Journal of Biomolecular Structure & Dynamics, ISSN 0739-1102
Volume 21, Issue Number 4, (2004)
©Adenine Press (2004)

Synonymous Codon Usage in *Lactococcus lactis*: Mutational Bias Versus Translational Selection

S. K. Gupta
T. K. Bhattacharyya
T. C. Ghosh*

Bioinformatics Centre
Bose Institute
P 1/12
C.I.T. Scheme VII M
Kolkata 700 054, India

http://www.jbsdonline.com

Abstract

In this study codon usage bias of all experimentally known genes of *Lactococcus lactis* has been analyzed. Since *Lactococcus lactis* is an AT rich organism, it is expected to occur A and/or T at the third position of codons and detailed analysis of overall codon usage data indicates that A and/or T ending codons are predominant in this organism. However, multivariate statistical analyses based both on codon count and on relative synonymous codon usage (RSCU) detect a large number of genes, which are supposed to be highly expressed are clustered at one end of the first major axis, while majority of the putatively lowly expressed genes are clustered at the other end of the first major axis. It was observed that in the highly expressed genes C and T ending codons are significantly higher than the lowly expressed genes and also it was observed that C ending codons are predominant in the duets of highly expressed genes, whereas the T endings codons are abundant in the quartets. Abundance of C and T ending codons in the highly expressed genes suggest that, besides, compositional biases, translational selection are also operating in shaping the codon usage variation among the genes in this organism as observed in other compositionally skewed organisms. The second major axis generated by correspondence analysis on simple codon counts differentiates the genes into two distinct groups according to their hydrophobicity values, but the same analysis computed with relative synonymous codon usage values could not discriminate the genes according to the hydropathy values. This suggests that amino acid composition exerts constraints on codon usage in this organism. On the other hand the second major axis produced by correspondence analysis on RSCU values differentiates the genes into two groups according to the synonymous codon usage for cysteine residues (rarest amino acids in this organism), which is nothing but a artifactual effect induced by the RSCU values. Other factors such as length of the genes and the positions of the genes in the leading and lagging strand of replication have practically no influence in the codon usage variation among the genes in this organism.

Key words: Synonymous codon usage, Highly expressed genes, Lowly expressed genes, Mutational bias, Translational selection, *Lactococcus lactis*, Correspondence analysis.

Introduction

Analysis of codon usage data has both practical and theoretical importance in understanding the basics of molecular biology (1-6). It is well known that synonymous codon usage bias is non-random and species specific (7). Moreover, codon usage patterns differ significantly among different genes within the same taxa (8). It has been widely accepted that compositional biases are the only dictator in shaping the codon usage variation among the genes in the extremely AT or GC rich unicellular organisms (9-11).

Codon usage variation among mammalian genes can be explained by the isochore organization of that genome (12-15). It has been suggested that translational selection determines the codon usage bias of highly expressed genes and subsequently it has been advocated that preferred codons in highly expressed genes are recognized by most abundant tRNAs (16-18). Very recently it has been observed that in

*Phone: +91-33-2334 6626
Fax: +91-33-2334 3886
Email: tapash@bic.boseinst.ernet.in

Pseudomonas aeruginosa, codon usage bias is mainly dictated by translational selection rather than the mutational biases though it is a high GC rich organism (19). In some unicellular organisms it was observed that both translational and compositional constraints are operational in dictating the codon usage variation among the genes in those organisms (16, 20-23). In *Borrelia burgdorferi* it was observed that replicational-transcriptional selection is responsible for the codon usage variation among the genes in this organism (6). Recently it has been reported that the cellular as well as the physical location of the gene products can also reflect the codon usage patterns (2, 24). It was also reported that in *Mycobacteria,* codon usage bias is dictated by the hydrophobicity of each gene (25).

Lactococcus lactis is an AT rich non-pathogenic Gram-positive bacterium and has been widely used in milk fermentation (26). The biochemistry and physiology of this microorganism generates a lot of curiosity among the biologists. In this study we have analyzed the codon usage data with all the experimentally known coding sequences with an aim to understand the genetic organization of this organism. Our results suggest that several factors are operational in dictating the codon usage variation among the genes in this organism.

