expression levels of a gene expressed in different expression libraries were estimated by calculating average expression values in all libraries considered (Supplementary Tables S2 and S3 contain library information). We sorted the expression values in each library in an ascending order, and then divided them into five groups, each containing 20% of the population.[26] Individual genes were assigned an expression rank from 1 (low expression) to 5 (high expression) according to the increase in average expression level.

Tissue specificity of a gene is measured by using tissue specificity index τ.[27,28] The τ of gene i is defined by

$$\tau_{\mathrm{H}} = \frac{\sum_{j=1}^{n_{\mathrm{H}}} \left(1 - [\log_2 S_{\mathrm{H}}(i,j) / \log_2 S_{\mathrm{H}}(i,\max)]\right)}{n_{\mathrm{H}} - 1},$$

where n_{H} is the number of tissues examined and $S_{\mathrm{H}}(i, \max)$ is the highest expression of gene i across the n_{H} tissues. The τ value ranges from 0 to 1, with higher values indicating higher variations in expressional level across tissues or higher tissue specificities. If a gene has expression in only one tissue, τ approaches 1. In contrast, if a gene is equally expressed in all tissues, $\tau = 0$.

We assigned housekeeping and tissue-specific genes by sorting our data set (4065 rice–Arabidopsis

homologous genes) according to increase in τ value and taking out genes from extreme 20% of population from both ends. Using the above criteria, we obtained 787 housekeeping and 770 tissue-specific genes. All our analysis were performed using 787 housekeeping and 770 tissue-specific genes of rice with its corresponding counterpart in Arabidopsis (Supplementary Tables S4 and S5 contain rice–Arabidopsis housekeeping and tissue-specific homologous gene-pairs).

2.3. *Sequence analysis*

Pair-wise synonymous (Ks) and non-synonymous (Ka) distance between the homologous genes of rice and Arabidopsis was calculated by using the method of Yang and Neilsen.[29]

The genetic robustness at codon level has been measured using CUB available at http://users.ox.ac.uk/~zool0643/codon/CUB.html.[30] According to this method proposed by Archetti, we have measured dissimilarity (D_{AA/AA^*}) between original (AA) and mutant amino acid (AA*) for each synonymous codon based on the McLachlan's matrix of chemical similarity.[31] Dissimilarity of a single amino acid (AA) is given by: $D_{AA/AA^*} = \omega_{AA/AA} - \omega_{AA/AA^*}$, where $\omega_{AA/AA}$ is the similarity of the amino acid AA to itself and ω_{AA/AA^*} is the similarity of AA to the mutant amino acid AA* obtained after an error at one of the positions of the original codon. Since $\omega_{AA/AA} > \omega_{AA/AA^*}$ for every amino acid, D_{AA/AA^*} is always positive, and since there are three possible mutants for each position, there are nine possible measures of D_{AA/AA^*} for each codon, corresponding to nine possible mutant codons. Their mean value is taken as a measure of distance (dissimilarity) between the original codon and its possible mutants. This mean value of dissimilarity is the measure of mean distance (MD) for each codon to its possible mutants. To calculate the degree of error minimization of a coding sequence, the correlation between the MD values and the corresponding relative synonymous codon usage (RSCU) is calculated for each synonymous family. If N is the number of degenerate synonymous codon families on which the correlation is calculated, and R is the sum of the correlations, the degree of error minimization is measured by $R_N = R/N$ (R_N ranging between -1 and $+1$). The R_N measures genetic robustness with the assumption that all the amino acids are weighted equally, irrespective of their frequency on the protein. If the value of each correlation is weighted (multiplied) by the frequency of the corresponding amino acid, then the measure is denoted by wR_N. Since MD is a measure of dissimilarity, the lower the value of R_N and wR_N, the higher the degree of error minimization.

The Zipfold program was used to predict free-folding energies for each native mRNA sequence available at http://dinamelt.bioinfo.rpi.edu/zipfold.php.

The transfer RNA gene copy number necessary to determine the major codons[32] for each amino acid in rice were taken from Xiyin et al.[33] and tRNA copy number for Arabidopsis was taken from http://lowelab.ucsc.edu/GtRNAdb/Athal/.

The Student's *t*-test was used to evaluate the significance of all the pair-wise differences. The statistical tests were performed using the SPSS (13.0) package.

