Gene 513 (2013) 163–173

Contents lists available at SciVerse ScienceDirect

Gene

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

Comparative genome sequence analysis of *Sulfolobus acidocaldarius* and 9 other isolates of its genus for factors influencing codon and amino acid usage

Kinshuk Chandra Nayak *

Bioinformatics Centre, Institute of Life Sciences, Department of Biotechnology, Govt. India, Nalco Square, Bhubaneswar, 751 023, India

ARTICLE INFO

Article history:
Accepted 21 October 2012
Available online 30 October 2012

Keywords:
Sulfolobus acidocaldarius
RSCU values
SCUO values
Correspondence analysis (COA)
Nc values
GC skews

ABSTRACT

In the present study, major constraints for codon and amino acid usage of *Sulfolobus acidocaldarius*, *Sulfolobus solfataricus*, *Sulfolobus tokodali*, *Sulfolobus islandis* and 6 other isolates from *islandicus* species of genus *Sulfolobus* were investigated. Correspondence analysis revealed high significant correlation between the major trend of synonymous codon usage and gene expression level, as assessed by the "Codon Adaptation Index" (CAI). There is a significant negative correlation between Nc (Effective number of codons) and CAI demonstrating role of codon bias as an important determinant of codon usage. The significant correlation between major trend of synonymous codon usage and GC3s (G + C at third synonymous position) indicated dominant role of mutational bias in codon usage pattern. The result was further supported from SCUO (synonymous codon usage order) analysis. The amino acid usage was found to be significantly influenced by aromaticity and hydrophobicity of proteins. However, translational selection which causes a preference for codons that are most rapidly translated by current tRNA with multiple copy numbers was not found to be highly dominating for all studied isolates. Notably, 26 codons that were found to be optimally used by genes of *S. acidocaldarius* at higher expression level and its comparative analysis with 9 other isolates may provide some useful clues for further *in vivo* genetic studies on this genus.

1. Introduction

Codon usage and codon preferences vary significantly within and between organisms (Akashi et al., 2007; Grantham et al., 1981; Guo and Yuan, 2009; Liu et al., 2010; Nayak, 2009, 2011; Ranjan et al., 2007; Sau and Deb, 2008; Sharp et al., 1988; Zhao et al., 2007; Zhou and Li, 2009). Traditionally, codon usage data have been used in a wide variety of areas (McInerney, 1998). It is often desirable to use codon usage information to reduce the redundancy of primers for the PCR (polymerase chain reactions) and optimized the codon usage tables for identifying those ORFs (open reading frames) that may encode proteins. Codon usage patterns also have been used to identify ORFs that probably do not code for functional proteins. The codon usage study has become one of the most important scientific issues in studying molecular genetic engineering for desirable translational efficiency and a useful tool to explore shifts in the mutation–selection balance across bacterial species with different life styles. Quantification of codon usage bias, especially at genomic scale, helps to understand evolution of living organisms, because bias results in exaggerated divergence of patterns of codon use between taxa. In addition, the presence of homogeneity in codon usage patterns among the genes of genomes was argued for the reason that they have not gone through purifying selection.

Synonymous codon usage bias was found to be correlated with many factors such as base compositional mutation bias (Hou and Yang, 2003; Karlin and Mrazek, 1996), level of gene expression (Duret and Mouchiroud, 1999; Peixoto et al., 2003; Romero et al., 2003; Sharp and Li, 1986) , gene length (Moriyama and Powell, 1998), transfer RNA (tRNA) abundance(Duret, 2000; Ohkubo et al., 1987) , protein structure (Gu et al., 2004) , codon–anticodon interaction (Xiufan et al., 2001) , hydropathy level of protein, amino acid conservation (Romero et al., 2000), etc. Among them, compositional mutation bias and natural selection with different relative importance in different species, mainly contributed to codon bias (Gu et al., 2004; Peixoto et al., 2003; Romero et al., 2003; Sharp et al., 1993). In some prokaryotic genomes, the codon usage pattern was attributable to the equilibrium between natural selection and compositional mutation bias. However, in extremely AT (A and T base composition) or GC (G and C base composition) rich unicellular genomes, compositional constraints played a dominant role in shaping codon usage variation among genes (Gupta et al., 2004;

Abbreviations: A, Adenosine; ABC, ATP-binding cassette; AU, Amino acid usage; C, Cytosine; CAI or *cai*, Codon Adaptation Index; COA, Correspondence Analysis; EMBOSS, European Molecular Biology Open Software Suite; FTP, File transfer protocol; G, Guanine; GAE, Genome analysis environment; GCSI, GC Skew Index; GRAVY, General average hydropathicity; ORF, Open reading frame; *P*-value, Probability or level of significance value; PCR, Polymerase chain reactions; *r*, Spearman correlation coefficient; RSCU, Relative synonymous codon usage; Saci, *Sulfolobus acidocaldarius*; SCUO, Synonymous Codon Usage Order; SPSS, Statistical Package for Social Science; T, Thymine; TIGR, The Institute of Genomic Research; tRNA, Transfer RNA (Ribo Nucleic Acid); | |, Absolute value; **, Statistical significance at $P<0.01$; *, Statistical significance at $P<0.05$.

* Tel.: +91 674 2301460x279; fax: +91 674 2300728.
E-mail addresses: kinshukils@hotmail.com, kinshuk@ils.res.in, kcnils@gmail.com.

0378-1119/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.gene.2012.10.024

Ohkubo et al., 1987; Wright and Bibb, 1992). Moreover, for some prokaryotic genomes the G + C base compositional bias and translational selection were considered to be the most important factors affecting the codon usage variation (Kanaya et al., 1999). In many cases translational selection determines the codon usage bias of highly expressed genes and the reason of this observation was subsequently found to be due to the presence of preferred codons in highly expressed genes which were recognized by most abundant tRNAs (Bennetzen and Hall, 1982; Ikemura, 1981, 1982). Adherence to these codon usages biases in prokaryotes is selectively advantageous and has been shown to be responsible for three to six fold differences in translation rates (Robinson et al., 1984) and up to ten-fold differences in the accuracy of translation (Precup and Parker, 1987). Therefore, analysis of codon usage data has both theoretical and practical significance in understanding the basics of molecular biology.

Sulfolobus acidocaldarius strain DSM639 is the type strain of the archaeal genus *Sulfolobus* and model organism of *Crenarchaeota*. It was the first hyperthermoacidophile to be characterized from terrestrial solfataras by Brock et al. (1972) and its complete genome sequence submitted recently. It grows optimally at 75 to 80 °C and pH 2 to 3, under strictly aerobic conditions, on complex organic substrates, including yeast extract, tryptone, and Casamino Acids and a limited number of sugars (Chen et al., 2005). It has been used for many seminal studies on archaea and crenarchaea. Thus it was employed for demonstrating similarity among archaeal and eukaryal transcription apparatuses (Bell et al., 2002; Langer et al., 1995; Puehler et al., 1989). In addition, its sensitivity to wide range of ribosomal antibiotics (Aagaard et al., 1994) and ease of transformation (Aagaard et al., 1996) have rendered *S. acidocaldarius* a focus for in vivo genetic studies.

S. acidocaldarius has also been used for studying genetic fidelity at high temperatures and is the only hyperthermophilic archaeon for which the rate and type of spontaneous mutation have been quantified in vivo (Grogan et al., 2001). This, together with presence of proteins responsible for chromatin folding, UV damage excision repair system and stable genome organization at high temperature has generated interest for its further genetic study. Moreover, thermostable enzymes secreted by these organisms have found commercial applications in the starch industry, in the pulp, petroleum, chemical and paper industry. Thus in-depth investigation about their genetics at genome level will help in mastering the cloning, finding suitable heterologous protein expression system and industrial exploitation of variety of genes which encode enzymes involved in starch hydrolysis, amino acid biosynthesis, protein hydrolysis, etc.

For this purpose, a comprehensive analysis of its codon and amino acid usage pattern for identifying different selection pressures was carried out at its genomic level using multivariate statistical techniques and non-parametric tests. In order to give detail genome characteristics of codon usage pattern at genus level, the study included *S. acidocaldarius, S. solfataricus* (She et al., 2001), *S. tokodali* (Kawarabayahi et al., 2001), *Sulfolobus islandis* (Reno et al., 2009) and 6 other isolates from *islandicus* species of this genus for comparative purpose. The result of this study will serve as an important research resource for codon optimization process of synthetic genes to get desirable translational efficiency, getting optimized gene construct having its adaptation to high temperature, phylogenetic or evolutionary studies of the crenarchaeal kingdom of *Archaea*, and other genetic studies of less complex archaeal systems in order to understand the corresponding, and more complex, systems in eukaryotes. Moreover, results for codon and amino acid usage profiles for these genomes will serve as important tools, which can be utilized to improve their function predictions and genome-environment mappings. The study has further investigated about translational efficiency considering that codon biases which might match the tRNA abundances to maximize speed and efficiency at the level of translation (Akashi, 1994; Bulmer, 1991; Ikemura, 1981, 1985; Sorensen et al., 1989) In addition, codon–anticodon base pairing preferences were analyzed for all genomes of genus *Sulfolobus* to explore possible role of translation regulation with reference to tRNA abundances. Studies on analysis of codon usage bias will help in understanding lifestyle and exploring reasons for their adaptation to extreme environment.

