J Mol Evol (2006) 62:645–658
DOI: 10.1007/s00239-005-0164-6

JOURNAL OF MOLECULAR EVOLUTION

Evolutionary Constraints on Codon and Amino Acid Usage in Two Strains of Human Pathogenic Actinobacteria *Tropheryma whipplei*

Sabyasachi Das, Sandip Paul, Chitra Dutta

Bioinformatics Centre, Indian Institute of Chemical Biology, 4 Raja S. C. Mullick Road, Kolkata 700 032, India

Received: 4 July 2005 / Accepted: 20 December 2005 [*Reviewing Editor:* Dr. Richard Kliman]

Abstract. The factors governing codon and amino acid usages in the predicted protein-coding sequences of *Tropheryma whipplei* TW08/27 and Twist genomes have been analyzed. Multivariate analysis identifies the replicational-transcriptional selection coupled with DNA strand-specific asymmetric mutational bias as a major driving force behind the significant interstrand variations in synonymous codon usage patterns in *T. whipplei* genes, while a residual intrastrand synonymous codon bias is imparted by a selection force operating at the level of translation. The strand-specific mutational pressure has little influence on the amino acid usage, for which the mean hydropathy level and aromaticity are the major sources of variation, both having nearly equal impact. In spite of the intracellular lifestyle, the amino acid usage in highly expressed gene products of *T. whipplei* follows the cost-minimization hypothesis. The products of the highly expressed genes of these relatively A + T-rich actinobacteria prefer to use the residues encoded by GC-rich codons, probably due to greater conservation of a GC-rich ancestral state in the highly expressed genes, as suggested by the lower values of the rate of nonsynonymous divergences between orthologous sequences of highly expressed genes from the two strains of *T. whipplei*. Both the genomes under study are characterized by the presence of two distinct groups of membrane-associated genes, products of which exhibit significant differences in primary and potential secondary structures as well as in the propensity of protein disorder.

Correspondence to: Chitra Dutta; *email:* cdutta@iicb.res.in, chitradutta@hotmail.com

Key words: Codon usage — Amino acid usage — Multivariate analysis — Replicational-transcriptional selection — Nucleotide substitution — Translational selection — Protein secondary structure — Protein disorder

Introduction

Whipple's disease is a rare multisystemic bacterial infection, involving the gastrointestinal tract as well as various other organs (Bentley et al. 2003). Symptoms include chronic diarrhea, intestinal bleeding, abdominal pain, weight loss, fatigue, and weakness (Ratnaike 2000; Dutly and Altwegg 2001). Cardiac or neurological problems, not always associated with digestive symptoms, have also been observed (Slowik and Szczudlik 2002). Patients are reported to have subtle immune defects and there are evidences that such altered immune functions play an important role in the manifestation of the disease (Ring et al. 2003).

The causative agent of Whipple's disease is an intracellular pathogenic actinobacteria *Tropheryma whipplei*, which localizes in the intestinal lamina propria of the host (Bentley et al. 2003). The bacterium is resistant to cultivation in vitro and seems to depend heavily on human cell-associated factors for growth. The complete genome sequences of two different strains of this human pathogen, *T. whipplei* TW08/27 and *T. whipplei* Twist, reveal several atypical characteristics of the organism (Bentley et al. 2003; Raoult et al. 2003). First, their genome sizes are small (< 1 Mb) and their G + C contents (∼46%) are low, compared to other actinomycete species, which have, in general, genome sizes ranging from 1 million to 8 million bp and high G + C contents (Casjens 1998). Second, the genomes bear all the traits of strictly host-adapted organisms, including pronounced deficiencies in energy metabolism and a lack of key biosynthetic pathways, indicating that like many other obligatory intracellular bacteria (Bentley et al. 2003; Raoult et al. 2003), *T. whipplei* also might have suffered gradual decay and loss of genes. Third, despite its small genome size, the bacterium devotes a large amount of coding capacity to biosynthesis of surface-associated features. It codes for a family of large surface proteins, some of which are associated with large amounts of noncoding repetitive DNA, unlike anything previously seen in bacterial genome sequences, and the genome exhibits a great deal of genetic variability, including phase variation and mostly directed toward the changes in cell surface proteins, indicating that immune evasion and host interaction play an important role in the lifestyle of this persistent bacterial pathogen (Bentley et al. 2003; Raoult et al. 2003). Such atypical characteristics of *T. whipplei* point toward the possible existence of special selection forces operative at the genome and/or proteome levels, unveiling of which calls for an in-depth analysis of the trends in codon and amino acid usage in the organism.

Synonymous codon usage in most of the unicellular organisms is primarily governed by directional mutational bias and translational selection (Sharp and Li 1987a; Pan et al. 1998; Zavala et al. 2002; Rocha 2004), though several other factors, such as context dependence (Berg and Silva 1997), replicational-transcriptional selection (McInerney 1998; Lafay et al. 1999; Romero et al. 2000; Das et al. 2005a), protein hydropathy (de Miranda et al. 2000), and ecological niches (Singer and Hickey 2003),may also have significant influences. However, in some bacterial genomes with an extremely high/low G + C-content (Ohkubo et al. 1987; Ohama et al. 1990; Wright and Bibb 1992; Andersson and Sharp 1996), no impact of translational selection is apparent on synonymous codon preferences and there are organisms like *Helicobacter pylori* where the codon usage is apparently influenced neither by simple mutational bias nor by translational selection (Lafay et al. 2000). The amino acid usage in microbial organisms may also be influenced by several factors like hydrophobicity, expressivity, and aromaticity of the respective proteins (Lobry and Gautier 1994) and cost minimization and conservation of GC-rich amino acids in highly expressed gene products etc. (Seligmann 2003; Das et al. 2005a, b; Chanda et al. 2005).

In some species, the mutational bias is markedly strand-specific (McInerney 1998; Lafay et al. 1999; Romero et al. 2000; Das et al. 2005a), and such organisms, in general, exhibit significantly different synonymous and/or nonsynonymous codon usage patterns in genes transcribed on the leading and lagging strands of replication. Replicational-transcriptional selection is a major source of variation in codon usage in those species—replicational selection is responsible for the higher number of genes on the leading strand and transcriptional selection for the enrichment of highly expressed and/or essential genes on that strand (Lopez and Philippe 2001; Das et al. 2005a; Rocha and Danchin 2003). In *Bacillus subtilis*, 78% of highly expressed genes and 94% of essential genes are preferentially positioned in the leading strand (Rocha and Danchin 2003). Interestingly enough, the organisms with strong strand-specific asymmetries studied so far are characterized by massive genome reduction and overall A + T richness, irrespective of their taxonomic positions. Examples include the spirochaetes *B. burgdorferi* and *T. pallidum* (McInerney 1998; Lafay et al. 1999), *Chlamydia trachomatis* among the Chlamydiae (Romero et al. 2000), the γ proteobacterium *Buchnera aphidicola* (Rispe et al. 2004), and the α proteobacteria *Bartonella* (Das et al. 2005a). However, not all A + T-rich reduced genomes exhibit pronounced compositional asymmetries in leading and lagging strands of replication. *H. pylori* (Lafay et al. 2000), *Wigglesworthia* (Herbeck et al. 2003), and *Wolbachia* (Wu et al. 2004) are the prime examples. The facts that *T. whipplei* is the only known reduced genome species (< 1 Mb) within the Actinobacteria and that the organization of the TW08/27 and Twist genomes differs by an inversion of a large chromosomal segment, symmetrical to the origin of replication (Raoult et al. 2003), have made it more intriguing to examine whether the reduction in genome size and/or the intracellular lifestyles of *T. whipplei* could exert any effect on its codon/amino acid preferences. Multivariate analyses carried out in the present study indicate that the codon and amino acid usage in this human pathogen might be a consequence of a complex balance among replicational-transcriptional selection, translational control, and other physicochemical properties of the gene products. The study, apart from providing an insight into the underlying selection pressures operative at the gene/protein level of *T. whipplei*, may also offer a better understanding of evolution of this host-adapted microorganism.

Materials and Methods

Actinobacterial Genome Sequences

All protein-coding sequences of the *T. whipplei* TW08/27 and Twist genomes were extracted from NCBI GenBank. Protein-coding sequences of 15 other actinobacterial species, complete genome sequences of which are available in NCBI GenBank (Version 145.0), were also extracted (listed in Table 1) for comparative analyses. In order to reduce sampling errors, the annotated genes with fewer than 100 codons were excluded from the analysis. The presumed duplicates, genes for transposase and integrase, and genes with internal stop codons and untranslatable codons were also excluded. Finally, 729 sequences for *T. whipplei* TW08/27 and 734 sequences for *T. whipplei* Twist were selected for analysis.