3. Results and discussion

3.1. *Influence of expression level in modulating synonymous substitution rates for both housekeeping and tissue-specific genes in rice*

Analysis of synonymous substitution patterns (Ks) between rice and Arabidopsis homologous genes pairs for both housekeeping and tissue-specific classes reveals that housekeeping genes are under stronger selective constraint as observed from their significantly lower average synonymous substitution rates ($Ks = 3.27$) ($P < 0.001$) when compared with tissue-specific genes ($Ks = 3.45$). Similar trend in evolutionary rates have been observed in earlier studies on mammalian genome.[15–17] It has already been demonstrated that housekeeping and tissue-specific genes comprise of both highly and lowly expressed genes.[18] In order to investigate the influence of expression level in modulating synonymous substitution rates of housekeeping and tissue-specific genes in rice, we measured synonymous substitution rates for highly and lowly expressed genes of both housekeeping and tissue-specific classes (Table 1). From the Table 1, it is obvious that synonymous substitution rate of highly expressed housekeeping genes ($Ks = 3.12$) is significantly ($P < 0.001$) lower than that of highly expressed tissue-specific genes ($Ks = 3.74$). In contrast, there is no significant difference in average synonymous substitution rate between lowly expressed housekeeping ($Ks = 3.34$) and lowly expressed tissue-specific genes ($Ks = 3.41$) (Table 1). The results imply that in rice genome selective constraint shaping synonymous codon usage of highly expressed genes varies depending on whether they belong to housekeeping or tissue-specific genes. Non-significant difference in average synonymous substitution rate between lowly expressed housekeeping and tissue-specific genes suggest that lowly expressed genes have been conserved during divergence between rice and Arabidopsis. However, while comparing synonymous substitution rates between highly and lowly expressed tissue-specific and housekeeping genes, an unusual trend have been observed.

Table 1. Average values of synonymous substitution rates for housekeeping and tissue-specific classes of genes in highly expressed (HEG) and lowly expressed genes (LEG) of rice

	Housekeeping	Tissue specific	Level of significance (b)
HEG	3.12	3.74	$P < 0.001$
LEG	3.34	3.41	NS
Level of significance (a)	$P < 0.05$	$P < 0.005$	

Level of significance (a) indicates significance of the difference between highly (HEG) and lowly (LEG) expressed housekeeping and tissue-specific genes of rice.
Level of significance (b) indicates significance of the difference between highly (HEG) expressed housekeeping and tissue-specific genes of rice and lowly (LEG) expressed housekeeping and tissue-specific genes of rice.
NS indicates not-significant.

In housekeeping genes (Table 1), we observed significantly lowered synonymous substitution rate in highly expressed genes (Ks = 3.12) ($P < 0.05$) (number of genes = 209) than lowly expressed genes (Ks = 3.34) (number of genes = 203). Interestingly, in tissue-specific genes of rice (Table 1), the average synonymous substitution rates were significantly lower in lowly expressed genes (Ks = 3.41) ($P < 0.005$) (number of genes = 512) when compared with highly expressed genes (Ks = 3.74) (number of genes = 99). It has been shown in previous studies that the synonymous substitution rate between *Escherichia coli* and *Salmonella typhimurium* is lower in highly than in weakly expressed genes, and it has been suggested that this is due to stronger selection for translational efficiency in highly expressed genes.[34] Recently, Drummond et al.,[35] working on yeast, demonstrated that expression level governs the rate of synonymous substitution and protein sequence evolution. In rice tissue-specific genes, our data suggest that high expression does not necessarily lead to lower synonymous substitution rates when compared with low expression. However, this also prompts us to explore relationship between expression level and translation selection for both housekeeping and tissue-specific genes in plants. Possibly, there may be some other selective force determining the synonymous substitution rate of highly expressed tissue-specific genes in rice.

3.2. *Co-adaptation of synonymous codon usage with the tRNA pool of housekeeping and tissue-specific homologous genes in rice and in Arabidopsis*

In an attempt to investigate the nature of selective constraint shaping synonymous codon usage of housekeeping and tissue-specific genes, we analyzed preferred codons in both the gene classes of rice (Table 2) and Arabidopsis (Table 3). Preferred codons are those that generally correspond to the most abundant tRNA species and they provide fitness benefits to highly expressed genes by enhancing translational efficiency.[36] The co-adaptation of tRNA content and codon usage for the optimal translation of highly expressed genes is well known in *Caenorhabditis elegans*.[37] To test translational selection in rice and Arabidopsis genome, we have identified those codons in both housekeeping and tissue-specific gene classes whose RSCU values are significantly higher in highly expressed genes than lowly expressed genes. We then investigated the correspondence between codon preferences in highly expressed genes and tRNA gene copy number in both rice and Arabidopsis. We obtained ten preferred codons in both housekeeping and tissue-specific gene classes (Table 2) in rice. We even considered revised wobble rules for eukaryotic genomes to estimate preferred codons in highly expressed housekeeping and tissue-specific genes.[38] These rules assume that GNN tRNAs pair with both C-ending and U-ending codons, whereas ANN tRNA genes are modified to inosine and decode both U-ending and G-ending codons. Following revised wobble rule, we observed 14 preferred codons in housekeeping rice genes. Similarly, in tissue-specific rice genes, there are 16 preferred codons that correspond to most abundant tRNA copy number. Our result indicates that translational selection driven by tRNA copy number to optimize synonymous codon usage of highly expressed genes equally influences both housekeeping and tissue-specific genes in rice which does not corroborate with unexpected lowering (Table 2) of synonymous substitution rates in lowly expressed tissue-specific genes. Same analysis was performed in Arabidopsis and it has been observed that in housekeeping genes, there are 10 codons that correspond to most abundant tRNA copy number, whereas in tissue-specific genes, there are only five codons that show perfect match with most abundant tRNA copy number (Table 3). However, after following revised wobble rules for eukaryotic genomes,[38] we obtained 17 codons in housekeeping genes that correspond to most abundant tRNA copy number, whereas in tissue-specific class, we observed only eight preferred codons that correspond to most abundant tRNA copy number (Table 3). Therefore, in Arabidopsis translational selection driven by tRNA copy number to optimize synonymous codon usage of highly expressed genes has a greater influence in housekeeping Arabidopsis genes.