Materials and Methods

The complete genomes of *Lactococcus lactis* have been downloaded from www.ncbi.nlm.nih.gov/genbank/genomes. Our own program developed in C was used to retrieve the coding sequences from the complete genome. To minimize sampling errors we have chosen only those sequences that are greater than or equal to 300 bp and have correct initial and termination codons. We have also excluded phage like and insertion element sequences from our analysis. Finally 1129 sequences were selected for data analysis.

Relative synonymous codon usage (RSCU) is defined as the ratio of the observed frequency of a codon to the expected frequency if all the synonymous codons for those amino acids are used equally (27). RSCU values greater than 1.0 indicate that the corresponding codons are used more frequently than the expected frequency whereas the reverse is true for RSCU value less than 1.0.

GC_{3s} is the frequency of (G+C) at the synonymous third positions of codons.

The effective number of codons used by a gene (N_C) is generally used to measure the bias of synonymous codons (28). The values of N_C range from 20 (when only one codon is used per amino acid) to 61 (when all codons are used in equal probability). The expected value of N_C under random codon usage is given by the following formula:

$$N_c = 2 + s + \{29/[s^2 + (1-s)^2]\};$$

Where $s = GC_{3s}$

All the parameters were calculated by using the programme CodonW 1.3 (available at www.molbiol.ox.ac.uk/cu). Correspondence analysis (CA) available in the CodonW program was used to investigate the major trend in codon usage variation among the genes (29). GC skew, defined as the ratio of (G-C) and along the DNA sequences was calculated using a sliding window of 60 kb and a step size of 6 kb.

Results and Discussion

Overall Codon Usage Analysis

Overall RSCU values of 1129 genes shown in Table I indicate that A and/or T ending codons are predominant in this organism. Since *L. lactis* is an AT rich genome

(26) it is expected that A and/or T ending codons will predominate in the coding regions of this organism. From overall RSCU values one can assume that compositional constraints are the only factor in shaping the codon usage variation among the genes in this organism. But overall RSCU values may hide some heterogeneity of codon usage bias among the genes that might be superimposed on the extreme genomic composition of a genome.

Table I

Overall codon usage data of *L. lactis* genes. RSCU represents relative synonymous codon usage values, calculated by summing over all the genes together. N is the number of codons, AA represents amino acid.

AA	Codon	N	RSCU	AA	Codon	N	RSCU
Phe	UUU	14566	(1.49)	Ser	UCU	6834	(1.57)
	UUC	4942	(0.51)		UCC	1111	(0.26)
Leu	UUA	13290	(1.92)		UCA	9440	(2.17)
	UUG	9034	(1.31)		UCG	1324	(0.30)
Tyr	UAU	10903	(1.56)	Cys	UGU	1371	(1.60)
	UAC	3031	(0.44)		UGC	347	(0.40)
ter	UAA	834	(2.22)	ter	UGA	191	(0.51)
ter	UAG	104	(0.28)	Trp	UGG	3788	(1.00)
Leu	CUU	11119	(1.61)	Pro	CCU	4910	(1.41)
	CUC	3206	(0.46)		CCC	958	(0.28)
	CUA	2756	(0.40)		CCA	6969	(2.01)
	CUG	2051	(0.30)		CCG	1048	(0.30)`
His	CAU	5627	(1.48)	Arg	CGU	7250	(2.83)
	CAC	1967	(0.52)		CGC	1747	(0.68)
Gln	CAA	12975	(1.73)		CGA	2178	(0.85)
	CAG	2028	(0.27)		CGG	894	(0.35)
Ile	AUU	23045	(2.16)	Thr	ACU	8625	(1.49)
	AUC	6400	(0.60)		ACC	2592	(0.45)
	AUA	2615	(0.24)		ACA	9481	(1.64)
Met	AUG	10715	(1.00)		ACG	2430	(0.42)
Asn	AAU	16328	(1.60)	Ser	AGU	5344	(1.23)
	AAC	4056	(0.40)		AGC	1998	(0.46)
Lys	AAA	25309	(1.71)	Arg	AGA	2889	(1.13)
	AAG	4206	(0.29)		AGG	420	(0.16)
Val	GUU	14698	(2.08)	Ala	GCU	13786	(1.69)
	GUC	4801	(0.68)		GCC	4805	(0.59)
	GUA	5211	(0.74)		GCA	10579	(1.30)
	GUG	3547	(0.50)		GCG	3433	(0.42)
Asp	GAU	16508	(1.46)	Gly	GGU	11462	(1.60)
	GAC	6060	(0.54)		GGC	3402	(0.47)
Glu	GAA	25459	(1.70)		GGA	10574	(1.47)
	GAG	4535	(0.30)		GGG	3264	(0.45)

