J Mol Evol (2002) 54:563–568
DOI: 10.1007/s00239-001-0040-y

JOURNAL OF MOLECULAR EVOLUTION

Trends in Codon and Amino Acid Usage in *Thermotoga maritima*

Alejandro Zavala,[1] Hugo Naya,[1] Héctor Romero,[1,2] Héctor Musto[1]

[1] Laboratorio de Organización y Evolución del Genoma, Departamento de Biología Celular y Molecular, Facultad de Ciencias, Iguá 4225, Montevideo 11400, Uruguay
[2] Departamento de Genética, Facultad de Medicina, Montevideo, Uruguay

Received: 22 June 2001 / Accepted: 1 October 2001

Abstract. The usage of synonymous codons and the frequencies of amino acids were investigated in the complete genome of the bacterium *Thermotoga maritima* using a multivariate statistical approach. The GC3 content of each gene was the most prominent source of variation of codon usage. Surprisingly the usage of UGU and UGC (synonymous triplets coding for Cys, the least frequent amino acid in this species) was detected as the second most prominent source of variation. However, this result is probably an artifact due to the very low frequency of Cys together with the nonbiased composition of this genome. The third trend was related to the preferential usage of a subset of codons among highly expressed genes, and these triplets are presumed to be translationally optimal. Concerning the amino acid usage, the hydropathy level of each protein (and therefore the frequency of charged residues) was the main trend, while the second factor was related to the frequency of usage of the smaller residues, suggesting that the cell economy strongly influences the architecture of the proteins. The third axis of the analysis discriminated the usage of Phe, Tyr, Trp (aromatic residues) plus Cys, Met, and His. These six residues have in common the property of being the preferential targets of reactive oxygen species, and therefore the anaerobic condition of *T. maritima* is an important factor for the amino acid frequencies. Finally, the Cys content of each protein was the fourth trend.

Key words: Codon usage — Amino acid usage — Amino acid frequency — *Thermotoga maritima* — GC3 content

Introduction

Until the availability of complete sequenced genomes, it was generally accepted that for unicellular organisms the factors that shape codon usage were limited to an equilibrium between mutational biases (toward G + C or A + T) and translational selection for elongation rate (acting mainly on highly expressed genes) and/or accuracy (reviewed by Sharp et al. 1995; Akashi and Eyre-Walker 1998). However, very recent results showed that the mutational biases are more complex than simply toward high or low levels of genomic G + C and that new factors may be at work during translation.

For example, in the spirochaetes *Borrelia burgdorferi* and *Treponema pallidum* it was shown that translational selection does not operate on highly expressed sequences and that the main factor shaping codon usage is the strong strand-specific mutational pressure. Indeed, the genomes of these prokaryotes are characterized by a strong GC skew [the quantity (G − C)/(G + C)], which switches sign at the origin of replication. As a consequence, the leading strand of replication is G- (and T-)-rich, and therefore genes placed on that strand display a strong bias toward those bases at the silent sites, while the opposite biases (toward C- and A-ending triplets) are found in genes placed on the lagging strand (McInerney 1998; Lafay et al. 1999). In other words, the major cause of variation in codon usage in these species is the location of a gene, either on the leading or on the lagging strand in replication. In *Mycoplasma genitalium,* it was reported that there is a systematic variation in GC3 around the genome, and therefore this factor strongly determines the pattern of codon usage (Kerr et al. 1997; McInerney 1997). For *Mycobacterium tuberculosis,* it

Correspondence to: Héctor Musto; *email:* hmusto@fcien.edu.uy

appears that codon usage is influenced by mutational bias and translational selection, and, interestingly, it was reported that the hydropathy level of each protein is influenced by the base composition at the synonymous sites. Indeed, in nearly all quartets there is an increase in G-ending codons and a decrease in C-ending triplets as long as hydrophobicity increases (de Miranda et al. 1999). A rather complex pattern was reported for *Chlamydia trachomatis,* where codon choices are the result of strand-specific mutational biases, natural selection acting at the level of translation, the hydropathy level of each protein, and amino acid conservation (Romero et al. 2000). Finally, in *Helicobacter pylori,* although the genome composition is not skewed and there is a low level of heterogeneity among genes, codon usage seems to be influenced neither by simple mutational biases nor by translational selection (Lafay et al. 2000). Therefore, it appears that as long as more complete prokaryotic genomes are analyzed, codon usage seems to be influenced by new factors, although at the moment all of them can be reduced to the "mutational bias–translational selection" paradigm (Romero et al. 2000).