2. Materials and methods

2.1. Coding sequences

Coding sequences *S. acidocaldarius, S. solfataricus*, *S. tokodali* and 6 other isolates from *islandicus* species were obtained from the GenBank FTP site (http://www.ncbi.nlm.nih.gov/ftp/). Sequences greater than 300 nucleotides in length were included directly to avoid sampling bias in codon usage calculations (Wright, 1990). The transfer RNA (tRNA) anticodon information of these organisms was obtained from TIGR annotation and program "tRNAscan-SE" at (http://gtrnadb.ucsc.edu/).

2.2. Index for codon usage and synonymous codon usage bias

GC_{3s} (frequency of codons ending in G + C at third synonymous positions, excluding Met, Trp and stop codons), relative synonymous codon usage (RSCU) (Sharp and Li, 1986), correspondence analysis (COA) (Greenance, 1984) of RSCU values, overall codon frequency, GRAVY (General average hydropathicity) and aromaticity scores of coding sequences were calculated using the program CODONW 1.3 (written by John Peden), available from (http://www.molbiol.ox.ac.uk/cu). The codons encoding for Met, Trp and stop codons were excluded from analysis. General average hydropathicity (or GRAVY score) for the gene product was calculated as the arithmetic mean of the hydropathic indices of each amino acid. The hydropathicity and aromaticity protein scores are indices of amino acid usage.

Nc, the effective number of codons (Wright, 1990) is a widely accepted measure that quantifies the magnitude of codon bias for an individual gene. It finds how a small subset of codons is used by a gene and its value range from 61(when all the codons are used with equal frequency) to 20 (when one codon is used per amino acid). This measure has been found to have a relationship to G + C base composition at the third synonymous position. In this study both the measures (Nc and GC_{3s} values) were calculated for all species.

"Codon Adaptation Index (CAI)" was used to estimate the extent of bias towards codons usage. It was known to be higher in highly expressed genes. CAI value lies between 0.0 and 1.0, higher indicate more bias (Sharp and Li, 1987). CAI has been proved to be the best gene expression value index and was extensively used as a measure of gene expression level (Naya et al., 2001; Wright and Bibb, 1992). In our analysis, 50S ribosomal proteins, heat shock protein, translation factor, transcriptional regulator, RNA polymerases, DNA repair ATPase and hypothetical Saci were found at extreme end of primary axes with CAI values ranging from 0.6 to 0.8. CAI values were calculated by *cai* program of EMBOSS (http://www.ch.embnet.org/EMBOSS/) package.

SCUO (synonymous codon usage order) measurement (Wan et al., 2004) was applied to compare the synonymous codon bias and GC variation in different codon positions among the three intestinal lactobacilli using CodonO program (http://sysbio.cvm.msstate.edu/CodonO/).

In order to normalize codon usage within genes of differing amino acid composition, relative synonymous codon usage (RSCU) values were calculated by dividing the observed codon usage by the expected when all codons for the same amino acid are used equally. RSCU values are defined as the ratio of observed frequency of a codon to the expected frequency, if all the synonymous codons for those amino acids were used equally (Sharp and Li, 1986). Converting codon usage values to RSCU values has the effect of 'normalizing'

comparisons across genes. This makes the codon usage value independent of amino acid composition of the sequences and identifies when a codon is being used more frequently than expected and when it is being used less frequently than expected. RSCU value of a codon >1 implies that the codon was used more frequently for the amino acid as compared to those having RSCU <1. Codons having RSCU values >1 and statistically significant between 10% genes located at extreme ends principal axis was considered as optimal and other codons with RSCU values <1, called rare codons were not included for optimal codon usage analysis. Therefore, here optimal codons were identified by comparing the codon usage frequencies between genes with high and low expression levels. RSCU values were calculated for each gene in the genome for 59 codons (omitting the two unbiased codons AUG and UGG out of 61 sense codons) and

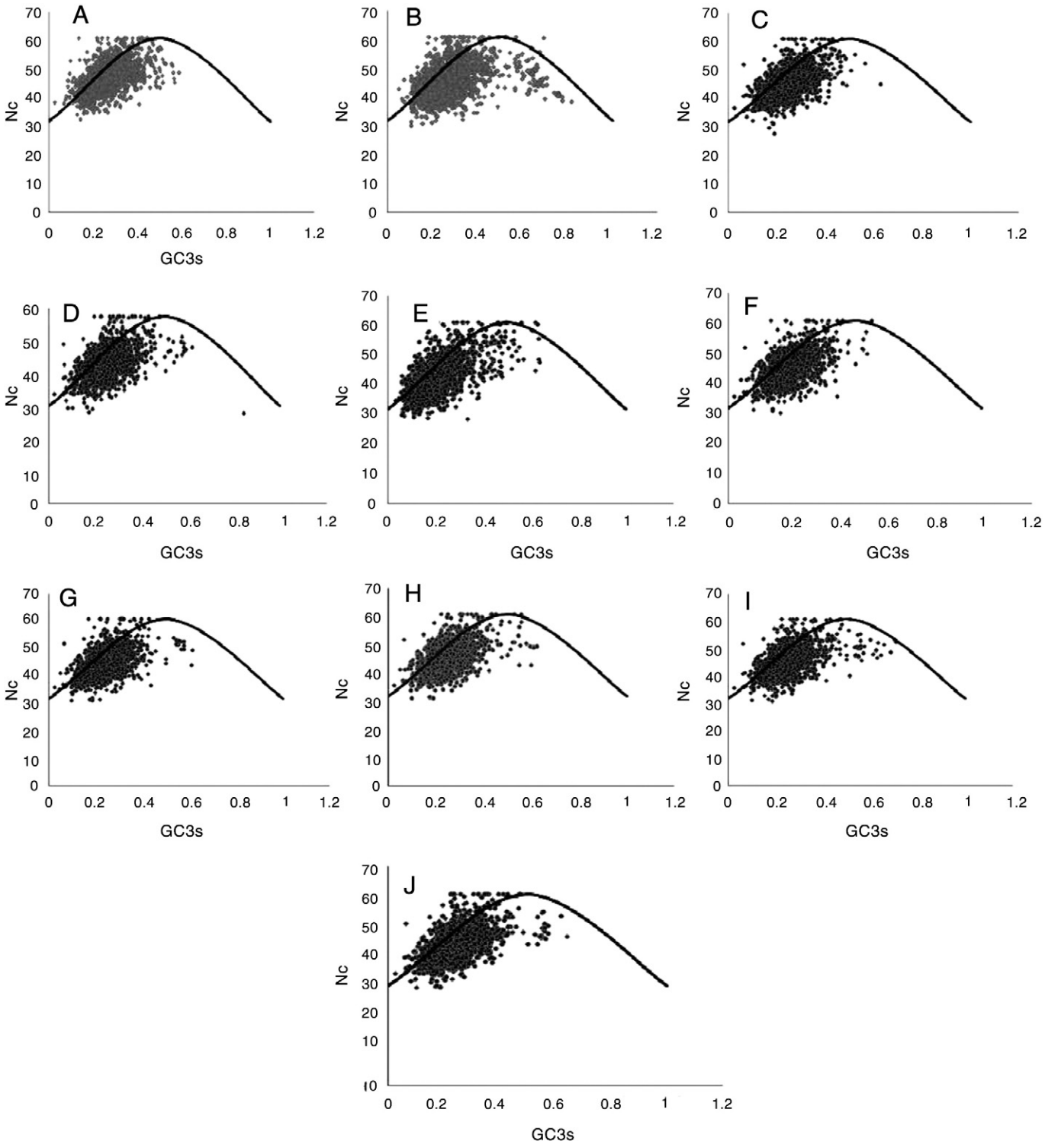

Fig. 1. Nc plots of A: *S. acidocaldarius* DSM 639; B: *S. solfataricus* P2; C: *Sulfolobus tokodaii*; D: *S. islandicus* L.D.8.5; E: *S. islandicus* L.S.2.15; F: *S. islandicus* M.14.25; G: *S. islandicus* M.16.27; H: *S. islandicus* M.16.4; I: *S. islandicus* Y.G.57.14; J: *S. islandicus* Y.N.15.51 genes. The continuous curve represents the expected curve between Nc and GC_{3s} during random codon usage. The dark squares represents (Nc, GC3s) value of each gene.

correspondence analysis of such values for all genes were plotted in a 59-dimensional hyperspace according to the usage of the 59 sense codons. The result was represented as 59 orthogonal axes.