Variation of Codon Usage Among the Genes

Two different indices, namely, effective number of codons used by gene (N_c) and (G+C) percentage at the synonymous third positions of codons (GC_{3s}) have been widely used to detect the codon usage variation among the genes. It was observed that N_c values range from 26.90 to 61.00 with a mean of 42.32 and s.d. 4.71. This indicates that there is a marked variation of codon usage in the genes of this organ-

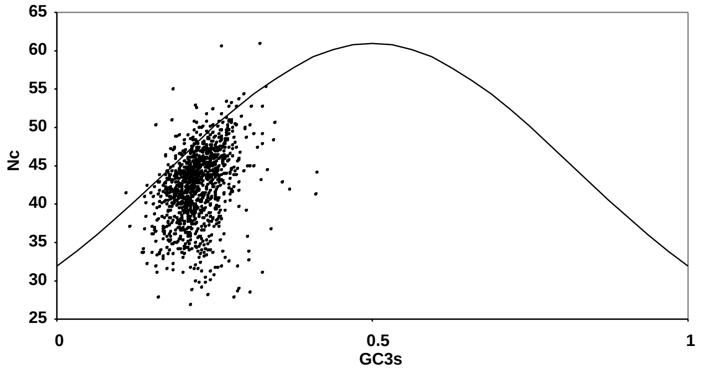

Figure 1: N_c plot of *L. lactis* genes. The continuous curve represents the expected curve between GC_{3s} and N_c under random codon usage.

ism. GC distributions at the synonymous third codon position demonstrate that GC_{3s} ranges from 11.20 to 41.10 with a mean of 22.28 and s.d. 3.48. These results suggest that apart from the compositional constraints other factors might have some influences in detecting the codon usage variation among the genes in this organism.

Various Factors in Determining the Codon Usage Variation Among the Genes in L. lactis

N_c plot (a plot of N_c versus GC_{3s}) was used to explore the codon usage variation among the genes in *L. lactis*. Wright suggested that a plot of N_c versus GC_{3s} could be used effectively to explore the codon usage variation among the genes (28). He argued that the comparison of actual distribution of genes, with the expected distribution under no selection could be indicative if codon usage bias of genes have some other influences other than compositional constraints. If the codon usage bias is completely dictated by GC_{3s} the values of N_c should fall on the expected curve between GC_{3s} and N_c. N_c plot of *L. lactis* shown in Figure 1 shows that a considerable number of points are lying on the expected curve towards the GC poor region, which certainly originates from the extreme compositional constraints. But it is also interesting to note that a majority of the points with low N_c values are lying well below the expected curve. This result suggests that a majority of genes in this organism have additional codon usage bias, which are independent of compositional constraints.

Multivariate Statistical Analysis

Multivariate statistical analysis has been widely used to study the codon usage variation among the genes in different organisms. Correspondence analysis is one of the multivariate statistical technique in which the data are plotted in a multidimensional space of 59 axes (excluding Met, Trp and stop codons) and then it determines the most prominent axes contributing the codon usage variation among the genes.

To examine if amino acid compositions exert any constraint on synonymous codon usage we have performed CA on simple codon count as well as on RSCU values. Figures 2 (a) and (b) show the positions of the genes along first and second major axes produced by CA on codon counts and on RSCU values respectively. It was

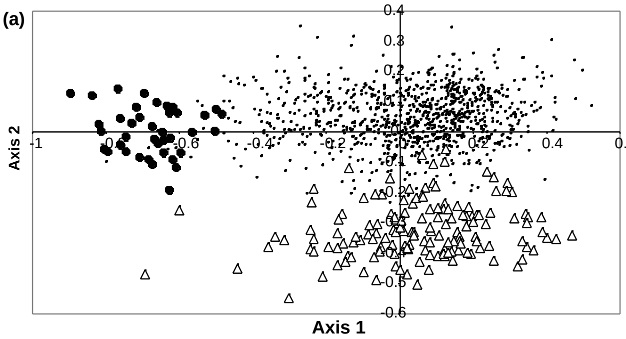

Figure 2 (a): Positions of *L. lactis* genes along the two major axes of variation in the correspondence analysis on codon count. Proteins having a gravy score >0.3 are represented as triangles; dark circles represent the highly expressed genes and other genes are represented as small dark squares.