3.3. *Selective constraint acting on mRNA secondary structure is responsible for regulating synonymous substitution rates in rice tissue-specific genes*

It has already been demonstrated that there is a selection for local RNA secondary structures in

Table 2. RSCU values of highly expressed (HEG) and lowly expressed (LEG) housekeeping and tissue-specific genes in rice

AA	Codons	RSCU (HEG)	RSCU (LEG)	tRNA copy number of *Oryza sativa*	AA	Codons	RSCU (HEG)	RSCU (LEG)	tRNA copy number of *Oryza sativa*
Phe	TTT	0.78 (0.57)	0.94 (0.72)	0	Ala	GCT	1.12 (0.63)	1.09 (0.76)	25
	TTC*	1.22 (1.43)	1.06 (1.28)	15		**GCC^{T}↗**	1.17 (1.43)	1.13 (1.34)	0
Tyr	TAT	0.81 (0.59)	1 (0.76)	0		GCA	0.82 (0.54)	0.95 (0.72)	11
	TAC*	1.19 (1.41)	1 (1.24)	16		GCG	0.88 (1.4)	0.84 (1.17)	13
His	CAT	0.96 (0.74)	1.04 (0.91)	0	Gly	GGT	1.01 (0.69)	1.05 (0.73)	0
	CAC*	1.04 (1.26)	0.96 (1.09)	11		**GGC***	1.3 (1.83)	1.05 (1.63)	24
Asn	AAT	0.88 (0.84)	1.14 (0.89)	0		GGA	0.86 (0.69)	0.95 (0.78)	13
	AAC*	1.12 (1.16)	0.86 (1.11)	14		GGG	0.82 (0.79)	0.95 (0.86)	8
Asp	GAT	1.05 (0.84)	1.18 (0.92)	0	Leu	TTA	0.37 (0.38)	0.55 (0.41)	7
	GAC*	0.95 (1.16)	0.82 (1.08)	28		TTG	0.98 (0.88)	1.22 (0.97)	9
Cys	TGT	0.69 (0.39)	0.98 (0.66)	0		CTT	1.3 (0.75)	1.28 (0.95)	19
	TGC*	1.31 (1.61)	1.02 (1.34)	10		**CTC*↗**	1.57 (1.94)	1.34 (1.74)	0
Gln	**CAA^{T}**	0.7 (0.87)	0.85 (0.78)	16		CTA	0.41 (0.49)	0.56 (0.51)	8
	CAG	1.3 (1.13)	1.15 (1.22)	13		CTG	1.37 (1.56)	1.05 (1.43)	6
Lys	AAA	0.56 (0.55)	0.79 (0.67)	10	Ser	TCT	1.14 (0.73)	1.21 (0.92)	17
	AAG*	1.44 (1.45)	1.21 (1.33)	22		TCC	1.2 (1.24)	1.17 (1.23)	0
Glu	GAA	0.7 (0.64)	0.83 (0.7)	15		TCA	1.06 (0.68)	1.19 (0.94)	10
	GAG*	1.3 (1.36)	1.17 (1.3)	29		TCG	0.76 (1.18)	0.64 (0.94)	7
Val	GTT	1.21 (0.7)	1.26 (0.87)	21		AGT	0.75 (0.84)	0.84 (0.68)	0
	GTC^{H}↗	1.09 (1.24)	0.92 (1.23)	0		AGC	1.1 (1.34)	0.95 (1.28)	13
	GTA	0.39 (0.32)	0.56 (0.39)	4	Arg	**CGT^{H}**	0.8 (0.61)	0.63 (0.56)	16
	GTG	1.32 (1.74)	1.27 (1.52)	10		**CGC^{T}↗**	1.37 (1.71)	1.26 (1.42)	0
Pro	CCT	1.16 (0.68)	1.14 (0.89)	16		CGA	0.39 (0.34)	0.48 (0.49)	4
	CCC^{T}↗	0.83 (0.98)	0.86 (0.8)	0		CGG	0.88 (1.05)	0.94 (1.12)	7
	CCA	1.1 (0.78)	1.16 (0.99)	11		AGA	0.96 (0.65)	1.19 (0.96)	9
	CCG	0.91 (1.56)	0.83 (1.32)	10		AGG	1.6 (1.63)	1.51 (1.45)	10
Thr	ACT	1.09 (0.79)	1.12 (0.84)	9	Ile	ATT	1.12 (0.76)	1.24 (0.98)	23
	ACC*↗	1.26 (1.44)	0.96 (1.23)	0		**ATC*↗**	1.32 (1.8)	1 (1.39)	0
	ACA	1.05 (0.7)	1.38 (0.96)	8		ATA	0.56 (0.45)	0.76 (0.63)	6
	ACG	0.61 (1.07)	0.54 (0.98)	0					