2.3. Correspondence analysis (COA)

Correspondence analysis (COA) (Greenance, 1984) of RSCU values was performed to identify the intra-genomic variation of amino acid compositions using CodonW 1.4.2 (developed by John Peden available from (http://www.molbiol.ox.ac.uk/cu)). It is a technique which creates a series of orthogonal axes to identify trends that explain the data variation, with each subsequent axis coordinates explaining a decreasing amount of variation. COA positions each gene and codon (or amino acid) on these axes and the ordination of the rows (genes) and columns (codons or amino acids) are super imposable. Therefore, major sources of synonymous codon usage variation were revealed from the correspondence analysis (COA) on relative synonymous codon usage (RSCU). It helps to find the difference in codon usage among genes and identifying the codons involved in such variations. In subsequent part of study axis 1 (RSCU) and axis 2 (RSCU) will represent first and second major axis of correspondence analysis on RSCU (or simply codon usages) whereas, axis 1 (AA) and axis 2 (AA) will represent first and second major axis of correspondence analysis on amino acid usages.

2.4. Statistical test

Chi-square test, involving a 2(rows)×2(columns) table that yields one degree of freedom, was used to examine the significance of codon usage difference between two datasets involving highly and lowly expressed genes estimated by CAI and Nc values. Correlation (Spearman's rank correlation) and variance analysis was carried out (with the level of significance $P<0.01$ or $P<0.05$) using SPSS version 13.0.

2.5. GC skew and DNA walk

GC skew, (G–C) / (G+C) is calculated across a genome as the sum of a series of sliding windows of specified length, the window size can be 1 or much larger for a complete genome. GC and AT skews have been widely used to predict termini and origins of replication in bacterial (Kunst et al., 1997; Mrazek and Karlin, 1998) and mammalian genomes (Touchon et al., 2005), transcription start sites in plants and fungi (Fujimori et al., 2005) as well as transcription regions in the human genome (Touchon et al., 2003). For our genomes, skew plot was constructed with default window size of 1 nucleotide.

In a "DNA walk" graph, sequences are plotted starting at $X=0, Y=0$. For each nucleotide (from position 1 in the sequence to the end), the position of the next point in the plot is calculated relative to the current position: for nucleotide C, G, T or A, the position moves north, south, east or west, respectively. If the current symbol is degenerate or a gap symbol, the position is unchanged. For a window size of k, every k' th point is actually drawn on the graph (but the calculations still include every nucleotide). A slider bar moves markers along the plots to locate specific regions of the sequences since the position of a particular nucleotide is solely dependent on the composition of the preceding nucleotide sequence, not on the position in the sequence. GC skew and DNA walk algorithms were used as described in the program Graph DNA (Thomas and Horspool, 2007).

2.6. Translational optimal codon

The anticodon table and copy number was generated by using program "tRNAscan-SE" (http://gtrnadb.ucsc.edu/).

Table 1
Bivariate correlation between axis 1, axis 2 and CAI, Nc, gene length and GC3s, for codon usage and between axis 1, axis2, GRAVY score and aromaticity scores of amino acid usage for all organisms of *Sulfolobus* genus. All correlations with $P<0.01$ were considered for discussion. The lower r value with $P<0.01$ indicated significant poor correlation. To avoid stochastic error all genes greater than 300 bp were included in the discussion.

Organism (Number of CDS > 100 codons)	Axes (% inertia) (RSCU) (AA)	CAI	Nc	GC3s	GC	Length	GRAVY (AA)	Aromaticity (AA)
Sulfolobus acidocaldarius	Axis1 8.8 26.2	−0.759**	0.596**	0.667**	0.575**	0.062**	−0.735**	−0.438**
DSM 639	Axis2 4.8 10.4	0.077**	−0.068**	−0.134**	−0.094**	−0.041	−0.154**	0.211**
(2002)								
Sulfolobus solfataricus P2	Axis1 13.8 24.5	0.791**	−0.537**	−0.714**	−0.496**	0.056**	0.762**	0.387**
(2757)	Axis2 5.8 11.2	0.089**	0.027	−0.059**	0.021	0.014	−0.005	0.225**
Sulfolobus	Axis1 10.4 24.0	−0.820**	0.684**	0.664**	0.363**	−0.083**	0.795**	0.423**
tokodali str. 7	Axis2 6.2 12.2	−0.167**	0.094**	0.088**	0.184**	0.053**	−0.019	−0.255**
(2562)								
Sulfolobus	Axis1 7.6 25.1	0.611**	−0.511**	−0.531**	−0.355**	0.027	0.759**	0.450**
islandicus L.D.8.5	Axis2 5.7 11.6	0.245**	−0.150**	−0.163**	0.021	0.061**	0.012	−0.296**
(2387)								
Sulfolobus islandicus	Axis1 6.8 25.2	0.455**	−0.438**	−0.426**	−0.312**	−0.001	−0.751**	−0.435**
L.S.2.15	Axis2 6.0 12.0	0.256**	−0.171**	−0.137**	0.070**	0.052*	−0.007	−0.310**
(2390)								
Sulfolobus islandicus	Axis1 7.0 25.3	−0.453**	0.452**	0.409**	0.361**	−0.008	−0.739**	−0.464**
M.14.25	Axis2 6.0 12.0	0.210**	−0.115**	−0.154**	−0.121**	0.019	−0.014	−0.276**
(2316)								
Sulfolobus islandicus	Axis1 7.3 25.0	−0.502**	0.458**	0.445**	0.395**	−0.003	−0.735**	−0.461
M.16.27	Axis2 5.6 12.2	−0.114**	0.046*	0.066**	0.043*	−0.001	0.033	−0.262**
(2353)								
Sulfolobus	Axis1 7.2 25.2	0.601**	−0.545**	−0.500**	−0.232**	0.023	−0.745**	0.444**
islandicus	Axis2 6.3 12.1	−0.181**	0.215**	0.194**	0.287**	0.017	−0.005	0.292**
M.16.4								
(2349)								
Sulfolobus islandicus	Axis1 8.6 24.9	−0.502**	0.458**	0.445**	0.395**	−0.003	−0.735**	−0.461**
Y.G.57.14	Axis2 5.6 11.8	−0.114	0.046*	0.066**	0.043*	−0.001	0.033	−0.262**
(2432)								
Sulfolobus islandicus	Axis1 8.0 24.5	0.576**	−0.459**	−0.459**	−0.439**	−0.023	−0.759**	0.421**
Y.N.15.51	Axis2 5.6 11.8	0.227**	−0.163**	−0.134**	−0.057**	0.039	0.020	−0.262**
(2432)								

** for $P<0.01$; * for $P<0.05$; AU: Amino acid usage.

3. Results and discussion

3.1. Overall codon usage, amino acid usage and nucleotide compositional constraint analysis

S. acidocaldarius constitute a circular chromosome with A + T content of 63.3%. Since it is a AT-rich organism, it was expected that A and /or U ending codons will predominate in the coding regions of this species. Indeed, six AT rich amino acids (F, Y, M, I, N, K) were U or A-ending codons and overall, this preference was also found for other codons with RSCU > 1(excluding Met, Trp and stop codons). Similar results were obtained for all other 9 isolates (Table S1, supplementary data). This non-randomness in the array of codon usage pattern could be attributed to mutational bias, or to natural selection favoring specific codons. Preference for certain codons may be due to an abundance of specific tRNA molecules available for translation. Therefore, we have carried out a comprehensive study of Nc plot analysis, COA of RSCU values, SCUO analysis and codon–anticodon matching to probe above possibilities.

The plot of effective number of codons (Nc) and G + C base composition at third synonymous position (GC_{3s}), called Nc plot, was effectively used to explore the codon usage variation in genome of an organism. Wright (1990) argued that the comparison of actual distribution of genes with expected distribution under no selection could be indicative if codon usage bias of genes has some other influences other than compositional constraints. It is interesting to note that in all isolates, although there were a small number of genes lying on or above the continuous Nc-plot curve, a majority of points with low Nc values were lying well below the expected curve and towards GC3s poor area (Fig. 1), suggesting that apart from the dominant role of A + T compositional constraints, other factors might have influence in dictating codon usage variation among genes (Gupta et al., 2004; Hou and Yang, 2003).

In order to know the role of mutational bias on synonymous codon usage, bivariate correlation analysis was carried out between Axis1 (RSCU) and GC3s for all isolates. The analysis showed (Table 1) significant ($P<0.01$) positive correlation coefficients for 5 isolates including the case of *S. acidocaldarius* and negative for other 5 organisms. To validate further about trend of mutational pressure on codon usage pattern, SCUO analysis was carried out for these isolates and the resulting plot supported it by showing a higher non-linear trend (with negative slope) for GC3 as compared to GC, GC1, and GC2 (Fig. 2). These results together suggest that, mutational pressure plays a dominant role in shaping codon usage pattern during evolution of all isolates and its asymmetric trend at third codon position might have provided a more flexible selection ability to these isolates during environmental adaptation process.

Interestingly, highest significant correlation was found between GC3s and CAI for *S. acidocaldarius* ($r=-0.8$, $P<0.01$) (Table 2). Furthermore, coefficients of correlation for other 9 isolates were found significant at higher degree (r value ranging from −0.648 to −0.756, $P<0.01$) (Table 2). It indicated that genes at higher expression level were highly influenced by mutational pressure with preference of A and/or T at wobble position. Above finding was further confirmed from highly significant positive correlation (r value ranging from 0.471 to 0.591, $P<0.01$) between Nc and GC3s (Table 2).