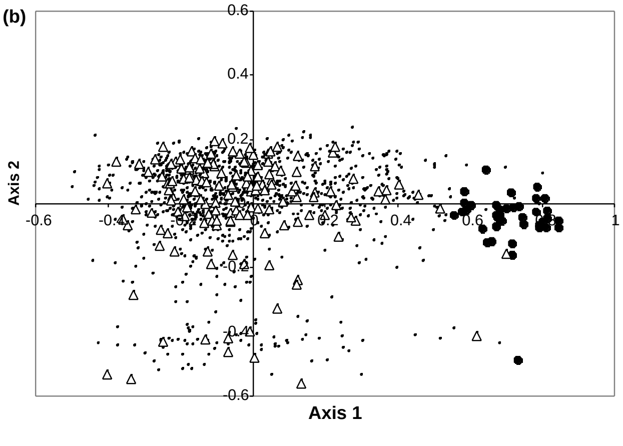

Figure 2 (b): Positions of *L. lactis* genes along the two major axes of variation in the correspondence analysis on RSCU values. Proteins having a gravy score >0.3 are represented as triangles; dark circles represent the highly expressed genes and other genes are represented as small dark squares.

observed that CA on codon count accounted for 16.90% and 7.80% of the total variation on the first and second major axes respectively whereas CA calculated on RSCU values shows 17.80% and 6.70% of the total variation on the first and second major axes respectively. Thus it can be said that in *L. lactis* genes there is a single major explanatory axis, which accounted the codon usage variation among the genes in this organism. It is interesting to note that in both the analyses (CA on codon count and on RSCU values), all the putatively highly expressed genes such as ribosomal proteins, elongation factors and outer membrane proteins are clustered on one side of first major axis and all other regulatory proteins are clustered on the other side of first major axis. The results show that gene expression levels are quite enough to discriminate genes according to their codon usage along the first major explanatory axis and amino acid compositions could not exert any constraints on this axis. We have not found any significant correlation between the positions of the genes along the first major axis produced by CA on codon count as well as on RSCU values with GC_{3s} levels. These results indicate that $(G+C)_{3s}$ levels have practically no effect in differentiating the genes according to the codon usage variation along the first major explanatory axis.

To investigate the differences between these two cluster of genes we have compared the codon usage variation between 10% of the genes located at the extreme right of axis 1 and 10% of the genes located at the extreme left of the axis 1 produced by CA on codon count. To estimate the codon usage variation between these two sets of genes we have performed chi square tests taking $P<0.01$ as significant criterion. Table II shows RSCU values for each codon for the two groups of genes. The asterisk represents the codons whose occurrences are significantly higher in the genes situated on the extreme left side of axis 1, compared to the genes present on the extreme right of the first major axis. It is important to note that out of 23 codons that are statistically over-represented in genes located on the extreme left side of axis 1 there is 8 C ending codons, 1 G ending codons and 8 T ending codons and 6 A ending codons. This actually represents 35% C ending codons, 4% G ending codons and 35% T ending codons and 26% A ending codons. It is interesting to note that 70% of the preferred triplets in the highly expressed genes are pyrimidine ending codons. A similar observation was reported in several organisms (23, 30, 31) and it was proposed that RNY codons are more advantageous for translation (32). It is interesting to note that most of the preferred C ending codons in the highly expressed genes occur in two codon family amino acids (5 out of 8), and none of the preferred T ending codons in the highly expressed genes are present in the duets indicating that C ending codons are predominant in the duets of highly expressed genes, whereas T ending codons are predominant in quartets or sextets. It was reported that NNY codons in duets are translated by a single GNN anticodon and could be advantageous for the translation of highly expressed genes (33). Moreover, it has also been shown that C ending codons are predominant in the highly expressed genes of *E.coli* (30). Very recently by analyzing the plastid genes it has been observed that in the highly expressed genes the overall biases of NNC codons are more prominent in duets (34). The predominance nature of C ending codons in the putatively highly expressed genes (particularly in duets) suggests that translational selection is also operating in codon usage variation among the genes in this organism. If the compositional constraints are the only dictator in shaping the codon usage variation among the genes then the base composition at the third position of codons in the preferred set of codons should also have A and/or T at their third codon positions.