Recently the complete genome sequence of *Thermotoga maritima* (Nelson et al. 1999), a hyperthermophilic bacterium which is probably one of the deepest lineages in the Eubacteria (Tiboni et al. 1991), has been reported. This genome is characterized by an average genomic G + C content of 46% and consists of 1860 kbp with 1846 predicted coding regions, but perhaps the most striking result that emerged from its sequencing was the finding that 24% of its genes seem to be derived from Archaea (Nelson et al. 1999; Logsdon and Faguy 1999).

In this paper we report the analysis of codon usage in *T. maritima* and show that the final pattern is the result of several factors, including mutational bias and natural selection. Furthermore, we study the main factors shaping the amino acid usage of the proteins encoded in this genome.

Materials and Methods

Sequences. The complete genome and coding sequences of *T. maritima* (Nelson et al. 1999) were obtained from ftp://ftp.tigr.org/pub/data/t_maritima/.

Methods. Codon usage, correspondence analysis (COA) (Greenacre 1984), GC3s (the frequency of codons ending in C or G, excluding Met, Trp, and stop codons), N_c [the "effective number of codons" (Wright 1990)], the relative synonymous codon usage (RSCU) (Sharp et al. 1986), and the codon adaptation index (CAI) (Sharp and Li 1987) were calculated using the program CodonW 1.3 (written by John Peden and available at ftp://molbiol.ox.ac.uk/Win95.codonW.zip). The CAI was calculated taking as a reference the codon usage of the ribosomal proteins. COA of RSCU values and amino acid frequencies was carried out to determine the major sources of variation among synonymous codons and amino acid usage, respectively.

The mean molecular weight (MMW) of the encoded amino acids for each protein was defined as

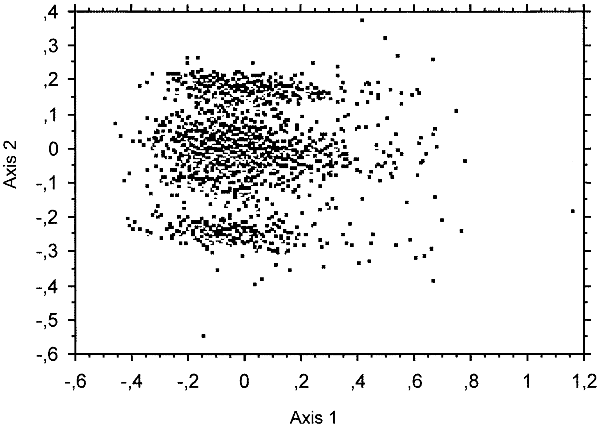

Fig. 1. Plot of the two most prominent axes generated from the COA of the RSCU values for the ORFs analyzed in this paper.

$$\mathrm{MMW} = (\Sigma_i \mathrm{MW})/N$$

where MW is the molecular weight of amino acid i and N is the number of amino acids in each protein.

To locate the genes in the leading or lagging strand of replication, the origin was assumed to be placed between position 155,060 and position 162,813, as suggested by Lopez et al. (2000).

Results and Discussion

Codon Usage

The pattern of codon usage in *T. maritima* was investigated by COA in a data set comprising 1846 ORFs. The removal of presumed duplicates and recent horizontally transferred genes from the data set (15% of the sequences) did not change the results significantly, therefore all analyses were carried out on the total described ORFs. Figure 1 shows the position of each gene on the plane defined by the first (horizontal) and second (vertical) axes in this analysis. These two main axes represented 8.3 and 4.8% of the total variance, respectively. Remarkably, the first principal component explained a rather low proportion of the total variability, which. suggests that in this bacterium the major trend in codon usage among the genes is not as strong as in other species such as *Escherichia coli, Bacillus subtilis, Haemophilus influenzae,* and *C. trachomatis,* where the major trend explained 15.5, 12.4, 11.1, and 9.9%, respectively. The position of each gene along the first axis was strongly correlated with the respective GC3s content of the sequences ($r = -0.84$, $p < 0.0001$), and the genes displaying the highest GC3s levels displayed the most negative values along that axis. When the frequencies of each base at the silent sites were considered separately, we found that the highest correlations were with the pyrimidines (Y). Indeed, the orders of the absolute r values were T3s (+0.79), C3s (−0.74), G3s (−0.45), and A3s (+0.40). When the sequences were sorted according to their GC3s