Determination of Leading and Lagging Strand Genes

The genes present in the leading and lagging strand were isolated on the basis of the reported location of oriC in *T. whipplei* Twist by Raoult et al. (2003). Based on the change in AT-skew signal using Oriloc (available at http://pbil.univ-lyon1.fr/software/Oriloc) (Frank and Lobry 2000) and the conservation of the dnaA-dnaN-recF gene cluster, the oriC region in *T. whipplei* TW08/27 is assumed to be located at 0 kB. The terminus is assumed at the second inflexion in AT-skew and thus allowed us to locate each coding sequence present in either the leading or the lagging strands of replication.

Multivariate Analyses of Relative Frequencies of Synonymous Codons and Amino Acids

To identify the major factors shaping variation in relative synonymous codon usage (RSCU) and relative amino acid usage (RAAU) among *T. whipplei* genes, we applied correspondence analyses (COA) using CODONW 1.4.2 (written by John Peden and available at www.molbiol.ox.ac.uk/win95.codonW.zip). For COA on RSCU values, the data were plotted in a multidimensional space of 59 axes (excluding AUG, UGG, and three stop codons) and then the most prominent axes contributing to the codon usage variation among the genes were determined. Similarly, in COA on RAAU, the variables were the frequencies of the 20 amino acids, which sum up to 1 for each gene.

Indexes Used to Identify the Trends of Variation in Codon and Amino Acid Usage

GC_{1+2} (G + C content at first and second codon positions), GC_{3S}, and GT_{3S} (frequency of G + C and G + T, respectively, at synonymous third codon positions) were calculated for each coding sequences in both strains of *T. whipplei*. Indexes like the total number of occurrences of each codon, RSCU (Sharp and Li 1987a), codon adaptive index (CAI; Sharp and Li 1987a), RAAU, average hydrophobicity (Gravy score) (Kyte and Doolittle 1982), aromaticity (Lobry and Gautier 1994), and average size/complexity quotient (Dufton 1997) of encoded proteins were also calculated to find out the factors influencing codon and amino acid usage. The CAI was calculated for all leading strand genes with respect to the RSCU values of the genes for ribosomal proteins (≥100 aa) present in that strand.

Prediction of Secondary Structure and Protein Disorder

Prediction of protein secondary structure was performed using the GOR IV algorithm (Garnier et al. 1996) from Expasy proteomics server and the disordered regions within proteins were predicted using GlobPlot (available at http://globplot.embl.de).

Estimation of Synonymous and Nonsynonymous Substitution Patterns in Highly and Lowly Expressed Genes

Orthologous genes of *T. whipplei* TW08/27 and *T. whipplei* Twist were obtained from NCBI GenePlot by filtering the symmetrical best hits of protein homologues and sequence alignments were performed using ClustalW. To examine the nucleotide substitution patterns, we estimated pairwise synonymous divergences (d_S) as well as nonsynonymous divergences (d_N) between the orthologues using the MEGA program (version 2.1), as described by Nei and Gojobori (1986).

Statistical Methods

In order to detect the significant differences between the two classes of genes (i.e., leading vs. lagging strand genes and highly vs. lowly expressed genes), if any, codon and amino acid abundances were compared by chi-square in 2 × 2 contingency tables having 1 degree of freedom. For each codon, the first and second row of the contingency table represented, respectively, the number of occurrences of the codon being analyzed and the total number of alternate synonyms for the corresponding residue in the two classes under comparison. For amino acid, rows in the contingency table were the count of a particular amino acid residue and the total count of the remaining residues. Linear regression analysis was used to find out the significance of association between the positions of sequences along major axes of COA and biological variables using STATISTICA (Version 6.0).

Results and Discussion

Asymmetrical Mutational Bias, Coupled with Replicational-Transcriptional Selection on Synonymous Codon Usage

Figures 1a and b show the position of the genes on the plane defined by the first and second major axes generated by COA on RSCU values of coding sequences in the TW08/27 and *Twist* genomes, respectively. The first principal axis accounts for 9.23% and 9.19%, while the second axis accounts for 5.15% and 5.07% of the total variability for TW08/27 and *Twist* genomes, respectively. The low percentage of variance explained by the first and second major axes indicates that no single axis accounts for much of the total variability in codon usage of *T. whipplei.*

Interestingly enough, in each plot, the genes transcribing in the leading and the lagging strands of replication are segregated in two discrete clusters with little overlap along axis 1. Similar scatter plots with two distinct clusters of points were observed earlier in case of genomes with pronounced strand-specific mutational bias (McInerney 1998; Lafay et al. 1999; Das et al. 2005a). When the positions of all synonymous codons are plotted on the plane defined by the first and second axes of COA on RSCU, most of the U- and G- ending codons are found to be distributed on the negative side of axis 1, whereas A- and C-ending codons are mostly located on the positive side

Table 1. Genome characteristics and correlations of GC_3 and GT_3 content with first two axes of COA on RSCU values of genes in 17 completely sequenced actinobacterial genomes

							Variation explained by COA on RSCU (%)		Correlation (r)			
							Axis 1	Axis 2	Axis 1		Axis 2	
Organism with strain	Accession No.	Size (Mb)	GC (%)	No. of tRNA genes	No. of rRNA operons	ORFs under study			GC_3	GT_3	GC_3	GT_3
Bifidobacterium longum NCC2705	AE014295	2.26	60.0	56	4	1641	24.85	6.52	**0.92**	−0.36	−0.24	0.19
Corynebacterium diphtheriae NCTC13129	BX248353	2.49	53.5	52	5	1952	23.10	12.24	**0.64**	**−0.74**	**0.55**	**0.55**
Corynebacterium efficiens YS-314	BA000035	3.15	63.1	56	5	2556	18.37	8.26	**0.86**	**−0.49**	−0.26	−0.32
Corynebacterium glutamicum ATCC13032	BA000036	3.31	53.8	60	6	2696	20.16	10.34	**−0.59**	**0.68**	0.07	**0.56**
Leifsonia xyli CTCB07	AE016822	2.58	67.7	45	1	1729	13.03	6.02	**−0.85**	0.29	0.17	**0.79**
Mycobacterium avium k10	AE016958	4.83	69.3	45	1	4134	15.50	5.01	**−0.90**	0.07	0.04	0.15
Mycobacterium bovis AF2122/97	BX248333	4.35	65.6	45	1	3526	9.02	5.62	**−0.87**	0.20	0.13	**0.63**
Mycobacterium leprae TN	AL450380	3.27	57.8	45	1	2418	7.23	6.26	**−0.74**	0.21	0.01	**0.78**
Mycobacterium tuberculosis CDC1551	AE000516	4.40	65.6	45	1	3620	9.20	5.71	**0.87**	−0.21	−0.13	**−0.67**
Mycobacterium tuberculosis H37Rv	AL123456	4.41	65.6	45	1	3642	8.54	5.66	**−0.86**	0.25	−0.16	**−0.64**
Nocardia farcinica IFM 10152	AP006618	6.29	70.7	53	3	5169	12.80	6.54	**0.81**	−0.24	−0.12	**−0.82**
Propionibacterium acnes KPA171202	AE017283	2.56	60.0	45	3	2097	13.06	7.63	**−0.82**	0.38	0.10	**0.81**
Streptomyces avermitilis MA-4680	BA000030	9.12	72.0	68	6	6944	11.31	6.65	**−0.78**	0.31	0.17	**0.83**
Streptomyces coelicolor A3(2)	AL645882	9.05	72.1	63	6	7183	14.19	6.76	**−0.79**	0.34	0.12	**0.82**
Symbiobacterium thermophilum IAM14863	AP006840	3.57	68.7	97	6	2956	19.15	4.71	**−0.88**	0.11	−0.04	−0.11
Tropheryma whipplei TW08/27	BX072543	0.93	46.3	51	1	729	9.23	5.15	−0.24	**−0.82**	0.01	−0.08
Tropheryma whipplei Twist	AE014184	0.93	46.0	50	1	734	9.19	5.07	−0.20	**−0.81**	0.10	0.06

Note. Correlations significant at $p < 0.0001$ are indicated in boldface.