In order to confirm our assumption that highly expressed genes are clustered along the first major axis we have calculated codon adaptation index (CAI) (36) of all the genes of *L. lactis*. CAI has been used widely to estimate the expressivities of genes by many workers and is now being considered a well-accepted measure of gene expressivities (30, 37-39, 40). CAI was calculated taking ribosomal proteins as a reference. A scatter diagram of the position of genes along the first major axis produced by CA on codon count and their corresponding CAI values was drawn (Fig.

Table II

Codon usage of highly and expressed genes of *L. lactis* genes. Superscript "a" denotes for highly expressed genes and "b" for lowly expressed genes. Asterisk represents the codons occurring significantly more often in the highly expressed genes than that of lowly expressed genes.

AA	Codon	RSCU[a]	N[a]	RSCU[b]	N[b]	AA	Codon	RSCU[a]	N[a]	RSCU[b]	N[b]
Phe	UUU	0.94	(659)	1.72	(1586)	Ser	UCU*	1.71	(557)	1.41	(543)
	UUC*	1.06	(739)	0.28	(257)		UCC	0.03	(9)	0.39	(152)
Leu	UUA	0.64	(339)	2.44	(1602)		UCA*	3.20	(1046)	1.68	(645)
	UUG*	1.76	(931)	1.02	(668)		UCG	0.09	(28)	0.38	(145)
	CUU*	2.82	(492)	1.20	(790)	Pro	CCU	1.28	(466)	1.61	(336)
	CUC*	0.64	(340)	0.36	(234)		CCC	0.04	(13)	0.42	(88)
	CUA	0.07	(35)	0.61	(398)		CCA*	2.61	(949)	1.59	(333)
	CUG	0.07	(36)	0.38	(251)		CCG	0.07	(27)	0.38	(80)
Ile	AUU*	1.70	(1482)	1.94	(1829)	Thr	ACU*	1.91	(1125)	1.44	(529)
	AUC*	1.29	(1124)	0.42	(400)		ACC	0.15	(86)	0.50	(184)
	AUA	0.01	(10)	0.64	(604)		ACA*	1.85	(1088)	1.54	(568)
Met	AUG	1.00	(1033)	1.00	(681)		ACG	0.09	(55)	0.52	(193)
Val	GUU*	2.50	(1976)	1.77	(741)	Ala	GCU*	1.94	(1820)	1.69	(688)
	GUC	0.41	(323)	0.70	(295)		GCC	0.38	(358)	0.53	(218)
	GUA	0.87	(688)	0.88	(370)		GCA	1.40	(1314)	1.29	(527)
	GUG	0.22	(174)	0.65	(272)		GCG	0.29	(268)	0.49	(198)
Tyr	UAU	1.04	(559)	1.73	(1124)	Cys	UGU	1.65	(85)	1.56	(156)
	UAC*	0.96	(519)	0.27	(174)		UGC	0.35	(18)	0.44	(44)
TER	UAA	2.92	(109)	1.71	(64)	TER	UGA	0.03	(1)	0.78	(29)
	UAG	0.05	(2)	0.51	(19)	Trp	UGG	1.00	(297)	1.00	(294)
His	CAU	0.97	(343)	1.64	(462)	Arg	CGU*	4.95	(1456)	1.14	(212)
	CAC*	1.03	(366)	0.36	(102)		CGC*	0.82	(240)	0.40	(75)
Gln	CAA*	1.97	(1158)	1.55	(1016)		CGA	0.10	(28)	1.20	(223)
	CAG	0.03	(18)	0.45	(292)		CGG	0.03	(09)	0.39	(72)
Asn	AAU	0.98	(835)	1.72	(1523)	Ser	AGU	0.44	(145)	1.70	(656)
	AAC*	1.02	(873)	0.28	(245)		AGC	0.54	(175)	0.44	(168)
Lys	AAA*	1.87	(2564)	1.60	(2049)	Arg	AGA	0.11	(31)	2.37	(441)
	AAG	0.13	(185)	0.40	(512)		AGG	0.00	(1)	0.50	(94)
Asp	GAU	1.23	(1448)	1.60	(1217)	Gly	GGU*	2.43	(1992)	1.14	(458)
	GAC*	0.77	(903)	0.40	(309)		GGC	0.32	(260)	0.45	(179)
Glu	GAA*	1.90	(2888)	1.52	(1764)		GGA	1.06	(869)	1.73	(694)
	GAG	0.10	(149)	0.48	(551)		GGG	0.19	(159)	0.68	(271)