Table 1. RSCU values and number of occurrences for Cys in the three groups of axis 2[a]

	Group 1		Group 2		Group 3	
	N	RSCU	*N*	RSCU	*N*	RSCU
UGU	808	1.96	1585	1.12	25	0.10
UGC	17	0.04	1241	0.88	453	1.90

[a] *N,* number of occurrences; RSCU, relative synonymous codon usage.

we could not detect any common biological trait among those genes displaying similar values except, as noted by Nelson et al. (1999), among a group of sequences encoding lipopolysaccharide biosynthesis proteins, spanning approximately from position 635,000 to position 666,000. Another group of genes with low GC3s, composed mainly of hypothetical proteins conserved among thermophilic prokaryotes and ABC transporters, is clustered around position 1,340,000. Since the majority of the predicted genes in these two regions are most similar to proteins in thermophilic prokaryotes, it seems reasonable to suppose that they were relatively recently acquired by horizontal transfer from thermophilic genomes more biased toward A + T. This phenomenon has been postulated to be very common in *Thermotoga* (Nelson et al. 1999; Logsdon and Faguy 1999).

A very striking result related to the COA of the RSCU data was that the second axis clearly split the sequences into three groups with little overlap among them (Fig. 1). To get a deeper insight into this result we plotted the position of each codon on the plane determined by the first and second axes and found that the triplets displaying the most extreme and opposite values on the second axis were UGU and UGC (synonymous codons coding for Cys), while all the other codons showed essentially no variation along that axis (not shown). This result was confirmed by the patterns of codon usage of the pooled sequences from each group: the only codons displaying different frequencies were again UGU and UGC, and the number of appearances and RSCU values for these codons are listed in Table 1. As can be seen, the first group (positive values along axis 2) is characterized by a strong bias toward UGU; there is a more random usage of both codons in the second group (around zero along axis 2), while in the third group (negative values along axis 2) the strong bias is toward UGC. Therefore, it is rather clear that the factorial load of the second axis comes mainly from triplets coding for Cys.

This result deserves some consideration, since, as far as we know, it is the first time that the relative variation in codon frequencies of only two synonymous codons is so strong that it becomes the second axis of the COA, explaining a relatively important percentage of the total variance among the genes. This observation is probably due to a mathematical artifact, as shown by the following observations. When the position of each sequence along the second axis is plotted against the total number of Cys's of the respective gene (Fig. 2), the distribution of points displays a particular "Christmas tree" grouping, showing the distribution of the different combinations of UGU and UGC for each Cys content. Hence, this behavior is due to two main factors. First is extremely low frequency of Cys in *Thermatoga:* in this organism, Cys represents only 0.7% of all amino acids, which implies only 2.2 residues per protein. Second is the nonbiased composition of the genome (G + C = 46%). Indeed, in a nonskewed genome, the frequencies of usage of two synonymous triplets (in this case, UGC and UGU) will be very similar. But if the amino acid is very scarce, then an important number of genes will code for either one or zero residue, which in terms of RSCU implies that for the rare amino acid most of the sequences can display RSCU values of only 0–0 (no usage), 2–0 (only UGC), or 0–2 (only UGU). In *T. maritima,* 49% of the sequences fall within this category. This distribution of RSCU values is increased when two Cys's are encoded (17% of the genes), since 50% of these sequences will use only one of the synonymous codons. If our interpretation is correct, then a similar result should be observed in every organism with a genomic G + C content of about 50% and with a low frequency of amino acids encoded by two synonymous triplets. Among the completed prokaryotic genomes, the only residue encoded by two synonymous codons and displaying frequencies <1% is Cys. *Thermoplasma acidophilum* (Ruepp et al. 2000) is characterized by a Cys content of 0.6% and a genomic G + C content of 46%, therefore, in these features it is similar to *T. maritima.* In this species, we found that the second axis of the COA again splits the sequences into three clear groups, which differ in their use of UGU and UGC (not shown). On the other hand, when the effect of an extreme GC content was superimposed on a low frequency of Cys, as is the case for *Ureaplasma urealitycum* (Cys = 0.66%, G + C = 27%) and *Borrelia burgdorferi* (Cys = 0.77%, G + C = 28%), this particular distribution was not found. In these cases the variability in the usage of UGU and UGC is very low, thus the contribution to the