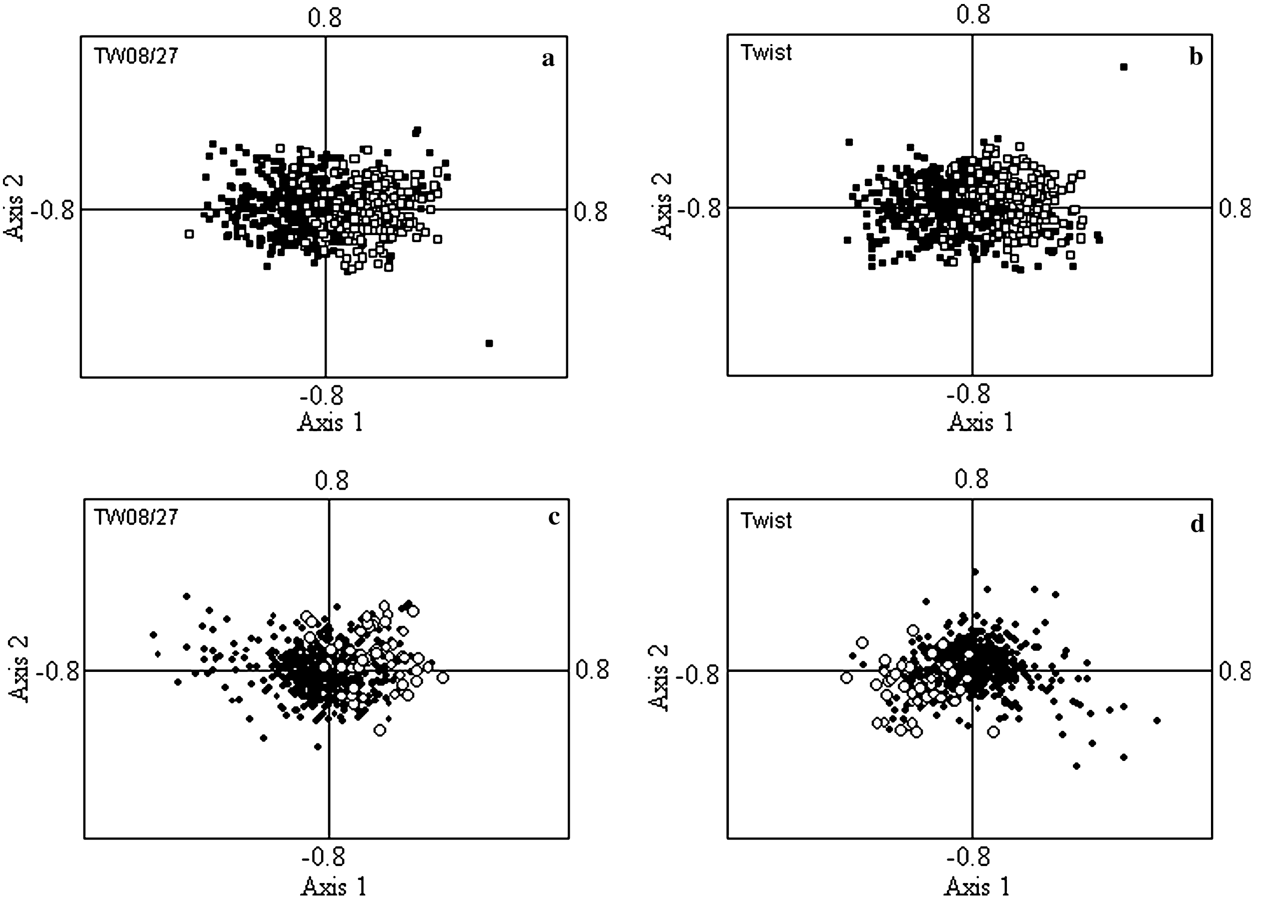

Fig. 1. Position of genes along the first two principal axes generated by COA on RSCU values of all genes from the (**a**) *T. whipplei* TW08/27 and (**b**) *T. whipplei* Twist genomes represented in the upper panel, and plots of the first two axes of COA on RSCU values of the genes located in the leading strand of the (**c**) TW08/27 and (**d**) Twist genomes are in the lower panel. The filled quadrangle and open quadrangle in a and b represent genes transcribed in the leading and the lagging strands of replication, respectively, whereas in c and d, the open and filled circles represent, respectively, the potential highly expressed genes and the other genes transcribed from the leading strand only.

(plot not shown). The codons CUG (Leu), GGG (Gly), AGU (Ser), CGC (Arg), GGC (Gly), GCC (Ala), and ACC (Thr) are the exceptions. It is also observed that the triplets occupying two extremes on the first axis are more represented by the NNA and NNU codons than by the NNC and NNG codons, suggesting that the relative contributions of the NNA and NNU codons to imparting the strand specificity in codon usage patterns are greater than those of NNC and NNG. Thus, in both species under study, the primary cause of variation in synonymous codon usage is whether a gene is transcribed from the leading strand or from the lagging strand of replication. To our knowledge, this is the first report that the strong strand-specific bias is the prime source behind the synonymous codon usage in any actinobacteria. Furthermore, in both strains of *Tropheryma*, the number of predicted coding sequences is significantly higher in the leading strand (73.8% for the TW08/27 and 74.1% for the Twist genome) than in the lagging strand. Although in *T. whipplei* complete information on essential genes is not available, it is important to note that the distribution of potential highly expressed genes (i.e., ribosomal proteins, transcription and translation processing factors, etc.) is also significantly skewed, most of them (>70%) being transcribed from the leading strand. All these observations indicate that replicational-transcriptional selection coupled with asymmetrical mutational bias might have been the primary source of intragenomic variations in codon usage pattern in *T. whipplei*.

In order to examine whether the strand-specific codon usage is a general characteristic of actinobacteria or whether it is a specific feature of *T. whipplei* only, synonymous codon usage patterns in all actinobacterial genomes sequenced so far were analyzed in the present study. Table 1 represents the comparison of codon usage patterns (as determined by COA on RSCU values of genes) of 17 completely sequenced actinobacteria. In all organisms except the two strains of *T. whipplei* and three species of *Corynebacterium*, the position of each gene along the first axis exhibits a much higher correlation with the respective GC_3

content than with GT_3, so that the genes characterized by the highest and the lowest GC_3 contents lie on the two extremes of that axis, and for most of these actinobacterial genomes, the second major axis exhibits a strong correlation with the GT_3 content of the genes. In *C. diphtheriae*, *C. glutamicum*, and *C. efficiens*, strong correlations of axis 1 are observed with GT_3 as well as with GC_3 content of the genes. The correlation of axis 1 with GT_3 is stronger in *C. diphtheriae* (average GC content, 53.5%) and *C. glutamicum* (average GC content, 53.8%) than that in *C. efficiens* (GC content, 63.1%). This result is consistent with the GC skew profiles of the respective genomes (Nishio et al. 2003). On the other hand, in *T. whipplei*, the axis 1 exhibits a strong correlation only with GT_3 ($r = -0.82$ and -0.81 at $p < 0.0001$ for the TW08/27 and Twist genomes, respectively). The genes representing the highest GT_3 content and displaying the most negative values along the first principal axis are present in the leading strand of replication (Figs. 1a and b). Thus the influence of strand-specific base compositional asymmetry on synonymous codon usage seems to be more pronounced in *T. whipplei* compared to other actinobacteria under study.

Table 2 shows the cumulative codon usage patterns in the genes located in the leading and lagging strands of the TW08/27 and Twist genomes. The chi-square test on occurrences of different codons in two replicating strands showed that there are 22 G-ending/U-ending codons, usages of which significantly increase ($p < 0.01$) in the leading strand genes, whereas usages of 24 and 25 C-ending/A-ending codons are significantly higher in the lagging strand of the TW08/27 and Twist genomes, respectively. The rest of the synonymous codons are used almost equally by both replicating strands. Unlike *B. burgdorferi* (McInerney 1998), not all these equally used codons are the rare codons. For example, CCG (Pro) and ACC (Thr), though used with relatively high frequencies, are almost equally used by the genes of two strands (Table 2). Although the TW08/27 and Twist genomes share >99% of identity at the nucleotide sequence level, it is noticeable that the genome organization differs by an approximately symmetrical inversion of a large chromosomal segment (Raoult et al. 2003). Due to such symmetrical inversion encompassing the origin or terminus, the locations of the genes do not change with respect to the leading and lagging strands (Eisen et al. 2000; Mackiewicz et al. 2001). The genomic coordinates of inversion region in the Twist genome are 182333–713982 bp. The Twist coordinates 182333 and 713982 bp correspond to the coordinates 712566 and 181874 bp, respectively, in the TW08/27 genome (Raoult et al. 2003). The inverted region in *T. whipplei* encompasses the replication terminus, which is located almost at the center of this region. Hence in *T. whipplei*, after inversion the location of the gene is conserved with respect to the two replicating strands, and as a consequence the trends in synonymous codon usage remain unaltered.