RSCU values within parenthesis represent tissue-specific genes of rice, and the values outside represent housekeeping rice genes. Arrows indicate the correspondence between codon and their isoaccepting tRNA based on revised wobble rules. Codons marked with asterisk hold a perfect correspondence with most abundant tRNA gene copy number in both housekeeping and tissue-specific genes. Codons marked with superscript H shows higher preference in highly expressed housekeeping genes. Codons marked with superscript T shows higher preference in highly expressed tissue-specific genes.

coding regions and this nucleic acid structure resembles the folding profiles of the coded proteins.[39] Further, it has been observed in *E. coli* the decrease of the stability of mRNA structure contributes to the increase of mRNA expression[40] suggesting possible relationships between synonymous codon usage and presence of some constraints upon mRNA secondary structure that subsequently regulate the gene expression levels. A significant increase ($P < 0.005$) of average mRNA folding energy was observed only in highly expressed tissue-specific genes, whereas there is no significant difference of mRNA folding energy between highly and lowly expressed housekeeping genes in rice. In order to determine whether selection acts on mRNA secondary structure to modulate synonymous substitution rates of tissue-specific genes, we performed correlation analysis between synonymous substitution rates of each gene with its corresponding mRNA folding energy. A significant strong positive correlation ($R_s = 0.307$, $P < 0.001$) indicates constraints on mRNA secondary structure influencing synonymous substitution rates in tissue-specific class of genes in rice. Thus, the influence of constraints acting on mRNA secondary

Table 3. RSCU values of highly expressed (HEG) and lowly expressed (LEG) housekeeping and tissue-specific genes in Arabidopsis

AA	Codons	RSCU (HEG)	RSCU (LEG)	tRNA copy number of Arabidopsis	AA	Codons	RSCU (HEG)	RSCU (LEG)	tRNA copy number of Arabidopsis
Phe	TTT	0.87 (1.05)	1.04 (1.13)	0	Ala	**GCT**^{H}	1.9 (1.8)	1.73 (1.76)	16
	TTC^{H}	1.13 (0.95)	0.96 (0.87)	16		**GCC**^{H}↗	0.7 (0.65)	0.57 (0.57)	0
Tyr	TAT	0.81 (0.97)	1.12 (1.13)	0		GCA	0.95 (1.1)	1.18 (1.07)	10
	TAC*	1.19 (1.03)	0.88 (0.87)	76		GCG	0.45 (0.45)	0.52 (0.6)	7
His	CAT	1.01 (1.18)	1.24 (1.32)	0	Gly	**GGT**^{H}↘	1.55 (1.32)	1.33 (1.33)	1
	CAC^{H}	0.99 (0.82)	0.76 (0.68)	10		GGC	0.5 (0.54)	0.46 (0.46)	23
Asn	AAT	0.84 (1)	1.04 (1.1)	0		GGA	1.47 (1.53)	1.57 (1.53)	12
	AAC^{H}	1.16 (1)	0.96 (0.9)	16		GGG	0.48 (0.61)	0.64 (0.69)	5
Asp	GAT	1.26 (1.28)	1.38 (1.38)	0	Leu	TTA	0.57 (0.69)	0.82 (0.87)	6
	GAC*	0.74 (0.72)	0.62 (0.62)	26		TTG	1.36 (1.39)	1.16 (1.38)	10
Cys	TGT	1.16 (1.11)	1.21 (1.21)	0		CTT	1.73 (1.57)	1.62 (1.53)	12
	TGC	0.84 (0.89)	0.79 (0.79)	15		CTC	1.25 (1.01)	1.12 (0.93)	1
Gln	CAA	0.94 (1.01)	1.14 (1.13)	8		CTA	0.48 (0.58)	0.68 (0.66)	10
	CAG^{H}	1.06 (0.99)	0.86 (0.87)	9		CTG	0.62 (0.76)	0.6 (0.64)	3
Lys	AAA	0.75 (0.86)	1.07 (1.03)	13	Ser	TCT	1.71 (1.68)	1.61 (1.65)	37
	AAG*	1.25 (1.14)	0.93 (0.97)	18		**TCC**^{H}↗	0.82 (1.64)	0.69 (0.7)	1
Glu	GAA	0.88 (0.97)	1.06 (1.09)	12		TCA	1.1 (1.23)	1.24 (1.28)	9
	GAG*	1.12 (1.03)	0.94 (0.91)	13		TCG	0.66 (0.58)	0.77 (0.59)	4
Val	GTT	1.74 (1.58)	1.69 (1.76)	15		AGT	0.87 (1)	0.94 (0.99)	0
	GTC*↗	0.83 (0.88)	0.7 (0.66)	0		AGC	0.83 (0.87)	0.75 (0.78)	13
	GTA	0.39 (0.42)	0.6 (0.63)	7	Arg	**CGT***	1.33 (1.28)	1 (0.97)	9
	GTG	1.04 (1.12)	1.01 (0.96)	8		**CGC**^{H}↗	0.46 (0.45)	0.34 (0.43)	0
Pro	CCT	1.56 (1.57)	1.52 (1.72)	16		CGA	0.53 (0.67)	0.69 (0.78)	6
	CCC	0.5 (0.41)	0.39 (0.46)	0		CGG	0.34 (0.34)	0.72 (0.5)	4
	CCA	1.32 (1.38)	1.33 (1.28)	45		AGA	1.87 (1.87)	2.16 (2.25)	9
	CCG	0.62 (0.64)	0.76 (0.54)	5		AGG	1.47 (1.39)	1.08 (1.08)	8
Thr	ACT	1.53 (1.47)	1.39 (1.44)	10	Ile	ATT	1.29 (1.26)	1.25 (1.24)	19
	ACC*↗	1.02 (1.01)	0.7 (0.83)	0		**ATC***↗	1.3 (1.14)	1.07 (0.98)	0
	ACA	0.96 (0.95)	1.33 (1.21)	8		ATA	0.42 (0.6)	0.68 (0.78)	5
	ACG	0.49 (1.58)	0.59 (0.52)	6					