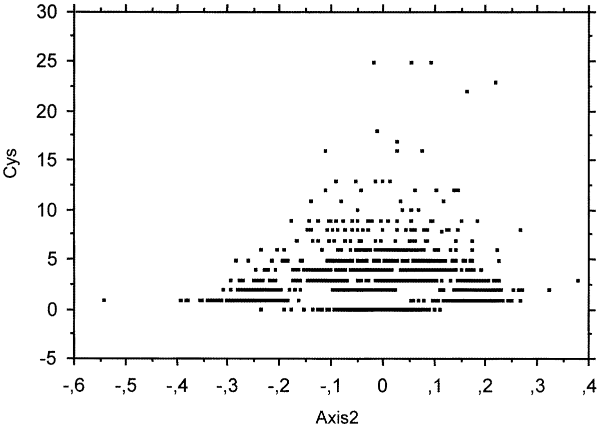

Fig. 2. Plot of the position of each sequence along the second axis of the COA versus the total number of Cys's of the respective gene.

total variance would not be enough to be extracted in the more prominent factors of the COA.

The third axis generated by the COA represented 4.1% of the total variability and seems to be related to expressivity, since at one extreme of the distribution were clustered sequences coding for ribosomal proteins, translation elongation factors, several transporters, cold-shock proteins, groES, single-stranded DNA binding proteins, etc., while genes presumably expressed at lowest levels were scattered throughout the distribution. Therefore, as in most unicellular species, the codon usage of highly expressed genes differs from the codon usage of the rest of the sequences. The most obvious explanation is that translational selection is shaping codon usage in this species.

To understand which synonymous triplets are most frequent among the highly expressed sequences, we compared the codon usage patterns of genes displaying the most extreme values at both ends of this axis. The results of this analysis are listed in Table 2. There are several codons whose usage is significantly higher among the highly expressed genes. For the Y-ending duets and Ile we found that the C-ending codon is always incremented, and these preferences are significant for UUC (Phe), GAC (Asp), UGC (Cys), and AGC (duet coding for Ser). For the quartets and the R-ending duets, the A-ending codons are always optimal, the only exception being CGA (Arg). It is interesting to note that in the case of duets and Ile, the most frequent codon among highly expressed sequences matches perfectly the only tRNA for the corresponding amino acid.

Finally, contrary to results reported for several bacteria (McInerney 1998; Lafay et al. 1999; Romero et al. 2000), we found that the position of each sequence in the leading (53.6% of the total) or in the lagging strand of replication does not affect the codon choices. This is probably due to the fact that this genome is characterized by a very weak GC skew. Indeed, in this species the variation for the skew around zero is ±3.9%, a very low value compared with the 12.5, 10.8, and 17.6% in *C. trachomatis, T. pallidum,* and *B. burgdorferi,* respectively; and in these prokaryotes, as noted above, the main factor shaping codon usage is the strong strand-specific mutational pressure.

Amino Acid Frequencies

To understand the sources of variation at the amino acid level, we applied a COA to the amino acid frequencies of each protein. As far as we know, this approach has been applied only twice to amino acid data. In *Escherichia coli* (Lobry and Gautier 1994) it was found that the three most important sources of variation were the hydrophobicity, expressivity, and aromaticity of the proteins. In the anaerobic unicellular eukaryote *Giardia lamblia* it was recently shown (Garat and Musto 2000) that the most relevant factors are related to the particular mechanism of defenses against reactive oxygen species, namely, the increment of sulfur-containing and aromatic residues. In addition, since the most abundant proteins tend to use smaller amino acids, the cell economy seems to be another prominent feature reducing energetic costs. Finally, as in *E. coli* the third trend in *Giardia* is correlated with the expressivity of each gene, indicating that in these species highly expressed sequences display a tendency to use a subset of the total amino acids preferentially.