It is worth mentioning at this point that depending on the relative kinetics of genomic rearrangements and the mutational bias, the intensities of strand asymmetries vary from organism to organism and high genome plasticity, in general, diminishes the base skews induced by asymmetric mutational biases in two replicating strands (Lopez and Philippe 2001). The genome of *T. whipplei* is characterized by a high degree of plasticity (Bentley et al. 2003; Raoult et al. 2003). The existence of a pronounced strand asymmetry in *T. whipplei* genome, therefore, contradicts the hypothesis made by Lopez and Philippe (2001). Probably in *T. whipplei*, the strand-specific asymmetric nature of the mutational bias has had enough time to conciliate, in spite of their high genome plasticity.

Evidence for Translational Selection in T. whipplei

In order to examine the possible effect of translational selection, if any, on codon selection by the highly expressed genes of *T. whipplei*, we have performed a COA on RSCU values of the genes transcribed from the leading strand of replication only (as most of the highly expressed genes (>70%) are located in leading strand). The potential highly expressed genes (i.e., genes encoding ribosomal proteins, transcription translation processing factors, heat shock proteins, etc.) that are transcribed from the leading strand mostly clustered at one extreme of axis 1, which represents about 8% of the total variance for both strains of *T. whipplei* (Figs. 1c and d). Therefore, the position of genes along the first axis of COA may be related to expressivity, though the influence of gene expressivity on synonymous codon selection is not expected to be strong enough, as axis 1 described a rather small amount of total variation. Consistent with this, when the pooled data from either extreme of the first axis are compared, it is found that there are 23 codons for TW08/27 and 25 codons for Twist genomes, usages of which are significantly higher ($p < 0.05$) among the potential highly expressed genes (Table 2). Among these preferred codons in potential highly expressed genes, 21 codons for the TW08/27 and 23 codons for the Twist genomes are G and U ending. However, in the quartets of Arg and Gly, C-ending triplets (CGC for Arg and GGC for Gly) are also significantly overrepresented in highly expressed genes. All these observations suggest that apart from the replication-transcriptional selection, a residual codon bias exists in highly expressed genes of *T. whipplei*, possibly due to translational selection.

Table 2. Codon usage of the leading and lagging strand genes and preferred codons in highly expressed sequences for the *T. whipplei* TW08/27 and *T. whipplei* Twist genomes

		T. whipplei TW08/27				*T. whipplei* Twist			
		Leading strand genes		Lagging strand genes		Leading strand genes		Lagging strand genes	
Amino acid	Codon	RSCU	*N*	RSCU	*N*	RSCU	*N*	RSCU	*N*
Phe	UUU	**1.48***	6093	1.36	1748	**1.49***	6149	1.35	1808
	UUC	0.52	2117	0.64†	826	0.51	2130	0.65†	865
Leu	UUA	0.64	2096	0.60	671	0.64	2130	0.61	703
	UUG	**1.08***	3528	0.88	987	**1.11***	3675	0.91	1054
	CUU	**1.93***	6329	1.66	1873	**1.93***	6404	1.65	1908
	CUC	0.77	2527	0.82	928	0.75	2486	0.81	936
	CUA	0.58	1886	0.80†	896	0.56	1855	0.79†	910
	CUG	1.00	3279	1.24†	1400	1.00	3321	1.24†	1441
Ile	AUU	**1.45***	5959	1.17	1946	**1.45***	6026	1.16	1962
	AUC	0.56	2293	0.62†	1028	0.57	2355	0.63†	1070
	AUA	0.99	4082	1.20†	1996	0.98	4067	1.21†	2040
Met	AUG	1.00	3307	1.00	1264	1.00	3407	1.00	1267
Val	GUU	**1.94***	7490	1.55	1978	**1.93***	7524	1.55	2031
	GUC	0.60	2328	0.63	808	0.60	2355	0.63	823
	GUA	0.69	2662	0.88†	1116	0.69	2708	0.86†	1122
	GUG	**0.77**	2959	0.94†	1194	**0.78**	3033	0.96†	1258
Ser	UCU	**1.60***	4105	1.18	986	**1.61***	4146	1.16	1000
	UCC	0.78	1997	0.77	643	0.76	1970	0.76	656
	UCA	1.15	2933	1.30†	1091	1.14	2928	1.28†	1109
	UCG	**0.76***	1953	0.68	569	**0.77***	1996	0.69	594
Pro	CCU	0.96*	2118	0.80	596	**0.96***	2130	0.80	611
	CCC	0.87	1908	0.83	615	0.88	1937	0.83	637
	CCA	1.15	2524	1.41†	1048	1.12	2474	1.39†	1066
	CCG	**1.03**	2263	0.96	717	**1.04**	2307	0.97	743
Thr	ACU	**0.97***	2506	0.80	792	**0.98***	2521	0.79	801
	ACC	1.06	2728	1.04	1031	1.04	2671	1.03	1045
	ACA	1.40	3621	1.57†	1564	1.38	3547	1.57†	1583
	ACG	**0.57**	1475	0.60	593	**0.59**	1507	0.61	617
Ala	GCU	1.08*	4254	0.87	1215	1.07*	4271	0.88	1262
	GCC	0.98*	3865	0.84	1171	0.98*	3914	0.81	1166
	GCA	1.18	4666	1.48†	2071	1.18	4730	1.47†	2107
	GCG	**0.77**	3025	0.82	1150	**0.77**	3061	0.84†	1203
Tyr	UAU	1.28*	3354	1.15	1120	**1.28***	3327	1.16	1158
	UAC	0.72	1883	0.85†	823	0.72	1889	0.84†	842
Ter	UAA	0.97	174	0.94	60	0.98	177	0.90	57
Ter	UAG	1.03	184	0.99	63	1.05	190	0.99	63
His	CAU	**1.28***	2494	1.16	786	**1.26***	2520	1.16	807
	CAC	0.72	1417	0.84†	575	0.74	1472	0.84†	587
Gln	CAA	0.76	2220	0.96†	1019	0.75	2220	0.96†	1040
	CAG	**1.24***	3616	1.04	1099	**1.25***	3666	1.04	1128
Asn	AAU	**1.35***	4334	1.27	1775	**1.36***	4388	1.26	1819
	AAC	0.65	2076	0.73†	1010	0.64	2055	0.74†	1060
Lys	AAA	1.06	4456	1.23†	2269	1.06	4513	1.22†	2328
	AAG	**0.94***	3976	0.77	1428	**0.94***	4024	0.78	1473
Asp	GAU	1.42*	7211	1.29	2229	1.42*	7283	1.29	2254
	GAC	0.58	2934	0.71†	1218	0.58	2967	0.71†	1245
Glu	GAA	0.96	4424	1.12†	2067	0.97	4499	1.12†	2116
	GAG	**1.04***	4753	0.88	1634	**1.03***	4814	0.88	1653
Cys	UGU	1.10	1309	1.07	465	1.09	1368	1.07	499
	UGC	0.90	1068	0.93	403	0.91	1131	0.93	434
Ter	UGA	1.00	180	1.07	68	0.98	177	1.11	70
Trp	UGG	1.00	1739	1.00	826	1.00	1770	1.00	870
Arg	CGU	**1.20***	2402	0.82	534	**1.20***	2450	0.80	544
	CGC	**1.16***	2321	0.94	615	**1.16***	2370	0.93	632
	CGA	0.52	1047	0.69†	448	0.53	1083	0.69†	468
	CGG	**0.64**	1285	0.64	419	**0.66**	1344	0.65	439
Ser	AGU	0.97	2479	1.07†	894	0.97	2508	1.08†	933

(Continued)

Table 2. Continued

		T. whipplei TW08/27				*T. whipplei* Twist			
		Leading strand genes		Lagging strand genes		Leading strand genes		Lagging strand genes	
Amino acid	Codon	RSCU	*N*	RSCU	*N*	RSCU	*N*	RSCU	*N*
	AGC	0.74	1883	1.01†	847	0.74	1912	1.03†	887
Arg	AGA	1.23	2459	1.59†	1035	1.20	2448	1.60†	1084
	AGG	**1.24**	2472	1.32	859	**1.24**	2534	1.33	899
Gly	GGU	1.44*	4977	1.17	1445	1.43*	5041	1.15	1463
	GGC	**1.06***	3662	0.92	1142	**1.06***	3732	0.92	1165
	GGA	0.76	2617	0.98†	1217	0.74	2616	1.01†	1281
	GGG	**0.75**	2608	0.92†	1140	**0.76**	2666	0.93†	1176

Note. N, total number of codons; RSCU, relative synonymous codon usage. Corresponding codons are significantly ($p < 0.01$) overrepresented among *leading strand genes and †lagging strand genes. Boldfaced values indicate that corresponding codons are more frequent ($p < 0.05$) in highly expressed sequences.