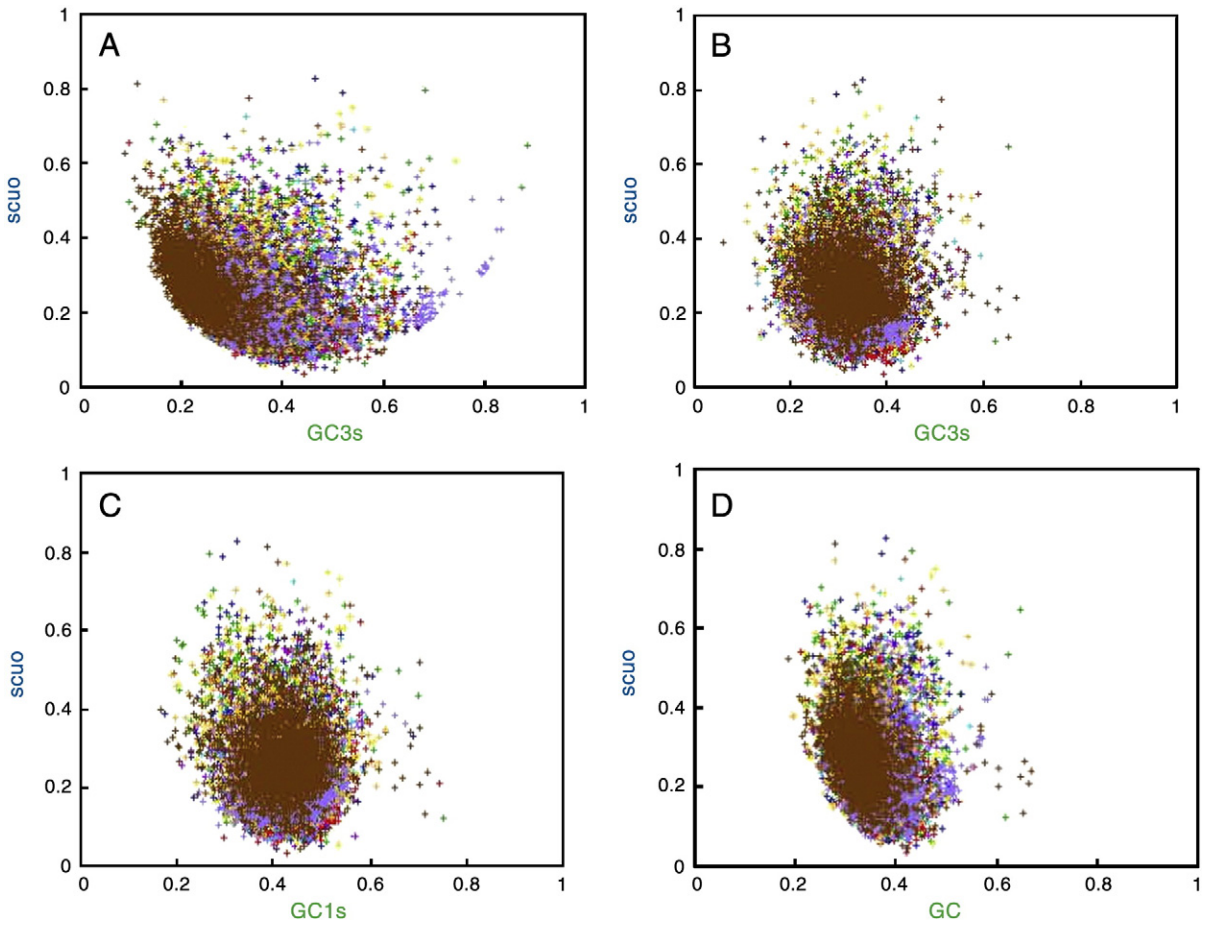

Fig. 2. Visualization of the correlation between synonymous codon usage bias (measured by SCUO) and GC composition, for organisms, *S. acidocaldarius* DSM 639, *S. solfataricus* P2, *Sulfolobus tokodaii, S. islandicus* L.D.8.5, *S. islandicus* L.S.2.15, *S. islandicus* M.14.25, *S. islandicus* M.16.27, *S. islandicus* M.16.4, *S. islandicus* Y.G.57.14 and *S. islandicus* Y.N.15.51.

Table 2
Bivariate correlation among axis1 (codon usage), CAI, Nc, GC3s. It includes all possible combinations between variables.

Organism		Axis1	CAI	Nc	GC3s
Sulfolobus acidocaldarius DSM 639	**Axis1**	1.00	−0.759**	0.596**	0.667**
	CAI		1.00	−0.650**	−0.800**
	Nc			1.00	0.573**
	GC3s				1.00
Sulfolobus solfataricus P2	**Axis1**	1.00	0.791**	−0.537**	−0.714**
	CAI		1.00	−0.532**	−0.724**
	Nc			1.00	0.471**
	GC3s				1.00
Sulfolobus tokodali str. 7	**Axis1**	1.00	−0.820**	0.684**	0.664**
	CAI		1.00	−0.687**	−0.756**
	Nc			1.00	0.591**
	GC3s				1.00
Sulfolobus islandicus L.D.8.5	**Axis1**	1.00	0.611**	−0.511**	−0.531**
	CAI		1.00	−0.581**	−0.665**
	Nc			1.00	0.508**
	GC3s				1.00
Sulfolobus islandicus L.S.2.15	**Axis1**	1.00	0.455**	−0.438**	−0.426**
	CAI		1.00	−0.586**	−0.666**
	Nc			1.00	0.515**
	GC3s				1.00
Sulfolobus islandicus M.14.25	**Axis1**		−0.453**	0.452**	0.409**
	CAI			−0.586**	−0.650**
	Nc				0.510**
	GC3s				
Sulfolobus islandicus M.16.27	**Axis1**	1.00	−0.502**	0.458**	0.445**
	CAI		1.00	−0.579**	−0.657**
	Nc			1.00	0.504**
	GC3s				1.00
Sulfolobus islandicus M.16.4	**Axis1**	1.00	0.601**	−0.545**	−0.500**
	CAI		1.00	−0.583**	−0.655**
	Nc			1.00	0.510**
	GC3s				1.00
Sulfolobus islandicus Y.G.57.14	**Axis1**	1.00	−0.502**	0.458**	0.445**
	CAI		1.00	−0.558**	−0.674**
	Nc			1.00	0.502**
	GC3s				1.00
Sulfolobus islandicus Y.N.15.51	**Axis1**	1.00	0.576**	−0.459**	−0.459**
	CAI		1.00	−0.565**	−0.648**
	Nc			1.00	0.514**
	GC3s				1.00

** for significance at $P<0.01$; * at $P<0.05$.

The COA of amino acid usage showed that, in all organisms, inertia shared by axis1 and axis2 is multiple times higher than synonymous codon usage. In order to identify amino acids that were commonly preferred during overall level of gene expression, a doughnut diagram was constructed for 20 amino acids of first two axes (Fig. 3). The diagram showed that, some amino acids, phenylalanine (aromatic and non-polar), tryptophan (aromatic), glutamic acid, arginine (long chained amino acid), alanine, isoleucine (non-polar), cysteine and proline are better represented by their weights as compared to rest of 12 amino acids. Since all these residues are favorable to thermo stabilization and carries higher weights in inertia, it could be assumed that their role is more significant than other residues having similar biophysical properties. The higher weights for usage of arginine (Fig. 3) in genes of all organisms may be due to its significant role in gene expression. This assumption hold true as the mechanism of thermostabilization is thought to depend on resonance stabilization effect of arginine and hence, arginine is assumed to contribute to protein thermostability, because it maintain ion pair more easily. The diagram shows the significant usage of cysteine. This amino acid is a small amino acid, but here it may be playing a dominant role in gene expression for these extremophiles, because it has ability to form disulphide bonds with other cysteines, which is considered to be important for thermostabilization. The reason for higher weights of nonpolar amino acids alanine, isoleucine may be due to their role in making protein core more hydrophobic, which could make the protein more stable at higher temperature. However, the weights for proline and glycine are not similar for all organisms, suggesting variability in rigidity and flexibility of their polypeptide chains. Proline is more rigid than other amino acids and having reducing entropy of the main chain, which makes the unfolding the chain less likely at higher temperatures. Unlike proline, glycine make the mainchain more flexible rather than more rigid, for the reason that, glycines do not have a side chain that restricts freedom of movement for the polypeptide chain. Since the diagram is constructed for only two primary axes, the above differences in terms of side chain interactions cannot be taken to be significant and further detail study on structural properties may reveal a better understanding of these results. The preferential usage of aromatic amino acid, tryptophan, can be thought to be important for stabilizing the fold of proteins because of their heavy compact side chains.