When applied to *Thermotoga,* we found that the first four main axes generated by the COA explained 48.4% of the total variance. Figure 3 shows the positions of the sequences along the first two axes. The first trend (18.8% of the variability) is positively correlated ($r = 0.85$, $p < 0.0001$) with the hydropathy level of each protein and therefore negatively correlated ($r = -0.92$, $p < 0.0001$) with the frequency of charged residues. The distribution of the sequences on the plane defined by the two principal axes of the analysis is also shown in Fig. 3. The main axis (horizontal) splits the sequences into two groups, the sequences displaying positive values on axis 1 being the most hydrophobic (probably membrane proteins; 11% of the total), while those with negative values are characterized by high frequencies of Lys and Arg (positively charged residues). It is interesting to note that, as mentioned above, in *E. coli* the first axis discriminates the same features (Lobry and Gautier 1994).

The second axis of the analysis (11.3% of the variability) correlated strongly ($r = 0.68$, $p < 0.0001$) with the MMWs of the amino acids used in each sequence, and in Fig. 3 the sequences displaying negative values on the second axis are the ones with the lowest MMWs. For *G. lamblia,* we have recently reported that the most highly expressed sequences are preferentially constructed with smaller residues, and we postulated that the rationale of this finding might be that smaller amino acids are energetically cheaper than big ones (Garat and Musto 2000). Three correlations suggest that this is the case in *T. maritima* too. First, there is a slight but significant correlation ($r = 0.10$, $p < 0.0001$) between the position of each gene along axis 3, generated from the RSCU data (which discriminates between highly and lowly expressed genes), and the position of the sequences along the axis 2, generated from the amino acid frequencies. The distribution of the sequences on the plane defined by these axes indicates that the most heavily expressed genes display lower MMWs. Second, the CAI of each sequence correlates significantly with the second axis of the amino acid frequencies ($r = 0.37$, $p > 0.0001$) and with the MMW of each protein ($r = 0.21$, $p < 0.0001$). In the two correlations, the distribution of points shows that the sequences with the highest CAI values (probably the most highly expressed) are constructed with smaller residues. Since preliminary results from our laboratory indicate that this phenomenon is

Table 2. Synonymous codon usage in highly and lowly expressed genes in *Thermotoga maritima*[a]

Amino acid	Codon	Highly expressed genes RSCU	Highly expressed genes *N*	Lowly expressed genes RSCU	Lowly expressed genes *N*
Phe	UUU	0.72	(675)	0.82	(850)
	UUC*	1.28	(1207)	1.18	(1232)
Tyr	UAU	0.58	(425)	0.61	(383)
	UAC	1.42	(1035)	1.39	(872)
His	CAU	0.66	(206)	0.75	(214)
	CAC	1.34	(420)	1.25	(355)
Asn	AAU	0.64	(481)	0.67	(415)
	AAC	1.36	(1018)	1.33	(833)
Asp	GAU	1.10	(1088)	1.21	(1033)
	GAC*	0.90	(894)	0.79	(674)
Cys	UGU	1.01	(159)	1.36	(193)
	UGC*	0.99	(156)	0.64	(91)
Gln	CAA*	0.59	(260)	0.45	(153)
	CAG	1.41	(627)	1.55	(520)
Lys	AAA*	1.15	(1892)	1.07	(1489)
	AAG	0.85	(1402)	0.93	(1301)
Glu	GAA*	1.38	(2435)	1.21	(1937)
	GAG	0.62	(1098)	0.79	(1259)
Ile	AUU	0.62	(632)	0.61	(556)
	AUC	1.29	(1313)	1.23	(1126)
	AUA	1.09	(1103)	1.16	(1060)
Val	GUU	1.30	(1098)	1.23	(1095)
	GUC*	0.81	(685)	0.70	(624)
	GUA*	0.55	(462)	0.41	(363)
	GUG	1.35	(1139)	1.66	(1471)
Pro	CCU	1.09	(444)	1.04	(378)
	CCC	0.78	(319)	1.28	(469)
	CCA*	1.40	(572)	0.73	(265)
	CCG	0.72	(295)	0.95	(348)
Thr	ACU	0.68	(319)	0.70	(293)
	ACC	1.01	(476)	1.10	(458)
	ACA*	1.54	(726)	0.91	(379)
	ACG	0.78	(368)	1.29	(540)
Ala	GCU	1.06	(690)	0.99	(538)
	GCC	0.87	(567)	1.07	(580)
	GCA*	1.21	(787)	0.79	(431)
	GCG	0.85	(550)	1.14	(620)
Gly	GGU*	1.23	(890)	1.05	(721)
	GGC	0.45	(326)	0.45	(308)
	GGA*	1.98	(1433)	1.80	(1236)
	GGG	0.33	(240)	0.70	(480)
Leu	UUA*	0.31	(193)	0.20	(129)
	UUG	0.72	(456)	0.76	(500)
	CUU	1.51	(949)	1.50	(983)
	CUC	2.04	(1285)	1.90	(1247)
	CUA*	0.21	(131)	0.15	(97)
	CUG	1.21	(763)	1.49	(978)
Ser	AGU	0.78	(257)	0.87	(331)
	AGC*	1.18	(389)	0.76	(288)
	UCU	1.33	(440)	1.37	(522)
	UCC	1.17	(388)	1.32	(501)
	UCA*	1.08	(356)	0.61	(230)
	UCG	0.47	(156)	1.07	(407)
Arg	AGA*	4.18	(1423)	2.18	(756)
	AGG	1.41	(479)	1.73	(601)
	CGU	0.17	(58)	0.49	(171)
	CGC	0.07	(25)	0.37	(129)
	CGA	0.14	(48)	0.51	(176)
	CGG	0.02	(8)	0.71	(246)