3) and it is interesting to note that there is a strong negative correlation (r = -0.950, P<0.001) between the positions of the genes along the first major axis and their corresponding CAI values, confirming that axis 1 is strongly correlated with the expression level of each sequence of *L. lactis*. We got very similar results when we performed the CA analysis on RSCU values. These results indicate the compositional constraints have no effect on the first major explanatory axis, but on the other hand expression levels of genes are the main player in dictating the codon usage variation among the genes in this organism.

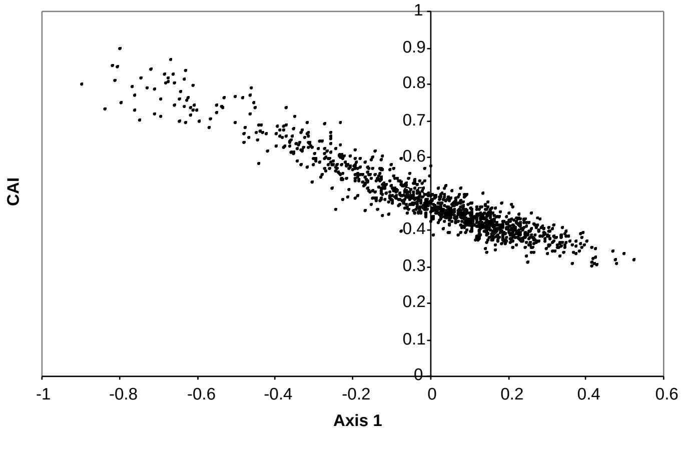

Figure 3: The scatter diagram of the *L. lactis* genes on the first major axis generated by correspondence analysis on codon count against their CAI values.

The position of the genes along the second major axis produced by CA on codon count is significantly negatively correlated with GC_{3s} (r = -0.241, P<0.01) and it is also interesting to note that the positions of genes along the second major axis are separated according to the hydrophobicity of the genes. These results are nothing but the superimposition of amino acid bias on codon usage bias. However, if we use RSCU values to compute CA, second major axis could not discriminate the genes according to the hydrophobicity values (shown in Fig 1(b)). These results suggest that to minimize the effect of amino acid composition to compute CA on RSCU values diminishes the quantity of information, as earlier observed in *Bacillus subtilis* (41). To see how the different codons are contributing towards codon usage variation among the genes in second major axis we have plotted the distribution of codons on the first two major axes both from CA on codon count and CA on RSCU values, which are shown in Figure 4 (a) and (b). From Figure 4 (a) it is obvious that all the codons are almost equally contributing towards the codon usage variation among the genes in second major axis, whereas Figure 4 (b) indicates that codons UGU and UGC (synonymous codons for cysteine) are extremely high but opposite in magnitude along the second axis. Recently it has been reported that synonymous codon usage of cysteine has a profound effect on the codon usage variation among the genes in *Thermotoga maritime* (42) and predicted that the amino acids whose occurrences are very rare in a non-skewed organism should have such impact on codon usage variation. It was also (42) argued that the absence of codon usage variation due to cysteine codons in *Ureaplasma urealitycum* and *Borrelia burgdorferi* is due to the effect of an extreme GC contents superimposed on a low frequency of Cys residues. Recently it has been demonstrated that a linkage between a particular amino acid (particularly the rarest amino acid) to the codon usage bias is nothing but an artifactual effect induced by the use of relative frequencies of codons. Codon uage variations due to cysteine codons along the second major axis in highly skewed *Lactococcus lactis* and also in case of *Thermotoga maritime* (42) are nothing but the introduction of another bias associated with the rarest amino acids in order to remove amino acid biases.