3.2. Gene expression level and codon bias

COA of RSCU values for *S. acidocaldarius* revealed that axis1 and axis 2 accounted for 8.2%, 5.9% of total inertia or codon usage variation respectively (Table 1). Since these two axes represent only partial extent of codon usage variation, it was postulated that there were some other major factors shaping codon usage variation in this organism. Our assumption was further supported from plot between axis 1 and axis2 of COA of RSCU values across all species and strains, where a lot of genes were scattered away from null point with higher RSCU values (Fig. 4).

Comparative analysis among other species and isolates showed similar results and inertia for them ranged from 6.5% (*S. islandicus* M.16.4) to 12.8% (*S. solfataricus* P2). The reason for higher contribution of inertia for *S. solfataricus* P2, *S. tokodali* and some other isolates might be due to more number of ORFs being included for analysis satisfying threshold value of 100 codons (shown in brackets under first column) (Table 1). However, the degree of correlation between axis1 and CAI for all species and isolates showed highly significant values ("$|r|$" ranging from 0.453 to 0.820, $P<0.01$). In addition, significant negative correlation (r value ranging from −0.532 to −0.687) was found between Nc and CAI for all species and isolates (Table 2). In these organisms, some highly expressed genes such as ribosomal proteins (50s, 30s) membrane proteins, heat shock proteins, RNA polymerases, DNA repair ATPase, hypothetical Saci, translation and transcription initiation factors, elongation factors, ABC transporters and many hypothetical proteins were located around extreme end of axis1.

Taken together our results suggest that, for all species and strains, gene expression level plays a dominant role in separating genes according to their codon usages or codon biases. Although, inertia shared by axis1 and axis2 COA for amino acid usage was comparatively very high (marked as second coordinate within brackets in column 2 of (Table 1)), correlation between axis1 and CAI was not found to be highly significant in any of these organisms at $P<0.01$ (r value between CAI and axis1 of amino acid usage: 0.290, *S. acidocaldarius*, −0.115, *S. solfataricus*; −0.131, *S. tokodaii*; −0.268, *S. islandicus* L.D.8.5; 0.263, *S. islandicus* L.S.2.15; 0.313, *S. islandicus* M.14.25; 0.319, *S. islandicus* M.16.27; 0.302, *S. islandicus* M.16.4; −0.071, *S. islandicus* Y.G.57.14; 0.208, *S. islandicus* Y.N.15.51). Thus amino acid usage was not highly influenced by gene expression level in our studied organisms.

3.3. Translational selection and optimal codons

To understand which of the triplets were selectively used among genes having high expression level (called optimal codons), 10% of the total genes (having lower Nc values) were selected from extreme ends of axis COA with extremely high and low CAI values ($0.5 \leq CAI \leq 1$). Using χ^2 test ($P<0.01$), dataset for all organisms was analyzed to know their codon preference patterns for genes at higher expression level and result was presented in Table S2 (supplementary). Optimal codons found at level of statistical significance $P<0.01$ were

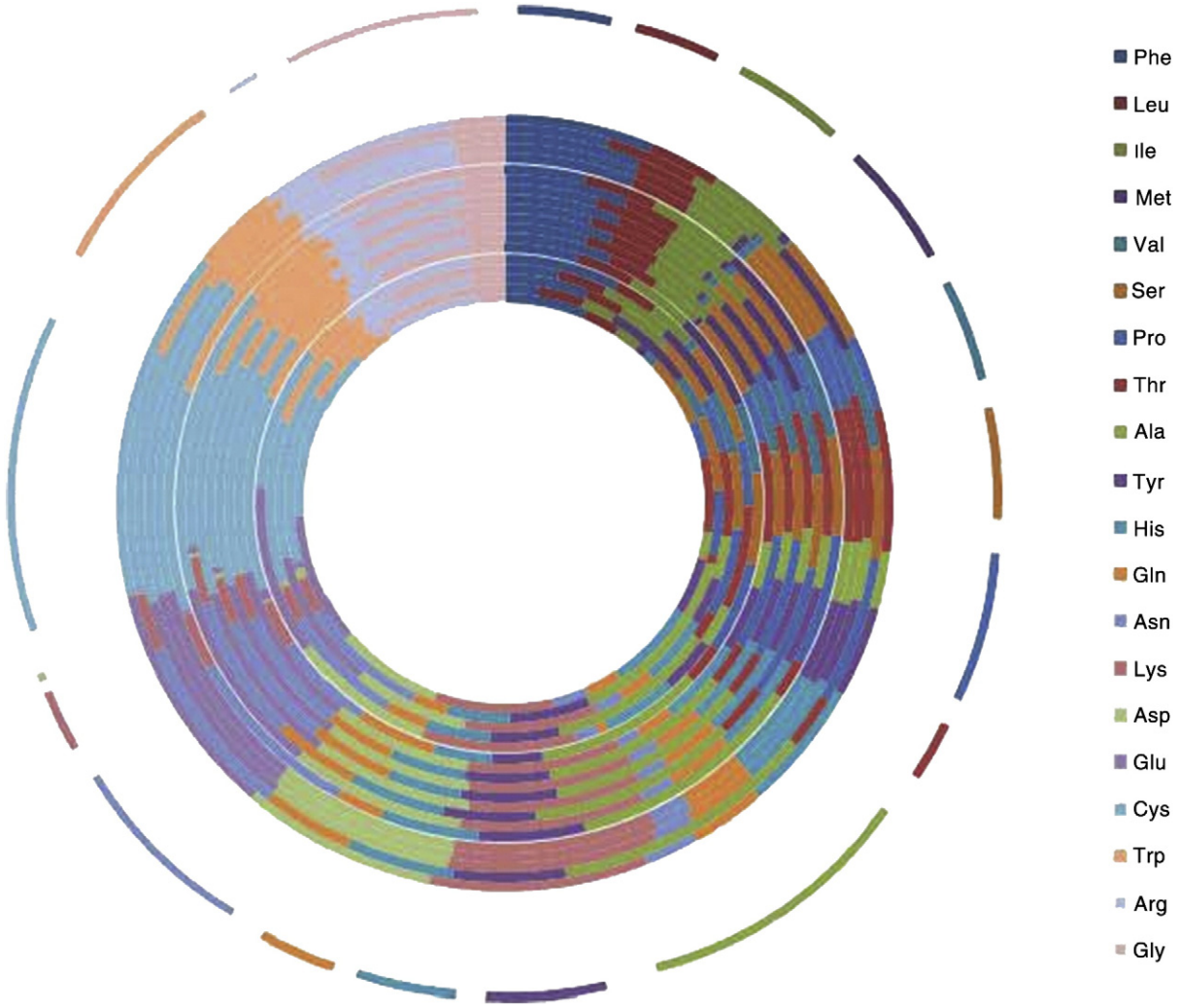

Fig. 3. The doughnut diagram of 20 amino acids representing coordinates values given by COA for first two axes of correspondence analysis of amino acid usage. The clustering of coordinates is, hence arc length is proportional to inertia shared by axis for respective amino acid usage. Color legends for 20 amino acids are as given below.

marked with ** and * for $P<0.05$. Interestingly, result showed that, for *S. acidocaldarius* all 26 optimal codons are A/U ending. Similar preference was obtained for all other organisms except in case of 3 codons (UGC**, *Sulfolobus islandicus* L.S.2.15; UGC**, *Sulfolobus islandicus* Y.G.57.14; GGC*, *Sulfolobus islandicus* Y.N.15.51). Codon preferences for ribosomal proteins (RSCU >1) was marked as bold. The result showed that most of preferable codons are translationally selected and matches with optimal codons for genes expressed at higher expression level. Furthermore, selection of A/U at third position was seen for all GC-rich amino acids, whereas, in AT-rich amino acids 1st, 2nd, 3rd, positions were mostly preferred to A/U. This preference pattern was found to be overall true in all our studied organisms (Table S2, supplementary data). Hence, genes at higher expression level were highly influenced by AT compositional bias in these organisms. Additionally significant negative correlation between CAI and GC3s values for these species and strains (all $r<-0.6$, $P<0.01$) completely supported the above observation (Table 2).

Moreover, for *S. acidocaldarius,* 11 out of 26 optimal codons match perfectly with putative most abundant isoacceptor tRNA (Table S2, supplementary) — we assume a correlation between the cellular levels of tRNAs and the copy number of tRNA genes (the anticodon table and copy number was generated by using program "tRNAscan-SE" (http://gtrnadb.ucsc.edu/)). These codons could be considered as translationally optimal (marked with # and copy numbers are given in brackets) (Table S2, supplementary), as was previously reported in *E. coli*, *Bacillus subtilis* and *Saccharomyces cerevisiae* (Dong et al., 1996; Ikemura, 1982; Kanaya et al., 1999; Percudani et al., 1997). For example, there are 6 Arg tRNA genes, with two copies from GCG and one copy from other three CCG, TCG, CCT and TCT. Here, GCG recognizes CGC and other three CCG, TCG, CCT and TCT were matched with CGG, CGA, AGG, and AGA respectively. In other cases, where the match was not perfect (as is the case for fourfold degenerate codons encoding Proline), the reason could be due to modification in first position of the anticodon. Although, the comparative analysis among other organisms shows that more than 50% of optimal codons are translationally optimal, there are some anti codons that match with non-optimal codons. In most of the cases translationally optimal codons are found to be A/U ending (Table S2, supplementary data). The table also showed many non-optimal codons having preference for translation are mostly GC rich and G/C ending, encoding for Proline, Alanine, Glycine, Tryptophan and Serine. This preference at lower expression level as compared to the higher ones may be due to presence of more abundant tRNA genes (for 4 and 6 box tRNA sets) favoring encoding of G/C ending codons.