RSCU values within parenthesis represent tissue-specific genes of Arabidopsis, and the values outside represent housekeeping Arabidopsis genes. Arrows indicate the correspondence between codon and their isoaccepting tRNA based on revised wobble rules. Codons marked with asterisk hold perfect correspondence with most abundant tRNA copy number in both housekeeping and tissue-specific genes. Codons marked with superscript H show significantly ($P < 0.05$) higher preference in highly expressed housekeeping genes.

structure modulates synonymous substitution rates in rice tissue-specific genes.

3.4. *Mutational bias regulates error minimization in both rice and Arabidopsis homologous set*

It is clear from our result that selective constraint shaping synonymous codon usage has taken a different turn in both housekeeping and tissue-specific highly expressed genes. Therefore, it is quite interesting to explore evolutionary forces acting on synonymous codon usage to optimize error minimization capacity of highly expressed housekeeping and tissue-specific genes in both the plants. The evolution of genetic code took place in such a way so that it can minimize errors due to mutation and mistranslation. The theory of error minimization for the evolution of genetic codes postulates that the codons are arranged in such a way that reduces errors.[41,42] Thus synonymous codons differ in their capacity to minimize the effects of errors due to mutation or mistranslation. In *Drosophila melanogaster*, the degree of error minimization is correlated with the degree of codon usage bias.[43] Later, it was reported that the

codon usage pattern of highly expressed genes in *E. coli* has been selected in such a way that mistranslation would have the minimum possible effects on the structure and function of the related proteins. Furthermore, according to Najafabadi et al.[44] frequencies of codons in highly expressed genes that correspond to most abundant tRNA copy number may have been under selection pressure for error minimization. For rice genome, we have calculated the error minimization capacity (wRn) of housekeeping and tissue-specific genes. We observed significant lowering of wRn ($P < 0.001$) for housekeeping genes ($wRn = -0.3322$) with respect to tissue-specific genes ($wRn = -0.2458$). This result indicates the presence of stronger selective constraint on codon usage of housekeeping genes to achieve greater degree of error minimization capacity. We compared wRn between highly and lowly expressed genes of housekeeping and tissue-specific categories of rice genome (Table 4). We observed significantly ($P < 0.001$) greater error minimizing capacity for highly expressed housekeeping genes than lowly expressed housekeeping genes. Surprisingly, in tissue-specific genes, we observed no significant difference of error minimization between highly and lowly expressed genes in rice. Thus, selection on codon usage for error minimization has hardly had any role in distinguishing highly and lowly expressed tissue-specific genes. Our observations for housekeeping genes are in consistent with the previous findings that highly expressed genes are those having a strong preference for codons to minimize the effect of errors by mutation and mistranslation.[30,44–47] We also performed the same analysis for Arabidopsis genes and observed that highly expressed genes in both housekeeping and tissue-specific categories have significantly ($P < 0.001$) greater error minimizing capacity than lowly expressed genes (Table 5). Therefore, selection acting on synonymous codon usage to optimize error minimization capacity of highly expressed genes equally influences both housekeeping and tissue-specific homologous genes of Arabidopsis. However, it is noteworthy that there is no significant difference in error minimizing capacity between highly expressed housekeeping and tissue-specific Arabidopsis genes. This discrepancy between translational selection driven by tRNA copy number and genetic robustness in both plants indicate that error minimizing capacity of highly expressed genes does not depend on selection based on tRNA abundance for both rice and Arabidopsis as observed in *E. coli*.[44,45] It is reasonable to assume from our results that frequencies of codons in highly expressed genes that correspond to most abundant tRNA copy number may not be under selection pressure for error minimization.