[a] Comparison of codon usage frequencies between highly and lowly expressed sequences, as discriminated by the third axis of the COA. *N*, number of occurrences; RSCU, relative synonymous codon usage.

* Significantly more frequent among highly expressed genes ($p < 0.01$) according to a χ^2 test.

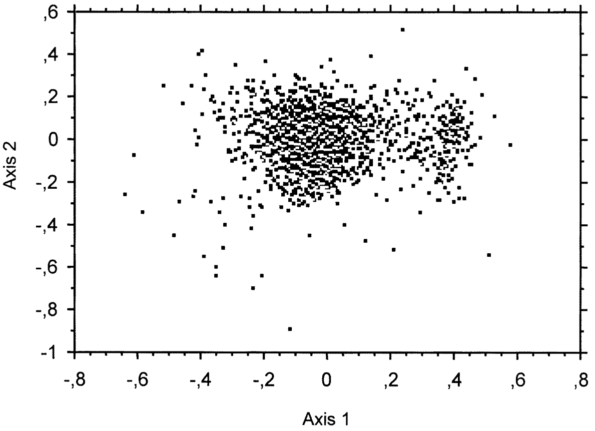

Fig. 3. Plot of the two most prominent axes generated from the COA of the amino acid frequencies for the proteins coded by the *T. maritima* genome.

very common, we propose that one of the main forces driving the construction of proteins is the energetic cost of the amino acids, and the most abundant proteins (encoded by highly expressed genes) tend to display the highest proportion of small (and therefore cheaper) residues.

The third axis of the analysis (10.6% of the variability) correlated significantly ($r = 0.51$, $p < 0.0001$) with the aromaticity of each protein. In *E. coli* the axis related to this feature was the third (Lobry and Gautier 1994), while in *G. lamblia* it was the second (Garat and Musto 2000). In trying to understand the biological causes of this correlation it should be taken into account, as noted by Lobry and Gautier (1994), that (a) aromatic residues (Phe, Tyr, and Trp) are variable among proteins (although usually rare), and (b) their biosynthesis is energetically expensive. Therefore, the relatively high percentage of variation related to the use of these particular residues might be the consequence of two opposite trends: a selective constraint against their use, to save energy, and their necessity in some proteins. If this is indeed the case, it is interesting to note that the cell economy might be related to two axes of the amino acid frequencies in *Thermotoga,* namely, the second and the third (see the previous paragraph). Together, these two axes represent almost 22% of the total variability.

Perhaps more interesting is that when the frequencies of Cys, Met, and His are added to the frequencies of the above-mentioned aromatic residues, the *r* value of the correlation with the position of the sequences along the third axis is higher ($r = 0.71$). These six residues have in common the property of being the preferential targets of reactive oxygen species and can act as sinks for radical fluxes through electron transfer between amino acids on the protein (Dean et al. 1993; Berlett and Stadtman 1997). It has been suggested that these residues could confer some resistance to reactive oxygen species in the anaerobic unicellular eukaryote *G. lamblia* (Garat and Musto 2000). It is not easy to understand this result in *T.*

maritima since its exposure to reactive oxygen species in its natural habitat is rather improbable. It is interesting to note that there is a slight but significant negative correlation between the frequency of usage of these residues and the CAI of each gene ($r = -0.20, p < 0.0001$), which suggests that the avoidance of these six amino acids is stronger among highly expressed genes.