In order to ascertain the influence of the residual selection putatively operating at the level of translation on synonymous codon bias in highly expressed genes, we used two different approaches. First, we calculated CAI for all leading strand genes and examined whether there exists any correlation of CAI values of the genes with their positions along the first principal axis of COA on RSCU values of the leading strand genes. The first axis exhibits a strong significant correlation ($r = 0.79$ and -0.77 at $p < 0.0001$ for the TW08/27 and Twist genomes, respectively) with the CAI values of genes. The genes displaying the higher values of CAI are the sequences with putative high expression levels, including genes coding for ribosomal proteins, subunits of RNA polymerase, heat shock proteins, etc. This suggests that there is a preferential use of a particular subset of synonymous codons by the highly expressed genes in *T. whipplei*.

The second approach is to compare the estimated d_S values between 50 putative highly and 50 putative lowly expressed orthologous genes between the TW08/27 and the Twist genomes taken from the pooled data of the two extremes of axis 1 generated by COA on RSCU for leading strand genes of both organisms. Although these two genomes are very closely related, the genes representing significantly low values of d_S are the potential highly expressed genes (i.e., ribosomal proteins, transcription and translation processing factors, heat shock proteins, etc.). The mean d_S value for high-expression genes is 0.0049, which is significantly lower (*t*-test, $p < 10^{-4}$) than that of low-expression genes (mean $d_S = 0.0058$), indicating that highly expressed genes have lower divergence at synonymous sites than lowly expressed genes of these two *T. whipplei* genomes, since their separation from the last common ancestor. This, in turn, is consistent with the hypothesis that the nucleotide selection force acting at the synonymous sites is stronger for highly expressed genes than that for lowly expressed ones, as reported previously for enterobacteria (Sharp and Li 1987b) and *Bartonella* (Das et al. 2005a). Therefore, it may be concluded that apart from the strand-specific mutational bias, a selection, putatively operating at the level of translation, also might have influenced the synonymous codon choice in genes of these two strains of *T. whipplei*.

Recent studies on complete microbial genomes indicated a close association among synonymous codon usage bias, tRNA abundance, number of rRNA operons, generation time, and genome size (Rocha 2004; dos Reis et al. 2004; Sharp et al. 2005). Evidence for apparently little or no translational selection has been reported for most of the species with reduced genomes examined so far, such as *B. burgdorferi* (McInerney 1998), *B. aphidicola* (Rispe et al. 2004), *H. pylori* (Lafay et al. 2000), *T. pallidum* (Lafay et al. 1999), and *Wigglesworthia* (Herbeck et al. 2003). Most endoparasitic and endosymbiotic bacteria are characterized by the presence of only one or two rRNA operons and a small number of tRNA genes, a long generation time, and a reduced genome and overall AT richness. Some of these bacteria such as *B. burgdorferi* (McInerney 1998), *B. aphidicola* Bpi strain (Rispe et al. 2004), *T. pallidum* (Lafay et al. 1999), *C. trachomatis* (Romero et al. 2000), and two species of *Bartonella* (Das et al. 2005a) exhibit strong base compositional asymmetries like *T. whipplei*. The genomes of the TW08/27 and Twist strains of *T. whipplei* have undergone a massive reduction in size as well as a decrease in overall GC content compared to their free-living actinobacterial relatives. Consistent with the characteristics of intracellular bacteria (Sharp et al. 2005; Rocha 2004), the in vitro doubling time of *T. whipplei* is remarkably longest among bacteria (Bentley et al. 2003), suggesting its slow growth rate. Both genomes contain only one rRNA operon and a limited number of tRNA genes

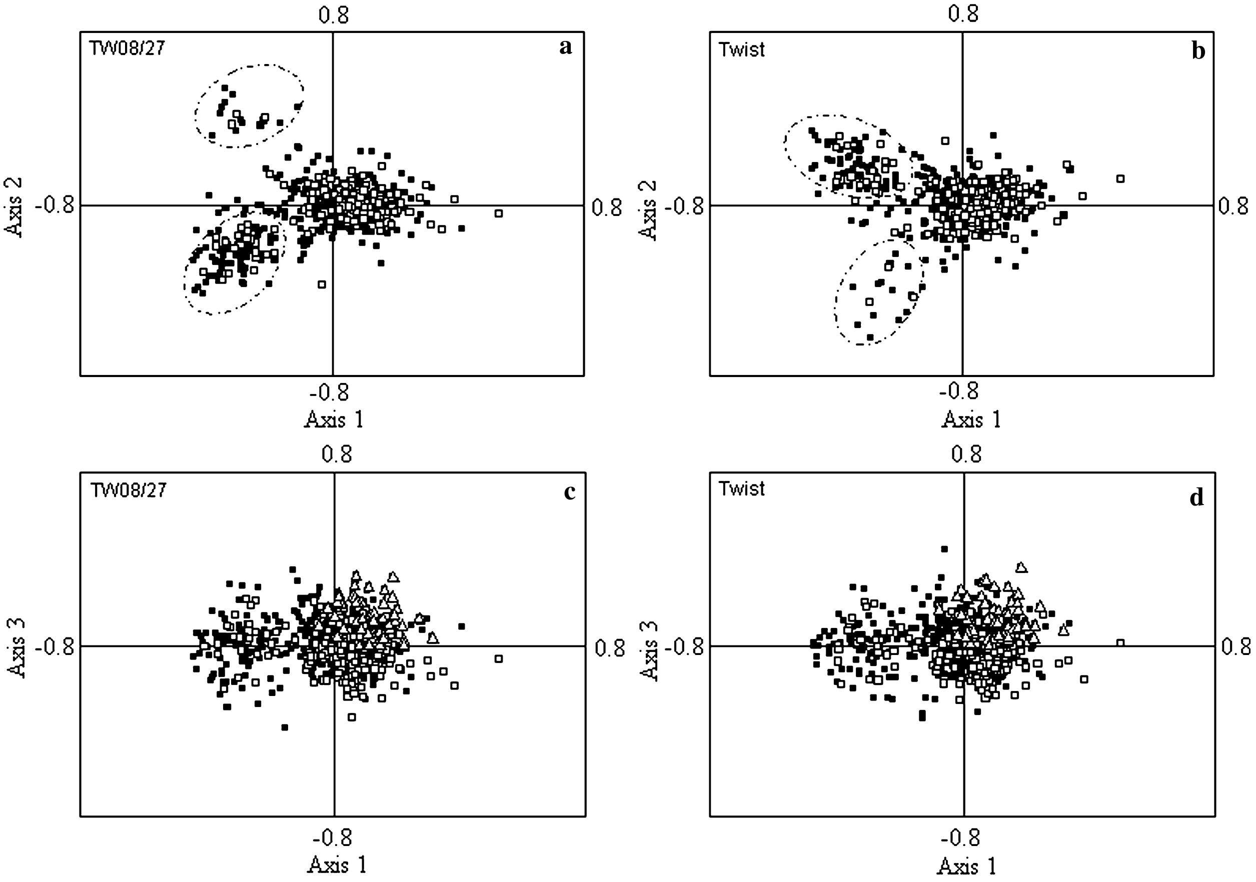

Fig. 2. Position of each gene along the first and second axes (**a**, **b**) and first and third axes (**c**, **d**) generated by COA on RAAU of encoded gene products for *T. whipplei* TW08/27 and *T. whipplei* Twist. The filled quadrangle and open quadrangle represent genes transcribed in the leading and lagging strands of replication, respectively. The large and small clusters of genes encoding membrane-associated proteins are marked by dashed-line ovals (in a and b). The open triangles (in c and d) represent potentially highly expressed genes.

(51 for the TW08/27 and 50 for the Twist genome). It is worth mentioning at this point that in spite of the relatively high GC content and large genome size of *Mycobacterium* species, the weakly selected codon usage bias found was probably due to the lower number of rRNA operons and tRNA genes (Sharp et al. 2005). Evidence for a weak translational selection in *T. whipplei* genomes is, therefore, consistent with the earlier observations on the organisms adapted in intracellular lifestyle.