Moreover, cysteine encoding C-ending codon, UGC, was preferentially selected in all organisms for translation, although its other synonym was found to be optimal. This selective choice of C-ending codon in AT-rich organisms may be due its vital role or ability to form disulphide bonds with other cysteins, which is considered to

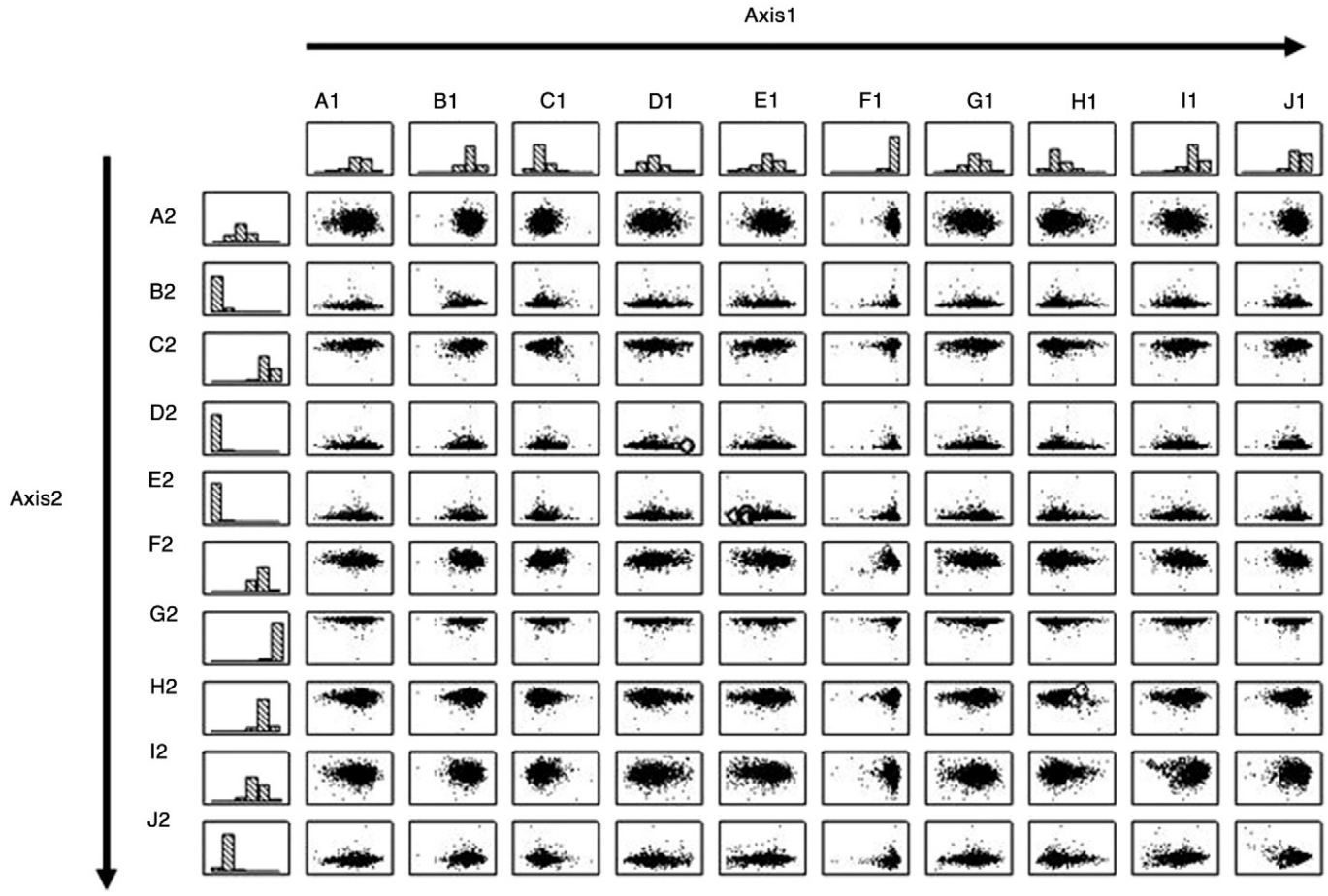

Fig. 4. Matrix plot of between axis 1(RSCU) values on X-axis and axis2 (RSCU) values on Y-axis for all organisms for *S. acidocaldarius* DSM 639 (A1 vs A2), *S. solfataricus* P2 (B1 vs B2), *Sulfolobus tokodaii* (C1 vs C2), *S. islandicus* L.D.8.5 (D1 vs D2), *S. islandicus* L.S.2.15 (E1 vs E2), *S. islandicus* M.14.25 (F1 vs F2), *S. islandicus* M.16.27 (G1 vs G2), *S. islandicus* M.16.4 (H1 vs H2), *S. islandicus* Y.G.57.14 (I1 vs I2) and *S. islandicus* Y.N.15.51 (J1 vs J2). This plot presents both distributions (histogram panel) and relationship (bubble dots) between variables.

be important for thermo stabilization. However, interestingly, except for few cases, single copy number of tRNA genes (numbers marked in brackets, Table S2, supplementary data) was found for all optimal (marked with **#) and non-optimal codons (marked with #) in the genus.

These results together suggest that, in our selected genomes, codon usage pattern of genes at higher expression level was dominantly influenced by translational selection. However, speed of translation of these optimal codons is not at very higher level, as there are few duplicate copies of tRNA genes available for translation. The possible explanation for few duplicate copies tRNA genes in all genomes may be given by arguing that, these organisms avoid mistranslation of proteins at higher expression level by lowering their translation speed. Because, it is not obvious that the fastest codon would always be most accurate, that is, there may be cases where accuracy and speed could be distinguished because they act in opposite directions (Higgs and Ran, 2008). Another reason for abundance of single copy numbers of tRNA genes across genomes may be due to co evolution of codon usage and tRNA content. These results of codon usages help to select optimal codons in codon optimization process for heterologous genes to get desired level of their protein expression levels.

3.4. Effect of other factors on codon and amino acid usage

Correlation coefficients between coordinates for genes on first axis of COA amino acid usage and GRAVY score were found to be highly significant for all organisms. Thus results taken together, it could be suggested that solubility of protein [positive GRAVY (hydrophobic), negative GRAVY (hydrophilic)] plays a dominant role in amino acid usage for organisms. Moreover, in all cases, similar variations in direction of correlation, was found for aromaticity scores at level significance $P<0.01$ and their degree of correlation was observed to be always more than 0.4 (although comparatively lower to respective GRAVY values) (Table 1), indicating its significant role in amino acid usage.

In *S. acidocaldarius* the correlation analysis among CDS length and first axis of COA for RSCU values, CAI and Nc values showed three correlations coefficients ($r=0.062$, $P<0.01$, -0.061, $P<0.01$, -0.004, not significant) respectively. These results are contrary to general tendency of more biased genes with longer CDS length towards higher expression level, as was seen in case of *E. coli* (Eyre-walker, 1996), *Pseudomonas aeruginosa* (Gupta and Ghosh, 2001) , *S. pneumoniae* (Hou and Yang, 2002) and *Yersinia pestis* (Hou and Yang, 2003). It has been previously reported that codon bias was affected by gene length to a certain degree (significant negative relationship) in eukaryotic organisms, such as in *Caenorhabditis elegans* (Marais and Duret, 2001), *Drosophila melanogaster* (Miyasaka, 2002). In this regard, Moriyama and Powell (1998) explained that if shorter proteins could perform similar functions to those of the longer ones, longer proteins become energy-expensive and disadvantageous. Thus, selection constraint acts to reduce the size of highly expressed genes, which determines a dominant relationship between codon usage bias and gene length. However, this preference of shorter length genes at higher expression levels was not found to be overall true in all studied species (r between CAI and gene length: 0.058,

$P<0.01$, *S. solfataricus*; 0.077, $P<0.01$, *S. tokodaii*; −0.013, not significant, *S. islandicus* L.D.8.5; −0.024, not significant, *S. islandicus* L.S.2.15; −0.014, not significant, *S. islandicus* M.14.25; −0.017, not significant, *S. islandicus* M.16.27; 0.009, not significant, *S. islandicus* M.16.4; 0.0, not significant, *S. islandicus* Y.G.57.14; −0.034, not significant, *S. islandicus* Y.N.15.51). These results suggest that, there were no universal rules about the relationship between codon bias and gene length in all studied genomes, and the real reason for this discrepancy is yet to be understood.