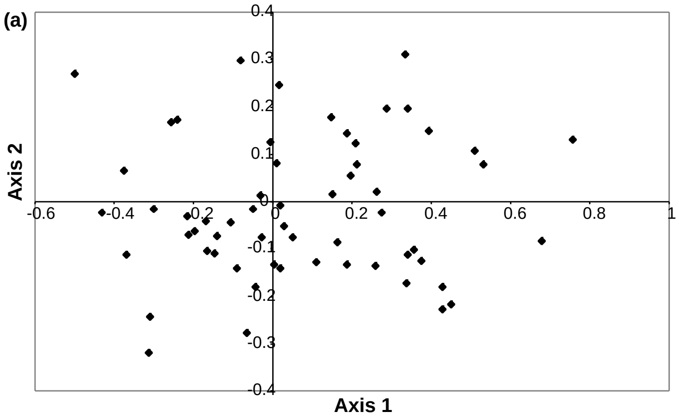

Figure 4 (**a**): Distribution of synonymous codons along the first and second major axes of the correspondence analysis on codon count.

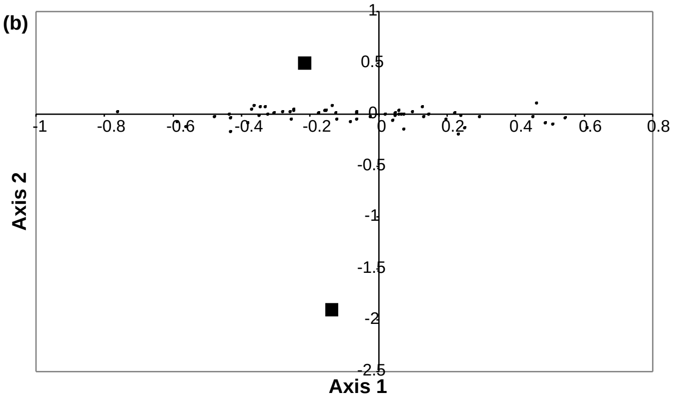

Figure 4 (**b**): Distribution of synonymous codons along the first and second major axes of the correspondence analysis on RSCU values. Large dark squares represent the synonymous codons for cysteine residue. The upper one is for UGU and the lower one for UGC.

We have also explored if the codon usage variation has any effect according to the position of each gene in the leading or in the lagging strand. GC skew was used to locate the leading and lagging strand in *L. lactis* genome. The distribution of genes on both leading and lagging strands are almost eqifrequent along the first major axis and also GT_3 levels are similar in both leading and lagging strand. These results suggest that the leading or lagging strand does not have any effect on codon usage variation among the genes in this organism.

In conclusion it can be said that though codon usage of *L. lactis* is determined by compositional constraints, translational selection is also operating in shaping the codon usage variation among the genes in this organism and it was also observed that C-ending codons are always preferred in duets whereas T ending codons are preferred in quartets of highly expressed genes. Hydrophobicity of genes is the second major factor in differentiating the codon usage variation among the genes in this organism. Length of the genes and the positions of the genes in the leading and lagging strand of replication have practically no influence in the codon usage variation among the genes in this organism.

Acknowledgements

Authors are thankful to the Department of Biotechnology, Government of India, for the financial help. We are also thankful to the anonymous referees for their helpful comments.