Major Sources of Variation in Proteome Composition

To investigate whether the DNA strand specific mutational biases have any impact on amino acid usage in *T. whipplei* gene products, we performed COA on relative amino acid usage (RAAU) of the encoded proteins. Figure 2 represents the position of each gene on the plane defined by the first two major axes of COA on RAAU of *T. whipplei* gene products. There is no clear segregation of the proteins encoded by the leading and lagging strands of the genomes under study. As G and T bases are prevalent in the leading strand, while C and A are overrepresented in the lagging strand, the amino acids encoded by either G + U-rich (such as Phe, Cys, Trp, Gly, and Val) or A + C-rich (such as Lys, Asn, Pro, Gln, His, and Thr) codons are expected to be prevalent in respective strands. When the cumulative amino acid usages of encoded gene products in two strands are compared separately, among these 11 residues, only 2 (Phe and Val) are significantly overrepresented ($p < 0.001$) in the leading strand and 3 residues (Lys, Asn, and Thr) are more abundant ($p < 0.001$) in the lagging strand. All other residues are used evenly by the genes of the two strands, indicating that the influence of the strand-specific mutational bias on the amino acid composition of *T. whipplei* gene products is not as pronounced as on their synonymous codon usage.

In both strains of *T. whipplei*, the first three axes generated by COA on amino acid usage explain about 42% of the total variability. The first and second major axes are strongly correlated both with hydrophobicity and with the aromaticity of encoded proteins (Table 3), implying hydropathy and aromaticity to be the major factors for amino acid variation

Table 3. Major trends in amino acid usage, as revealed by COA on amino acid frequencies of the encoded proteins

	T. whipplei TW08/27			*T. whipplei* Twist		
Axis	Variability explained (%)	Source of variation	Correlation coefficient (*r* value)*	Variability explained (%)	Source of variation	Correlation coefficient (*r* value)*
1	18.9	Gravy Score	−0.83	18.4	Gravy Score	−0.87
		Aromaticity	−0.68		Aromaticity	−0.68
2	14.8	Gravy Score	−0.59	14.1	Gravy Score	0.46
		Aromaticity	−0.45		Aromaticity	0.38
3	8.8	Size/complexity	−0.79	8.8	Size/complexity	−0.74
		GC_{1+2}	0.65		GC_{1+2}	0.66

*All correlations are significant at $p < 0.0001$.

Table 4. Amino acid usage and mean value of potential secondary structures of membrane-associated proteins in the large and small clusters for the *T. whipplei* TW08/27 and *T. whipplei* Twist genomes

Amino acid or predicted secondary structure	*T. whipplei* TW08/27		*T. whipplei* Twist	
	Large cluster (%)	Small cluster (%)	Large cluster (%)	Small cluster (%)
Phe	8.17*	3.17	7.87*	3.32
Leu	14.36*	8.24	14.46*	8.07
Ser	7.54	10.62†	7.66	10.27†
Tyr	2.97	4.70†	2.96	4.70†
Cys	1.07*	0.59	1.16*	0.59
Trp	1.81*	0.68	1.77*	0.79
Pro	4.27	8.07†	4.36	7.92†
His	1.33	1.73	1.39	1.47
Gln	1.83	3.20†	1.96	3.20†
Arg	4.56*	3.14	4.82*	3.18
Ile	8.76*	4.76	8.46*	4.77
Met	2.25*	0.50	2.18*	0.56
Thr	4.67	13.78†	4.67	13.93†
Asn	2.54	3.57†	2.57	3.44†
Lys	2.80	4.08†	2.83	3.97†
Val	9.29	8.74	9.25	8.99
Ala	9.43*	6.41	9.27*	6.43
Asp	2.33	3.67†	2.43	3.63†
Glu	1.97	2.10	2.08	2.16
Gly	7.71	8.09	7.49	8.46
α-Helix	32.17*	9.44	31.92*	9.93
ß-Sheet	24.29	29.95†	24.96	29.62†
Random coil	43.54	60.61†	43.12	60.45†

Note. Values significantly ($p < 0.001$) more frequent in *large or †small cluster gene products.

in *T. whipplei* proteins. The first axis, therefore, separates the genes encoding proteins having high values of mean hydropathy and mean aromaticity on the left and the others on the right side of that axis (Fig. 2). Furthermore, in the TW08/27 genome low-hydrophobic and low-aromatic proteins display positive values on axis 2, whereas in the Twist genome such proteins are represented by negative values on the same axis (Fig. 2).

It is worth noting that there are two distinct clusters of proteins near the left end of axis 1 (Fig. 2). A careful investigation reveals that the small cluster contained the genes for membrane-associated proteins including WiSP family members and a few hypothetical proteins, whereas genes coding for integral membrane proteins, several transporters, subunits of cytochrome *c*, translocase protein *TatC*, etc., are present in the large cluster. Although the genes present in both clusters are mainly membrane-associated proteins, the amino acid usage profiles exhibited by them are quite distinct from one another. Membrane proteins are known to have very different amino acid compositions (Lobry and Gautier 1994; Pascal et al. 2005). The products of the genes present in the large cluster are significantly overrepresented ($p < 0.001$) by Phe, Leu, Cys, Trp, Arg, Ile, Met, and Ala, whereas Ser, Tyr, Pro, Gln, Thr, Asn, Lys, and Asp present in significantly higher amounts ($p < 0.001$) in the proteins encoded by small cluster genes (Table 4).

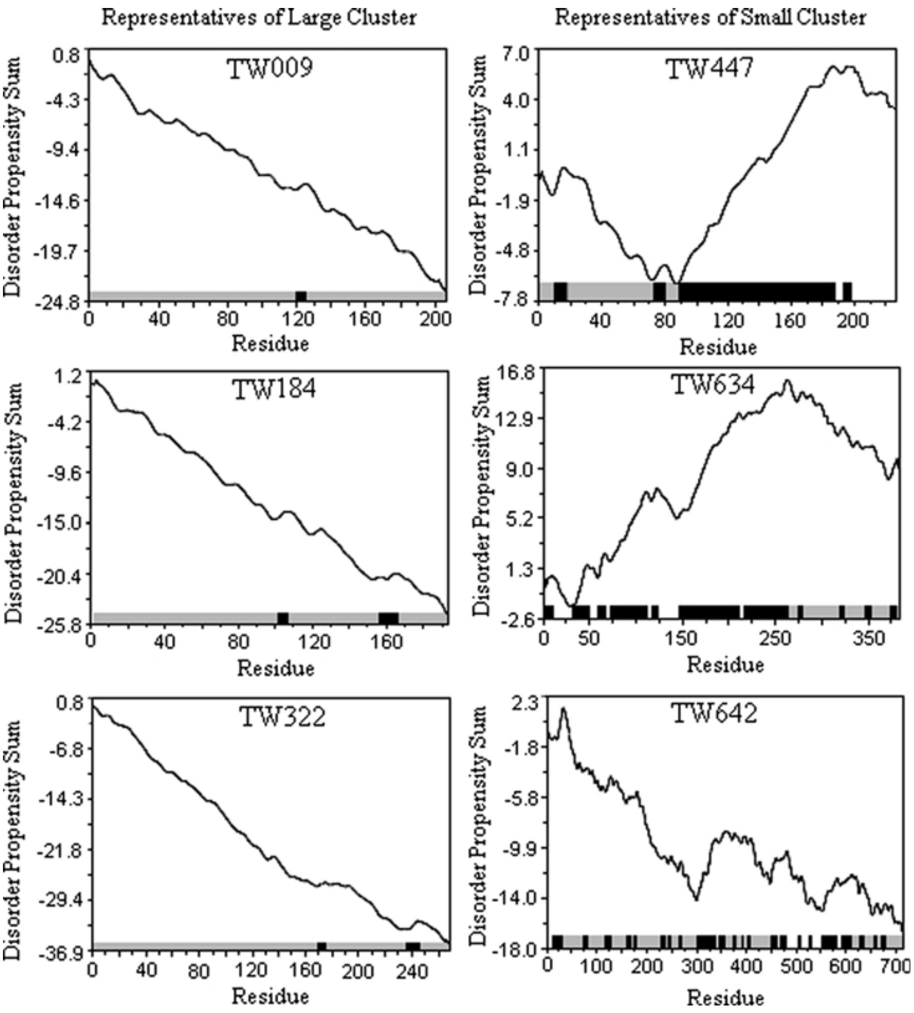

Fig. 3. GlobPlot of the membrane-associated proteins encoded by genes taken from large and small clusters generated by COA on RAAU for the *T. whipplei* TW08/27 genome. Plots in the left and right panels represent large and small clusters, respectively. Black color indicates disordered regions (lack of regular secondary structures) and gray color indicates ordered (globular) regions that typically contain regular secondary structures.