However, according to Archetti[43] if genetic robustness is correlated with GC composition then mutational bias is a reason behind the observed pattern of error minimization. In order to investigate if observed pattern of error minimization in rice and Arabidopsis is due to mutational bias, we measured GC_3 level for both highly and lowly expressed homologous genes of housekeeping and tissue-specific genes in rice and Arabidopsis. A significant difference in average GC_3 ($P < 0.001$) level has been observed between highly and lowly expressed genes of both housekeeping and tissue-specific homologous genes of Arabidopsis (Table 6). Correlation analysis was performed between GC content and error minimization capacity of both housekeeping and tissue-specific genes of Arabidopsis. A significant strong negative correlation has been observed between error

Table 4. Average error minimization values (wRn) of housekeeping and tissue-specific classes of genes in highly expressed (HEG) and lowly expressed genes (LEG) of rice

	Housekeeping (wRn)	Tissue specific (wRn)
HEG	−0.39463	−0.26440
LEG	−0.28266	−0.24700
Level of significance	$P < 0.001$	NS

Level of significance between highly expressed (HEG) and lowly expressed (LEG) housekeeping and tissue-specific genes of rice is shown. NS indicates average values of error minimization (wRn) not significant between highly and lowly expressed tissue-specific genes of rice.

Table 5. Average error minimization values (wRn) of housekeeping and tissue-specific classes of genes in highly expressed (HEG) and lowly expressed genes (LEG) of Arabidopsis

	Housekeeping (wRn)	Tissue specific (wRn)
HEG	−0.1937	−0.1514
LEG	−0.0059	0.0531
Level of significance	$P < 0.001$	$P < 0.001$

Level of significance between highly expressed (HEG) and lowly expressed (LEG) housekeeping and tissue-specific genes of Arabidopsis is shown.

Table 6. Average GC_3 values for housekeeping and tissue-specific classes of genes in highly expressed (HEG) and lowly expressed genes (LEG) of Arabidopsis

	Housekeeping	Tissue specific
HEG	45.64	45.34
LEG	41.88	40.46
Level of significance	$P < 0.005$	$P < 0.001$

Level of significance between highly expressed (HEG) and lowly expressed (LEG) housekeeping and tissue-specific genes of Arabidopsis is shown.

Table 7. Average GC_3 values for housekeeping and tissue-specific classes of genes in highly expressed (HEG) and lowly expressed genes (LEG) of rice

	Housekeeping	Tissue specific
HEG	69.12	68.71
LEG	62.17	65.84
Level of significance	$P < 0.001$	NS

Level of significance between highly expressed (HEG) and lowly expressed (LEG) housekeeping and tissue-specific genes of rice is shown. NS indicates average values of GC_3 not significant between highly and lowly expressed tissue-specific genes of rice.

minimization capacity and GC content of both housekeeping ($R_s = -0.541$, $P < 0.001$) and tissue-specific genes ($R_s = -0.499$, $P < 0.001$) in Arabidopsis (Supplementary Tables S6–S9 contain Arabidopsis housekeeping and tissue-specific homologous genes and their corresponding GC_3 and error minimization values). However, in rice, there is no significant difference of GC_3 between highly and lowly expressed tissue-specific genes (Table 7). Rather, we observed a significant difference in average GC_3 level only between highly and lowly expressed housekeeping genes in rice (Table 7). There is a significant ($P < 0.001$) increment of GC content in highly expressed housekeeping genes of rice genome; consistent with this, we found that synonymous substitution rate of GC-rich rice housekeeping genes ($Ks = 2.54$) is significantly ($P < 0.001$) lower than GC-poor housekeeping genes ($Ks = 3.63$). In addition, it has been further estimated that the synonymous substitution rate (Ks) is negatively correlated ($R_s = -0.216$, $P < 0.01$) with GC content at third codon position in housekeeping set of genes in rice. The result suggests that increment of GC in highly expressed housekeeping genes is under selection to optimize synonymous substitution rates.

Correlation analysis was again performed between GC content and error minimization capacity of housekeeping genes in rice. A significant strong negative correlation ($R_s = -0.606$, $P < 0.001$) has been observed between error minimization capacity and GC content of housekeeping genes in rice. These lead us to conclude that in plants it is the mutational bias that regulates error minimization of highly expressed genes.