In order to know the effect of replication associated mutational pressure on protein coding sequences, cumulative GC skew was drawn using the program GraphDNA (http://www.virology.ca). "DNA walk" method was used to know the nucleotide distribution using the same program. The skew diagram (Fig. 5A) suggested that there are multiple origins of replication, because it does not show any characteristic V/inverted V shape for bi-directional replication between singular ori and ter. The result for two replication origins was reported for *S. acidocaldarius* using two dimensional gel analyses, whereas *Z*-curve

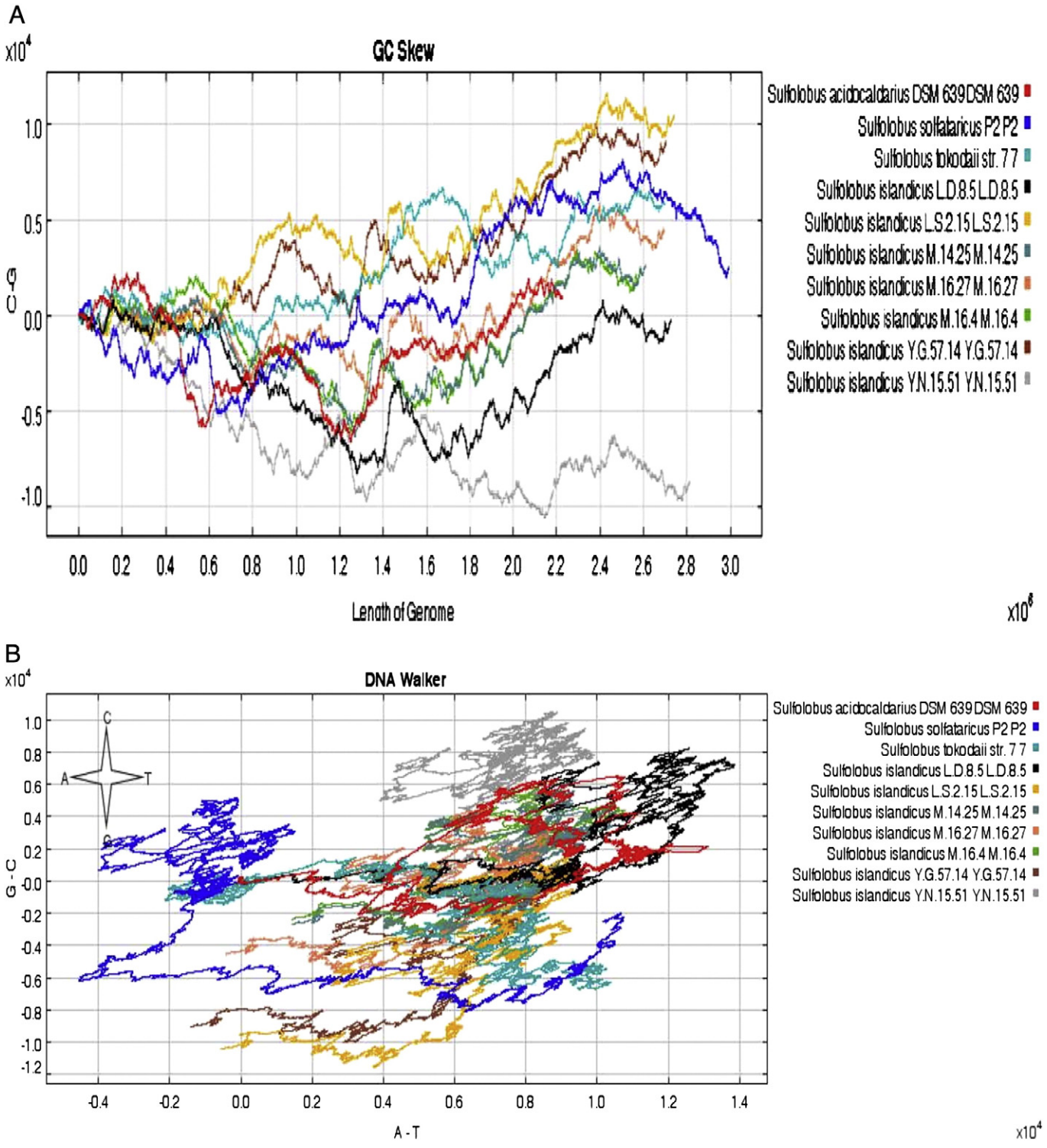

Fig. 5. A. GC skew for completed genomes *Sulfolobus* genus (color legends are at right side of the skew diagram). B. DNA walk graph for completed genomes *Sulfolobus* genus (color legends are at right side of the skew diagram).

analysis suggested its multiple origin of replication (Chen et al., 2005). Here the study showed GC skew and DNA walk (Fig. 5B) analysis results for all species of genus *Sulfolobus* with reference to above findings. These results together suggest that codon usage in all *Sulfolobus* genomes was not affected by preferences in nucleotide substitution in two strands of DNA molecule and hence no indication of role of mutational pressure on DNA asymmetry for protein coding sequences in all studied organisms. These findings were further confirmed from calculating GCSI (GC skew Index), (Arakawa and Tomita, 2007) using the G-language, GAE (publicly available from http://www.g-language.org/). GC skew index for all organisms were found either less than to 0.05 (threshold value) or marginally greater than 0.05, suggesting that mutational pressure does not play a dominant role on their DNA asymmetry.

4. Conclusion

In the current study, we analyze the role of different evolutionary constraints that influence codon and amino acid usage pattern in *S. acidocaldarius, S. solfataricus*, *S. tokodali, Sulfolobus islandis* and 6 other isolates from *islandicus* species. We find that gene expression level, compositional mutational bias, translational selection and gene length are operative for shaping codon usage variation in studied organisms. However, considering the degree of correlation, gene expression level and mutational bias dominates over all other factors. For amino acid usage, aromaticity and hydropathicity plays vital role. At the same time, the study shows that codon choices in amino acids for genes at higher expression level i.e. optimal codons are not always same as translationally optimal codons and vice versa in all organisms and this information may be helpful in design of heterologous protein expression in these organisms, because translational selection is operative for most of the non-optimal codons. Interestingly, we find some translationally preferred GC-rich codons for genes at higher expression level and genes located towards right part of the expected Nc plot curve. Since these coordinates fall below the expected curve, there are probably still other factors influencing codon usage variation in these organisms and thus, it remains open issues that need to be further studied to elucidate the determinants of codon usage pattern.

The study has given a most comprehensive analysis of codon usages patterns and has provided a basic understanding of mechanisms for codon usage bias in different species and strains of *Sulfolobus* genus, which could be useful in further study of their evolutionary mechanism, cloning and heterologous expression of its functionally important proteins.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.gene.2012.10.024.

Acknowledgments

I would like to thank the Department of Biotechnology, Ministry of Science and Technology, Govt. of India for financial assistance to carry out research program.