These two groups of proteins also differ in their predicted secondary structures. For the members of the small cluster, most of which are WiSP family proteins, propensities for formation of ß-sheet structures are much higher than that for formation of α-helices, while the proteins encoded by the large cluster genes have higher propensities for formation of α-helix structures than ß-sheets (Table 4). Proteins of these two clusters also differ in content of potentially disordered structures. Disordered regions in proteins can be predicted by the lack of regular secondary structures and increased flexibility in the polypeptide chain, whereas ordered regions (often termed globular) typically contain regular secondary structures packed into a compact globule (Linding et al. 2003; Fuxreiter et al. 2004). As the probable coil-forming regions are significantly higher in the small cluster proteins in both strains of *T. whipplei* (Table 4), disordered structures are more commonly found in the proteins of the small cluster compared to those comprising the large cluster (Fig. 3). Recent investigations have indicated that disordered structures are usually more favored by proteins involved in regulatory functions and binding of various ligands (Fuxreiter et al. 2004; Fink 2005). Therefore, it may be presumed that the proteins in the small cluster, which might play important roles in interactions with the host and/or immune evasion (Bentley et al. 2003; Raoult et al. 2003), would be overrepresented by disordered structures. Members of the other cluster, containing fewer disordered regions and exhibiting higher propensities for α-helical regions, are primarily involved in transport and other membrane-associated processes. Most of the variations in cell surface proteins might be potentially important for *T. whipplei* to evade the host's immune response and probably evolved during the course of parasitic adaptation. Compositions of the membrane-associated proteins of these two clusters are, therefore, distinguished from each other in a fine coordination among function, structure, and stability.

Another factor that influences the variations in amino acid usage is gene expressivity, as indicated by the presence of most of the potential highly expressed genes near the positive extreme of the axis 3 (Figs. 2c and d). Recently it was reported that a decrease in the average size/complexity quotient of encoded proteins with a potentially high expression level points toward

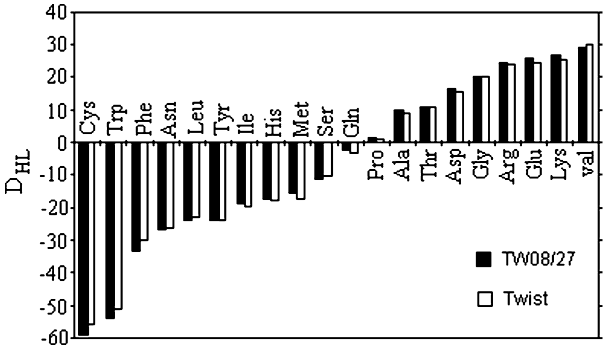

Fig. 4. Difference in relative amino acid usage of highly and lowly expressed genes of *T. whipplei* TW08/27 (filled bars) and *T. whipplei* Twist (open bars). The differences were derived as D_{HL} = [(Freq.H/Freq.L) − 1] × 100}.

a decrease in biosynthetic cost of highly expressed gene products (Das et al. 2005c). The strong negative correlation between the third major axis and the average size/complexity quotient of encoded proteins (Table 3) suggests that the highly expressed genes in *T. whipplei* have a tendency to avoid the heavier residues including the aromatic ones. This supports the cost minimization hypothesis (Seligmann 2003), which claims that highly expressed genes tend to use small and energetically less expensive amino acids in their encoded proteins. Considering the fact that both strains of *T. whipplei* are obligatory intracellular parasites, which lack the enzymes required for amino acid biosynthesis and obtain all amino acid residues by salvage from the host (Bentley et al. 2003; Raoult et al. 2003), it is surprising to see that they follow the cost minimization hypothesis. However, recently Peixoto et al. (2004) mentioned that parasitic organisms may follow the cost minimization hypothesis as a strategy for genome-level adaptation to the host environment, as utilization of the less expensive and smaller residues by the highly expressed genes can minimize the energy exhaustion of the host and help them thereby to maintain the sustained infection, having the least chance of elimination by the host.

Relation Among Gene Expression, Protein Conservation, and Amino Acid Usage

To understand the possible effect of gene expression on amino acid usage, we compared the average RAAU profiles of putative highly and lowly expressed gene products of *T. whipplei* (Fig. 4). Depending on the GC content of the nonsynonymous codon positions of the respective codons, the amino acid residues may be classified into three groups: (i) GC-rich (Gly, Ala, Arg, and Pro), (ii) AU-rich (Phe, Tyr, Met, Ile, Asn, and Lys), and (iii) unbiased amino acids (the rest) (Foster et al. 1997). On average, the codons, which are AU-rich at nonsynonymous sites, tend to encode relatively expensive amino acids (Akashi and Gojobori 2002). It is interesting to note that in both strains of *T. whipplei*, except for Pro, all residues encoded by GC-rich codons are present in significantly higher amounts ($p < 0.001$) in highly expressed gene products, while usages of the residues encoded by all AU-rich codons, except Lys, are significantly higher ($p < 0.001$) in lowly expressed gene products (Fig. 4). Increased usage of GC-rich or decreased usage of AU-rich codons in highly expressed genes of *T. whipplei* is supported by the significant positive correlation of GC_{1+2} with the coordinates of axis 3 of COA on RAAU (Table 3). This observation is in accordance with the studies on other intracellular bacteria, such as *Buchnera*, *Blochmannia floridanus*, and *Wigglesworthia*, exhibiting higher usage GC-rich amino acids in high-expression genes (Schaber et al. 2005).

Both the genomes under study have a relatively low G + C-content (~46%) compared to other actinomycete species, and hence, a question arises: Why do the highly expressed genes of these species exhibit higher usage of the residues encoded by G + C-rich codons? This might have happened for either of the following two reasons: (i) the highly expressed genes of *T. whipplei* are more conserved at the amino acid level than their lowly expressed counterparts, and hence, they have retained a GC-richer composition which is closer to their putative ancestral state, or (ii) all genes are undergoing substitutions at a comparable rate irrespective of their level of expression, but due to some functional advantages, the highly expressed genes exhibit a positive selection in favor of the residues encoded by GC-rich codons. In order to examine which of these two possibilities is more likely to be true, we compared the estimated pairwise nonsynonymous divergences (d_N) between all orthologues of putative highly and lowly expressed genes preidentified from the two extremes of axis 1 generated by COA on RSCU of the leading strand genes of *T. whipplei* TW08/27 and *T. whipplei* Twist. The mean d_N for highly expressed genes (0.021) is significantly lower (*t*-test, $p < 10^{-4}$) than that for lowly expressed genes ($d_N = 0.057$), suggesting that the amino acids of highly expressed genes are more conserved in *T. whipplei*. A similar inference was drawn previously in several organisms like *E. coli*, *S. typhimurium*, and *P. falciparum* (Sharp 1991; Chanda et al. 2005). It may be inferred therefore that in *T. whipplei*, the overall protein conservation in highly expressed genes is much higher and it also influences the discrimination in amino acid usage between highly and lowly expressed genes. The enhanced usage of the residues encoded by GC-rich codons in the highly expressed gene-products might have been possible due to more conservation at the amino acid level in highly expressed genes than the low-expression genes and hence, they retain GC-rich compositions which are closer to their putative ancestral state.

In summary, the present study reveals that the patterns of synonymous codon and amino acid usage in *T. whipplei* are the result of several factors. As observed earlier in many other species characterized by genome reduction and overall A + T-richness (McInerney 1998; Romero et al. 2000; Das et al. 2005a), in *T. whipplei* also, replicational-transcriptional selection coupled with asymmetric mutational bias is the primary cause of intragenomic variations in codon usage pattern. There is a residual intrastrand bias in synonymous codon usage by the highly expressed genes, possibly due to the presence of translational selection. However, the influence of strand-specific mutational pressure is not as pronounced at the level of amino acid usage. Hydrophobicity and aromaticity seem to be the major sources of variation, both having a nearly equal influence on amino acid usage. In spite of the intracellular lifestyle, the amino acid preferences in highly expressed gene products of *T. whipplei* follow the cost minimization hypothesis. Another interesting finding is that the products of the highly expressed genes prefer to use the residues encoded by GC-rich codons, although the *T. whipplei* genome, on average, has only a 46% G + C content. The analysis presented here indicates that this might be due to greater conservation of a relatively GC-rich ancestral state in the highly expressed genes. Even the energetic cost of amino acid residues (Akashi and Gojobori 2002) might play a significant role in retaining the GC-rich residues in highly expressed genes. The study also sheds light on the diverse compositional and structural characteristics of two groups of membrane-associated proteins that might play distinct roles in host interactions.

Acknowledgments. This work was supported by the Council of Scientific and Industrial Research (Project CMM 0017) and Department of Biotechnology, Government of India (Grant BT/BI/04/055-2001).