3.5. Conclusion

In this work, we studied how selective constraint shape synonymous codon usage of housekeeping and tissue-specific homologous genes in both rice and Arabidopsis. We observed that there is difference in codon usage pattern between housekeeping and tissue-specific genes in both rice and Arabidopsis genes. Although, previous studies on Drosophila and rodents favor selectionist model for error minimization at protein level,[30] we demonstrated that mutational bias is responsible for the observed pattern of error minimization. We argue that error minimization at protein level has taken a different turn after the divergence of plants and animals. Moreover, our results show that housekeeping genes are under stronger selective constraint than that of the tissue-specific genes. Translational selection driven by tRNA copy number is responsible for optimizing codon usage variation in housekeeping genes. On the contrary, in housekeeping genes, selection acting on mRNA secondary structural stability of tissue-specific genes has a greater influence to modulate codon usage variation. Lavner and Kotlar[48] argued that selection may act on codon bias to reduce elongation rate by favoring non-optimal codons in lowly expressed genes. In the present study, influence of mRNA secondary structural stability on codon usage variation of tissue-specific genes might be the consequence of favoring non-optimal codons in lowly expressed tissue-specific genes. Thus, our study unambiguously suggests that two sets of genes in rice and Arabidopsis (housekeeping and tissue specific) have evolved under contrasting evolutionary constraints.

Acknowledgements: Authors are also thankful to Dr Nakai Kenta and two anonymous reviewers for their fruitful constructive comments in improving the manuscript.

Supplementary Data: Supplementary data are available online at www.dnaresearch.oxfordjournals.org.

Funding

Authors are thankful to Department of Biotechnology, Government of India for financial help.

References

1. International Rice Genome Sequencing Project 2005, The map-based sequence of the rice genome, *Nature*, **436**, 793–800.
2. The *Arabidopsis* Genome Initiative 2000, Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*, *Nature*, **408**, 796–815.
3. Bernardi, G. 2004, *Structural and Evolutionary Genomics: Natural Selection in Genome Evolution*, Elsevier Amsterdam: The Netherlands.
4. Wang, H. C. and Hickey, D. A. 2007, Rapid divergence of codon usage patterns within the rice genome, *BMC Evol. Biol.*, **7**, 1–10.