References and Footnotes

1. J. W. Fickett. *Nucl. Acids Res. 10*, 5303-5318 (1982).
2. H. Chiapello, E. Ollivier, C. Landes-Devauchelle, P. Nitschke, and J.-L. Risler. *Nucl. Acids Res. 27*, 2848-2851 (1999).
3. A. Martin, J. Bertranpetit, J. L.Oliver, and J. R. Medina. *Nucl. Acids Res. 17*, 6181-6189 (1989).
4. A. T. Lloyd and P. M. Sharp. *Nucl. Acids Res. 20*, 5289-5295 (1992).
5. E. N. Moriyama and D. L. Hartl. *Genetics 134*, 847-858 (1993).
6. J. O. McInerney. *Proc. Natl. Acad. Sci. 95*, 10698-10703 (1998).
7. R. Grantham, C. Gautier, M. Gouy, M. Jacobzone, and R. Mercier. *Nucl. Acids Res. 9*, r43-r74 (1981).
8. K. Wada, S. Aota, R. Tsuchiya, F. Ishibashi, T. Gojobori, and T. Ikemura. *Nucl. Acids Res. 18* (Suppl.), 2367-2411 (1990).
9. S. Ohkubo, A. Muto, Y. Kawauchi, F. Yamao, and S. Osawa. *Mol. Gen. Genet. 210*, 314-322 (1987).
10. F. Wright and M. J. Bibb. *Gene 113*, 55-65 (1992).
11. S. K.Gupta. T. K. Bhattacharyya, and T. C.Ghosh. *Ind. J. Biochem. & Biophys. 39*, 35-48 (2002).
12. G. Bernardi. *Ann. Rev. Genet. 29*, 445-476 (1995).
13. S. Aota, and T. Ikemura. *Nucl. Acids Res. 14*, 6345-6355 (1986).
14. G. Bernardi, B. Olofsson, J. Filipski, M. Zerial, J. Salinas, G. Cuny, M. Meunier-Rotival, and F. Rodier. *Science 228*, 953-956 (1985).
15. H. Musto, H. Romero, A. Zavala, and G. Bernardi. *J. Mol. Evol. 49*, 325-329 (1999).
16. T. Ikemura. *J. Mol. Biol. 146*, 1-21 (1981).
17. T. Ikemura. *J. Mol. Biol. 158*, 573-587 (1982).
18. J. L. Bennetzen and B. D. Hal. *J. Biol. Chem. 257*, 3026-3031 (1982).
19. S. K. Gupta and T. C. Ghosh. *Gene 273*, 63-70 (2001).
20. M. Gouy and C. Gautier. *Nucl. Acids Res. 10*, 7055-7074 (1982).
21. S. G. Andersson and P. M. Sharp. *Microbiology 142*, 915-925 (1996).
22. H. Romero, A. Zavala, and H. Musto. *Gene 242*, 307-31 (2000).
23. T. C. Ghosh, S. K. Gupta, and S. Majumdar. *Int. J. Parasitol. 30*, 715-722 (2000).
24. A. R. Kerr, J. F. Peden, and P. M. Sharp. *Mol. Microbiol. 25*, 1177-1179 (1997).
25. A. B. de Miranda, F. Alvarez-Valin, K. Jabbari, W. M. Degrave, and G. Bernardi. *J. Mol. Evol. 50*, 45-55 (2000).
26. A. Bolotin, P. Wincker, S. Mauger, O. Jaillon, K. Malarme, J. Weissenbach, S. D. Ehrlich, and A. Sorokin. *Genome Res. 11*, 731-53 (2001).
27. P. M. Sharp, and W-H. Li. *Nucl. Acids Res. 14*, 7737-7749 (1986).
28. F. Wright. *Gene 87*, 23-29 (1990).
29. M. J. Greenacre. London: Academic press (1984).
30. G. Gutierrez, L. Marquez, and A. Marin. *Nucl. Acids Res. 24*, 2525-2527 (1996).
31. H. Musto, H. Romero, A. Zavala, K. Jabbari, and G. Bernardi. *J. Mol. Evol. 49*, 27-35 (1999).
32. J. C.Shepherd. *Proc. Natl. Acad. Sci. USA 78*, 1596-1600 (1981).
33. S. Osawa, T. Jukes, K. Watanabe, and A. Muto. *Microbiol. Rev. 56*, 229-264 (1992).
34. B. R. Morton and B. G. So. *J. Mol. Evol. 50*, 184-193 (2000).

35. E. N. Moriyama and J. R. Powell. *Nucl. Acids Res. 26*, 3188-3193 (1998).
36. P. M. Sharp and W-H. Li. *Nucl. Acids Res. 15*, 81-1295 (1987).
37. Y. Nakamura and S. Tabata. *Microbiol. Comp. Genomics 2*, 299-312 (1997).
38. A. Pan, C. Dutta, and J. Das. *Gene 215*, 405-413 (1998).
39. E. R. Tiller and R. A. Collins. *J. Mol. Evol. 50*, 249-257 (2000).
40. S. K.Gupta. T. K. Bhattacharyya, and T. C.Ghosh. *Biochem Biophys Res Commun 269*, 2-696 (2000).
41. G. Perriere and J. Thioulouse. *Nucl. Acids Res. 30*, 4548-4555 (2002).
42. A. Zavala, H. Naya, H. Romero, and H. Musto. *J. Mol. Evol. 54*, 563-568 (2002).

Date Received: August 4, 2003

Communicated by the Editor Ramaswamy H Sarma