References

Akashi H, Gojobori T (2002) Metabolic efficiency and amino acid composition in the proteomes of *Escherichia coli* and *Bacillus subtilis*. Proc Natl Acad Sci USA 99:3695–3700

Andersson SG, Sharp PM (1996) Codon usage and base composition in *Rickettsia prowazekii*. J Mol Evol 42:525–536

Bentley SD, Maiwald M, Murphy LD et al. (2003) Sequencing and analysis of the genome of the Whipple's disease bacterium *Tropheryma whipplei*. Lancet 361:637–644

Berg OG, Silva PJ (1997) Codon bias in *Escherichia coli*: the influence of codon context on mutation and selection. Nucleic Acids Res 25:1397–1404

Casjens S (1998) The diverse and dynamic structure of bacterial genomes. Annu Rev Genet 32:339–377

Chanda I, Pan A, Dutta C (2005) Proteome composition in *Plasmodium falciparum*: higher usage of GC-rich nonsynonymous codons in highly expressed genes. J Mol Evol 61:513–523

Das S, Paul S, Chatterjee S, Dutta C (2005a) Codon and amino acid usage in two major human pathogens of genus *Bartonella*—Optimization between replication-transcriptional selection, translational control and cost minimization. DNA Res 12:91–102

Das S, Pan A, Paul S, Dutta C (2005b) Comparative analyses of codon and amino acid usage in symbiotic island and core genome in nitrogen-fixing symbiotic bacterium *Bradyrhizobium japonicum*. J Biomol Struct Dyn 23:221–232

Das S, Ghosh S, Pan A, Dutta C (2005c) Compositional variation in bacterial genes and proteins with potential expression level. FEBS Lett 579:5205–5210

de Miranda AB, Alvarez-Valin F, Jabbari K, Degrave WM, Bernardi G (2000) Gene expression, amino acid conservation, and hydrophobicity are the main factors shaping codon preferences in *Mycobacterium tuberculosis* and *Mycobacterium leprae*. J Mol Evol 50:45–55

dos Reis M, Savva R, Wernisch L (2004) Solving the riddle of codon usage preferences: a test for translational selection. Nucleic Acids Res 32:5036–5044

Dufton MJ (1997) Genetic code synonym quotas and amino acid complexity: Cutting the cost of proteins? J Theor Biol 187:165–173

Dutly F, Altwegg M (2001) Whipple's disease and "Tropheryma whippelii." Clin Microbiol Rev 14:561–583

Eisen JA, Heidelberg JF, White O, Salzberg SL (2000) Evidence for symmetric chromosomal inversions around the replication origin in bacteria. Genome Biol 1:research 0011.1–0011.9

Fink AL (2005) Natively unfolded proteins. Curr Opin Struct Biol 15:35–41

Foster PG, Jermiin LS, Hickey DA (1997) Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria. J Mol Evol 44:282–288

Frank AC, Lobry JR (2000) Oriloc: prediction of replication boundaries in unannotated bacterial chromosomes. Bioinformatics 16:560–561

Fuxreiter M, Simon I, Friedrich P, Tompa P (2004) Preformed structural elements feature in partner recognition by intrinsically unstructured proteins. J Mol Biol 338:1015–1026

Garnier J, Gibrat JF, Robson B (1996) GOR method for predicting protein secondary structure from amino acid sequence. Methods Enzymol 266:540–553

Herbeck JT, Wall DP, Wernegreen JJ (2003) Gene expression level influences amino acid usage, but not codon usage, in the tsetse fly endosymbiont *Wigglesworthia*. Microbiology 149: 2585–2596

Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. J Mol Biol 157:105–132

Lafay B, Lloyd AT, McLean MJ, Devine KM, Sharp PM, Wolfe KH (1999) Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases. Nucleic Acids Res 27:1642–1649

Lafay B, Atherton JC, Sharp PM (2000) Absence of translationally selected synonymous codon usage bias in *Helicobacter pylori*. Microbiology 146:851–860

Linding R, Russell RB, Neduva V, Gibson TJ (2003) GlobPlot: exploring protein sequences for globularity and disorder. Nucleic Acids Res 31:3701–3708

Lobry JR, Gautier C (1994) Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes. Nucleic Acids Res 22:3174–3180

Lopez P, Philippe H (2001) Composition strand asymmetries in prokaryotic genomes: mutational bias and biased gene orientation. CR Acad Sci III 324:201–208

Mackiewicz P, Mackiewicz D, Kowalczuk M, Cebrat S (2001) Flip-flop around the origin and terminus of replication in prokaryotic genomes. Genome Biol 2:interactions 1004.1–1004.4

McInerney JO (1998) Replicational and transcriptional selection on codon usage in *Borrelia burgdorferi*. Proc Natl Acad Sci USA 95:10698–10703

Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3:418–426

Nishio Y, Nakamura Y, Kawarabayasi Y, Usuda Y, Kimura E, Sugimoto S, Matsui K, Yamagishi A, Kikuchi H, Ikeo K, Gojobori T (2003) Comparative complete genome sequence analysis of the amino acid replacements responsible for the thermostability of *Corynebacterium efficiens*. Genome Res 13:1572–1579

Ohama T, Muto A, Osawa S (1990) Role of GC-biased mutation pressure on synonymous codon choice in *Micrococcus luteus*, a bacterium with a high genomic GC- content. Nucleic Acids Res 18:1565–1569

Ohkubo S, Muto A, Kawauchi Y, Yamao F, Osawa S (1987) The ribosomal protein gene cluster of *Mycoplasma capricolum*. Mol Gen Genet 210:314–322

Pan A, Dutta C, Das J (1998) Codon usage in highly expressed genes of *Haemophillus influenzae* and *Mycobacterium tuberculosis*: translational selection versus mutational bias. Gene 215:405–413

Pascal G, Medigue C, Danchin A (2005) Universal biases in protein composition of model prokaryotes. Proteins 60:27–35

Peixoto L, Fernandez V, Musto H (2004) The effect of expression levels on codon usage in *Plasmodium falciparum*. Parasitology 128:245–251

Raoult D, Ogata H, Audic S, Robert C, Suhre K, Drancourt M, Claverie JM (2003) *Tropheryma whipplei* Twist: a human pathogenic Actinobacteria with a reduced genome. Genome Res 13:1800–1809

Ratnaike RN (2000) Whipple's disease. Postgrad Med J 76:760–766

Ring S, Schneider T, Marth T (2003) Mucosal immune response to *Tropheryma whipplei*. Int J Med Microbiol 293:69–76

Rispe C, Delmotte F, van Ham RC, Moya A (2004) Mutational and selective pressures on codon and amino acid usage in *Buchnera*, endosymbiotic bacteria of aphids. Genome Res 14:44–53

Rocha EP (2004) Codon usage bias from tRNA's point of view: redundancy, specialization, and efficient decoding for translation optimization. Genome Res 14:2279–2286

Rocha EP, Danchin A (2003) Essentiality, not expressiveness, drives gene-strand bias in bacteria. Nat Genet 34:377–378

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

Schaber J, Rispe C, Wernegreen J, Buness A, Delmotte F, Silva FJ, Moya A (2005) Gene expression levels influence amino acid usage and evolutionary rates in endosymbiotic bacteria. Gene 352:109–117

Seligmann H (2003) Cost-minimization of amino acid usage. J Mol Evol 56:151–161

Sharp PM (1991) Determinants of DNA sequence divergence between *Escherichia coli* and *Salmonella typhimurium*: codon usage, map position, and concerted evolution. J Mol Evol 33:23–33

Sharp PM, Li WH (1987a) The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications. Nucleic Acids Res 15:1281–1295

Sharp PM, Li WH (1987b) The rate of synonymous substitution in enterobacterial genes is inversely related to codon usage bias. Mol Biol Evol 4:222–230

Sharp PM, Bailes E, Grocock RJ, Peden JF, Sockett RE (2005) Variation in the strength of selected codon usage bias among bacteria. Nucleic Acids Res 33:1141–1153

Singer GA, Hickey DA (2003) Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content. Gene 317:39–47

Slowik A, Szczudlik A (2002) Whipple's disease—a rare cause of neurological symptoms and disorders. Neurol Neurochir Pol 36:959–970

Wright F, Bibb MJ (1992) Codon usage in the G + C–rich Streptomyces genome. Gene 113:55–65

Wu M, Sun LV, Vamathevan J et al. (2004) Phylogenomics of the reproductive parasite *Wolbachia pipientis* wMel: a streamlined genome overrun by mobile genetic elements. PLoS Biol 2:0327–0341

Zavala A, Naya H, Romero H, Musto H (2002) Trends in codon and amino acid usage in *Thermotoga maritima*. J Mol Evol 54:563–568