5. Montero, L. M., Salinas, J., Matassi, G. and Bernardi, G. 1990, Gene distribution and isochore organization in the nuclear genome of plant, *Nucleic Acids Res.*, **18**, 1859–1867.
6. Carels, N. and Bernardi, G. 2000, Two classes of genes in plants, *Genetics*, **154**, 1819–1825.
7. Guo, X., Bao, J. and Fan, L. 2007, Evidence of selectively driven codon usage in rice: implications for GC content evolution of Gramineae genes, *FEBS Lett.*, **581**, 1015–1021.
8. Wong, G. K., Wang, J., Tao, L., et al. 2002, Compositional gradients in Gramineae genes, *Genome Res.*, **12**, 851–856.
9. Sharp, P. M., Averof, M., Lloyd, A. T., Matassi, G. and Peden, J. F. 1995, DNA sequence evolution: the sounds of silence, *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.*, **349**, 241–247.
10. Ponger, L., Duret, L. and Mouchiroud, D. 2001, Determinants of CpG islands: expression in early embryo and isochore structure, *Genome Res.*, **11**, 1854–1860.
11. D'Onofrio, G. 2002, Expression patterns and gene distribution in the human genome, *Gene*, **300**, 155–160.
12. Vinogradov, A. E. 2003, Isochores and tissue-specificity, *Nucleic Acids Res.*, **31**, 5212–5220.
13. Arhondakis, S., Auletta, F., Torelli, G. and D'Onofrio, G. 2004, Base composition and expression level of human genes, *Gene*, **325**, 165–169.
14. Lercher, M. J., Urrutia, A. O., Pavlicek, A. and Hurst, L. D. 2003, A unification of mosaic structures in the human genome, *Hum. Mol. Genet.*, **12**, 2411–2415.
15. Duret, L. and Mouchiroud, D. 2000, Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate, *Mol. Biol. Evol.*, **17**, 68–74.
16. Hastings, K. E. 1996, Strong evolutionary conservation of broadly expressed protein isoforms in the troponin I gene family and other vertebrate gene families, *J. Mol. Evol.*, **42**, 631–640.
17. Hughes, A. L. and Hughes, M. K. 1995, Self peptides bound by HLA class I molecules are deprived from highly conserved regions of a set of evolutionary conserved proteins, *Immunogenetics*, **41**, 257–262.
18. Zhang, L. and Li, W. H. 2004, Mammalian housekeeping genes evolve more slowly than tissue-specific genes, *Mol. Biol. Evol.*, **21**, 236–239.
19. Plotkin, J. B., Robins, H. and Levine, A. J. 2004, Tissue-specific codon usage and the expression of human genes, *Proc. Natl. Acad. Sci. USA*, **101**, 12588–12591.
20. Semon, M., Lobry, J. R. and Duret, L. 2006, No evidence for tissue-specific adaptation of synonymous codon usage in humans, *Mol. Biol. Evol.*, **23**, 523–529.
21. Mukhopadhyay, P., Basak, S. and Ghosh, T. C. 2007, Nature of selective constraints on synonymous codon usage of rice differs in GC-poor and GC-rich genes, *Gene*, **400**, 71–81.
22. Altschul, S. F., Madden, T. L., Schaffer, A. A., et al. 1997, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.*, **25**, 3389–3402.
23. Banerjee, T., Gupta, S. K. and Ghosh, T. C. 2006, Compositional transitions between *Oryza sativa* and *Arabidopsis thaliana* genes linked to the functional change of encoded proteins, *Plant Sci.*, **170**, 267–273.
24. Nakano, M., Nobuta, K., Vemaraju, K., Tej, S. S., Skogen, J. W. and Meyers, B. C. 2006, Plant MPSS databases: signature-based transcriptional resources for analyses of mRNA and small RNA, *Nucleic Acids Res.*, **34**, D731–D735.
25. Meyers, B. C., Tej, S. S., Vu, T. H., et al. 2004, The use of MPSS for whole-genome transcriptional analysis in *Arabidopsis*, *Genome Res.*, **14**, 1641–1653.
26. Ren, X. -Y., Vorst, O., Fiers, M. W. E. J., Stiekema, W. J. and Nap, P. 2006, In plants, highly expressed genes are the least compact, *Trends Genet.*, **22**, 528–532.
27. Liao, B. Y. and Zhang, J. 2006, Low rates of expression profile divergence in highly expressed genes and tissue-specific genes during mammalian evolution, *Mol. Biol. Evol.*, **23**, 1119–1128.
28. Yanai, I., Benjamin, H., Shmoish, M., et al. 2005, Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification, *Bioinformatics*, **21**, 650–659.
29. Yang, Z. and Nielsen, R. 2000, Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models, *Mol. Biol. Evol.*, **17**, 32–43.
30. Archetti, M. 2004, Selection on codon usage for error minimization at the protein level, *J. Mol. Evol.*, **59**, 400–415.
31. McLachlan, A. D. 1971, Tests for comparing related amino-acid sequences Cytochrome c and cytochrome c 551, *J. Mol. Biol.*, **61**, 409–424.
32. Kotlar, D. and Lavner, Y. 2006, The action of selection on codon bias in the human genome is related to frequency, complexity, and chronology of amino acids, *BMC Genom.*, **7**, 67.
33. Xiyin, W., Xiaoli, S. and Bailin, H. 2002, The transfer RNA genes in *Oryza sativa* L. ssp. *Indica*, *Sciences in China Series C*, **45**, 504–511.
34. Berg, O. G. and Martelius, M. 1995, Synonymous substitution-rate constants in *Escherichia coli* and *Salmonella typhimurium* and their relationship to gene expression and selection pressure, *J. Mol. Evol.*, **41**, 449–456.
35. Drummond, D. A., Raval, A. and Wilke, C. O. 2006, A single determinant dominates the rate of yeast protein evolution, *Mol. Biol. Evol.*, **23**, 327–37.
36. Ikemura, T. 1992, Transfer RNA in protein synthesis, In: Hatfield, D. L., Lee, B. J. and Pirtle, R. M. (eds.), CRC: Boca Raton, FL, pp. 87–111.
37. Duret, L. 2000, tRNA gene number and codon usage in the *C. elegans* genome are co-adapted for optimal translation of highly expressed genes, *Trends Genet.*, **16**, 287–289.
38. Percudani, R. 2001, Restricted wobble rules for eukaryotic genome, *Trends Genet.*, **17**, 133–135.
39. Biro, J. C. 2006, Indications that "codon boundaries" are physico-chemically defined and that protein-folding information is contained in the redundant exon bases, *Theor. Biol. Med. Model*, **3**, 28.
40. Jia, M. and Li, Y. 2005, The relationship among gene expression, folding free energy and codon usage bias in *Escherichia coli*, *FEBS Lett.*, **579**, 5333–5337.

41. Woese, C. R. 1965, On the evolution of the genetic code, *Proc. Natl. Acad. Sci. USA*, **54**, 1546–1552.
42. Epstein, C. J. 1966, Role of the amino-acid 'code' and of selection for conformation in the evolution of proteins, *Nature*, **210**, 25–28.
43. Archetti, M. 2006, Genetic robustness and selection at the protein level for synonymous codons, *J. Evol. Biol.*, **19**, 353–365.
44. Najafabadi, H. S., Goodarzi, H. and Torabi, N. 2005, Optimality of codon usage in *Escherichia coli* due to load minimization, *J. Theor. Biol.*, **237**, 203–209.
45. Najafabadi, H. S., Lehmann, J. and Omidi, M. 2007, Error minimization explains the codon usage of highly expressed genes in *Escherichia coli*, *Gene*, **387**, 150–155.
46. Bulmer, M. 1991, The selection-mutation-drift theory of synonymous codon usage, *Genetics*, **129**, 897–907.
47. Akashi, H. 1994, Synonymous codon usage in Drosophila melanogaster: natural selection and translational accuracy, *Genetics*, **136**, 927–935.
48. Lavner, Y. and Kotlar, D. 2005, Codon bias as a factor in regulating expression via translation rate in the human genome, *Gene*, **345**, 127–138.