The fourth axis of the COA on the amino acid frequencies (7.7% of the variability) correlates strongly ($r = 0.71, p < 0.0001$) with the Cys content of each protein. As mentioned in the section on codon usage, this residue is the least frequent of all amino acids in this species (0.72% of the total residues). However, the variability in the percentage of this residue among the proteins is relatively high, ranging from 0% (26% of the total proteins) to more than 12%. In the group of proteins characterized by a Cys content >5%, there are ribosomal proteins which display zinc finger domains and several iron binding proteins in whose structure the Cys is indispensable.

Finally, we analyzed the amino acid frequencies in the genes placed on the leading or lagging strand of replication, and as expected given the weak GC skew of this genome (see the previous section), no significant difference was found.

In summary, we found that the patterns of synonymous codon usage and amino acid frequencies in *T. maritima* are the result of several factors. At the codon usage level we detected an increment of several codons among highly expressed genes, therefore translational selection apparently operates in this bacterium. An interesting result was that the variation of usage of only two synonymous codons (UGU and UGC, coding for Cys) was detected by the COA as one of the most prominent sources of variation; this result is probably an artifact due to the very low frequency of usage of Cys together with the nonbiased composition of this genome. At the amino acid level, first, the hydropathy (and the frequency of charged residues) is the main factor driving the architecture of the proteins, a feature that appears to be universal. Second, the economy of the cell seems to be very important since presumably highly expressed sequences are constructed with smaller residues and avoid aromatic amino acids. Third, the anaerobic condition of *Thermotoga* influences the frequency of usage of the amino acids that are the preferential targets of reactive oxygen species. Finally, the fourth trend appears to be related to the usage of Cys, which is, as mentioned, the least frequent amino acid.

References

Akashi H, Eyre-Walker A (1998) Translational selection and molecular evolution. Curr Opin Genet Dev 8:688–693

Berlett BS, Stadtman ER (1997) Protein oxidation in aging, disease, and oxidative stress. J Biol Chem 272:20313–20316

Dean RT, Gieseg S, Davies MJ (1993) Reactive species and their accumulation on radical-damaged proteins. Trends Biochem Sci 18:437–441

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

Garat B, Musto H (2000) Trends of amino acid usage in the proteins from the unicellular parasite *Giardia lamblia.* Biochem Biophys Res Commun 279:996–1000

Greenacre M (1984) Theory and applications of correspondence analysis. Academic, London

Kerr AR, Peden JF, Sharp PM (1997) Systematic base composition variation around the genome of *Mycoplasma genitalium,* but not *Mycoplasma pneumoniae.* Mol Microbiol 125:1177–1179

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

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

Logsdon JM, Faguy DM (1999) *Thermotoga* heats up lateral gene transfer. Curr Biol 9:R747–R751

Lopez P, Forterre P, le Guyader H, Philippe H (2000) Origin of replication of *Thermotoga maritima.* Trends Genet 16:59–60

McInerney JO (1997) Prokaryotic genome evolution as assessed by multivariate analysis of codon usage patterns. Microb Comp Genomics 2(1):1–10

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

Nelson KE, Clayton RA, Gill SR, et al. (1999) Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima.* Nature 399:323–329

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

Ruepp A, Graml W, Santos-Martinez ML, Koretke KK, Volker C, Mewes HW, Frishman D, Stocker S, Lupas AN, Baumeister W (2000) The genome sequence of the thermoacidophilic scavenger *Thermoplasma acidophilum.* Nature 407:508–513

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

Sharp PM, Tuohy TM, Mosurski KR (1986) Codon usage in yeast: Cluster analysis clearly differentiates highly and lowly expressed genes. Nucleic Acids Res 14:5125–5143

Sharp PM, Averof M, Lloyd AT, Matassi G, Peden JF (1995) DNA sequence evolution: The sounds of silence. Philos Trans R Soc Lond B Biol Sci 349:241–247

Tiboni O, Cantoni R, Creti R, Cammarano P, Sanangelantoni AM (1991) Phylogenetic depth of *Thermotoga maritima* inferred from analysis of the fus gene: Amino acid sequence of elongation factor G and organization of the *Thermotoga* str operon. J Mol Evol 33:142–151

Wright F (1990) The ‘effective number of codons’ used in a gene. Gene 87:23–29