References

Aagaard, C., et al., 1994. A spontaneous point mutation in the single 23S rRNA gene of the thermophilic archaeon *Sulfolobus acidocaldarius* confers multiple drug resistance. J. Bacteriol. 176, 7744–7747.
Aagaard, C., et al., 1996. General vectors for archaeal hyperthermophiles: strategies based on a mobile intron and a plasmid. FEMS Microbiol. Rev. 18, 93–104.
Akashi, H., 1994. Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. Genetics 136, 927–935.
Akashi, H., Goel, P., John, A., 2007. Ancestral Inference and the Study of Codon Bias Evolution: Implications for Molecular Evolutionary Analyses of the *Drosophila melanogaster* Subgroup. PLoS One 2, 10.
Arakawa, K., Tomita, M., 2007. The GC skew index: a measure of genomic compositional asymmetry and the degree of replicational selection. Evol. Bioinformatics Online 3, 159–168.
Bell, S.D., et al., 2002. The interaction of alba, a conserved archaeal chromatin protein and its regulation by acetylation. Science 296, 148–151.
Bennetzen, J.L., Hall, B.D., 1982. Codon selection in yeast. J. Biol. Chem. 257, 3026–3031.
Brock, T.D., et al., 1972. *Sulfolobus*: a new genus of sulfur oxidising bacteria living at low pH and high temperature. Arch. Microbiol. 84, 54–68.
Bulmer, M., 1991. The selection–mutation–drift theory of synonymous codon usage. Genetics 129, 897–907.
Chen, L., et al., 2005. The genome of *Sulfolobus acidocaldarius*, a model organism of the Crenarchaeota. J. Bacteriol. 187, 4992–4999.
Dong, H., Nilsson, L., Kurland, C.G., 1996. Co-variation of tRNA abundance and codon usage in *Escherichia coli* at different growth rates. J. Mol. Biol. 260, 649–663.
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.
Duret, L., Mouchiroud, D., 1999. Expression pattern and surprisingly, gene length shape codon usage in *Caenorhabditis, Drosophlila,* and *Arabidopsis*. Proc. Natl. Acad. Sci. U. S. A. 96, 4482–4487.
Eyre-Walker, 1996. A Synonymous codon bias is related to gene length in *Escherichia coli* selection for translational accuracy. Mol. Biol. Evol. 13, 867–872.
Fujimori, S., Washio, T., Tomita, M., 2005. GC-compositional strand bias around transcription start sites in plants and fungi. BMC Genomics 6, 26.
Grantham, R., et al., 1981. Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res. 9, r43–r74.
Greenance, M.J., 1984. Theory and application of correspondence analysis. Academic Press, London, p. 223.
Grogan, D.W., Carver, G.T., Drake, J.W., 2001. Genetic fidelity under harsh conditions: analysis of spontaneous mutation in the thermoacidophilic archaeon *Sulfolobus acidocaldarius*. Proc. Natl. Acad. Sci. U. S. A. 98, 7928–7933.
Gu, W., et al., 2004. The relationship between synonymous codon usage and protein structure in *Escherichia coli* and *Homo sapiens*. Biosystems 73, 89–97.
Guo, F., Yuan, J., 2009. Codon usages of genes on chromosome, and surprisingly, genes in plasmid are primarily affected by strand-specific mutational biases in *Lawsonia intracellularis*. DNA Res. 16, 91–104.
Gupta, S.K., Ghosh, T.C., 2001. Gene expressivity is the main factor in dictating the codon usage variation among the genes in *Pseudomonas aeruginosa*. Gene 273, 63–70.
Gupta, S.K., Bhattacharyya, T.K., Ghosh, T.C., 2004. Synonymous codon usage in *Lactococcus lactis*: mutational bias versus translational selection. J. Biomol. Struct. Dyn. 21, 527–536.
Higgs, P.G., Ran, W., 2008. Coevolution of codon usage and tRNA genes leads to alternative stable states of biased codon usage. Mol. Biol. Evol. 25, 2279–2291.
Hou, Z.C., Yang, N., 2002. Analysis of factors shaping *S. pneumoniae* codon usage. Acta Genet. Sin. 29, 747–752.
Hou, Z.C., Yang, N., 2003. Factors affecting codon usage in *Yersinia pestis*. Acta Biochim. Biophys. Sin. 35, 580–586.
Ikemura, T., 1981. Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes. J. Mol. Biol. 151, 389–409.
Ikemura, T., 1982. Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. J. Mol. Biol. 158, 573–597.
Ikemura, T., 1985. Codon usage and tRNA content in unicellular and multicellular organisms. Mol. Biol. Evol. 2, 13–34.
Kanaya, S., et al., 1999. Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. Gene 238, 143–155.
Karlin, S., Mrazek, J., 1996. What drives codon choices in human genes? J. Mol. Biol. 262, 459–472.
Kawarabayahi, Y., et al., 2001. Complete genome sequence of an aerobic thermoacidophilic Crenarchaeon, *Sulfolobus tokodaii* strain7. DNA Res. 31, 123–140.
Kunst, F., et al., 1997. The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. Nature 390, 249–256.
Langer, D., et al., 1995. Transcription in archaea: similarity to that in eucarya. Proc. Natl. Acad. Sci. U. S. A. 92, 5768–5772.
Liu, H., et al., 2010. Analysis of synonymous codon usage in *Zea mays*. Mol. Biol. Rep. 37, 677–684.
Marais, G., Duret, L., 2001. Synonymous codon usage, accuracy of translation, and gene length in *Caenorhabditis elegans*. J. Mol. Evol. 52, 275–280.
McInerney, J.O., 1998. Replicational and transcriptional selection on codon usage in *Borrelia burgdorferi*. Proc. Natl. Acad. Sci. U. S. A. 95, 10698–10703.
Miyasaka, H., 2002. Translation initiation AUG context varies with codon usage bias and gene length in *Drosophila melanogaster*. J. Mol. Evol. 55, 52–64.
Moriyama, E.N., Powell, J.R., 1998. Gene length and codon usage bias in *Drosophila melanogaster, Saccharomyces cerevisiae* and *Escherichia coli*. Nucleic Acids Res. 26, 3188–3193.
Mrazek, J., Karlin, S., 1998. Strand compositional asymmetry in bacterial and large viral genomes. Proc. Natl. Acad. Sci. U. S. A. 95, 3720–3725.
Naya, H., et al., 2001. Translational selection shapes codon usage in the GC-rich genomes of *Chlamydomonas reinhardtii*. FEBS Lett. 501, 127–130.
Nayak, K.C., 2009. Mutational bias and gene expression level shape codon usage in *Thermobifida fusca* YX. In Silico Biol. 9, 337–353.
Nayak, K.C., 2011. Comparative study on factors influencing the codon and amino acid usage in *Lactobacillus sakei* 23K and 13 other *lactobacilli*. Mol. Biol. Rep. 38, 1–11.
Ohkubo, S., et al., 1987. The ribosomal protein gene cluster of *Mycoplasma capricolum*. Mol. Gen. Genet. 210, 314–322.
Peixoto, L., et al., 2003. The strength of translational selection for codon usage varies in the three replicons of *Sinorhizobium melioti*. Gene 320, 109–116.

Percudani, R., Pavesi, A., Ottonello, S., 1997. Transfer RNA gene redundancy and translational selection in *Saccharomyces cerevisiae*. J. Mol. Biol. 268, 322–330.
Precup, J., Parker, J., 1987. Missense misreading of asparagine codons as a function of codon identity and context. J. Biol. Chem. 262, 11351–11355.
Puehler, G., et al., 1989. Archaebacterial DNA-dependent RNA polymerases testify to the evolution of the eukaryotic nuclear genome. Proc. Natl. Acad. Sci. U. S. A. 86, 4569–4573.
Ranjan, A., Vidyarthi, A.S., Poddar, R., 2007. Evaluation of codon bias perspectives in phage therapy of *Mycobacterium tuberculosis* by multivariate analysis. In Silico Biol. 7, 423–431.
Reno, M., et al., 2009. Biogeography of the *Sulfolobus islandicus* pan-genome. Proc. Natl. Acad. Sci. U. S. A. 106, 8605–8610.
Robinson, M., et al., 1984. Codon usages can affect efficiency of translation of genes in *Escherichia coli*. Nucleic Acids Res. 12, 6663–6671.
Romero, H., Zavala, A., Musto, H., 2000. Codon usage in *Chlamydia trachomatis* is the result of strand-specific mutational biases and a complex pattern of selective forces. Nucleic Acids Res. 28, 2084–2090.
Romero, H., et al., 2003. The influence of translational selection on codon usage in fishes from the family Cyprinidae. Gene 317, 141–147.
Sau, K., Deb, A., 2008. Temperature influences synonymous codon and amino acid usage biases in the phages infecting extremely thermophilic prokaryotes. In Silico Biol. 9, 1–9.
Sharp, P.M., Li, W.H., 1986. An evolutionary perspective on synonymous codon usage in unicellular organisms. J. Mol. Evol. 24, 28–38.
Sharp, P.M., Li, W.H., 1987. The codon adaptation index —a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res. 15, 1281–1295.
Sharp, P.M., et al., 1988. Codon usage patterns in *Escherichia coli, Bacillus subtilis, Saccharomyces cerevisiae, Schizosaccharomyces pombe, Drosophila melanogaster* and *Homo sapiens*; a review of the considerable within-species diversity. Nucleic Acids Res. 16, 8207–8211.
Sharp, P.M., et al., 1993. Codon usage: mutational bias, translational selection, or both? Biochem. Soc. Trans. 21, 835–841.
She, Q., et al., 2001. The complete genome of the crenarchaeon *Sulfolobus solfataricus* P2. Proc. Natl. Acad. Sci. U. S. A. 98, 7835–7840.
Sorensen, M.A., Kurland, C.G., Pedersen, S., 1989. Codon usage determines translation rate in *Escherichia coli*. J. Mol. Biol. 207, 365–377.
Thomas, J.M., Horspool, D., 2007. GraphDNA: a Java program for graphical display of DNA composition analyses. BMC Bioinformatics 8, 21.
Touchon, M., et al., 2003. Transcription-coupled TA and GC strand asymmetries in the human genome. FEBS Lett. 555, 579–582.
Touchon, M., et al., 2005. Replication-associated strand asymmetries in mammalian genomes: toward detection of replication origins. Proc. Natl. Acad. Sci. U. S. A. 102, 9836–9841.
Wan, X.F., et al., 2004. Quantitative relationship between synonymous codon usage bias and GC composition across unicellular genomes. BMC Evol. Biol. 4, 4–19.
Wright, F., 1990. The "effective number of codons" used in a gene. Gene 87, 23–29.
Wright, F., Bibb, M.J., 1992. Codon usage in the G+C-rich *Streptomyces* genome. Gene 113, 55–65.
Xiufan, S., et al., 2001. Is there a close relationship between synonymous codon bias and codon–anticodon binding strength in human genes. Chin. Sci. Bull. 12, 1015–1019.
Zhao, S., et al., 2007. The factors shaping synonymous codon usage in the genome of *Burkholderia mallei*. J. Genet. Genomics 34, 362–372.
Zhou, M., Li, X., 2009. Analysis of synonymous codon usage patterns in different plant mitochondrial genomes. Mol. Biol. Rep. 36, 2039–2046.