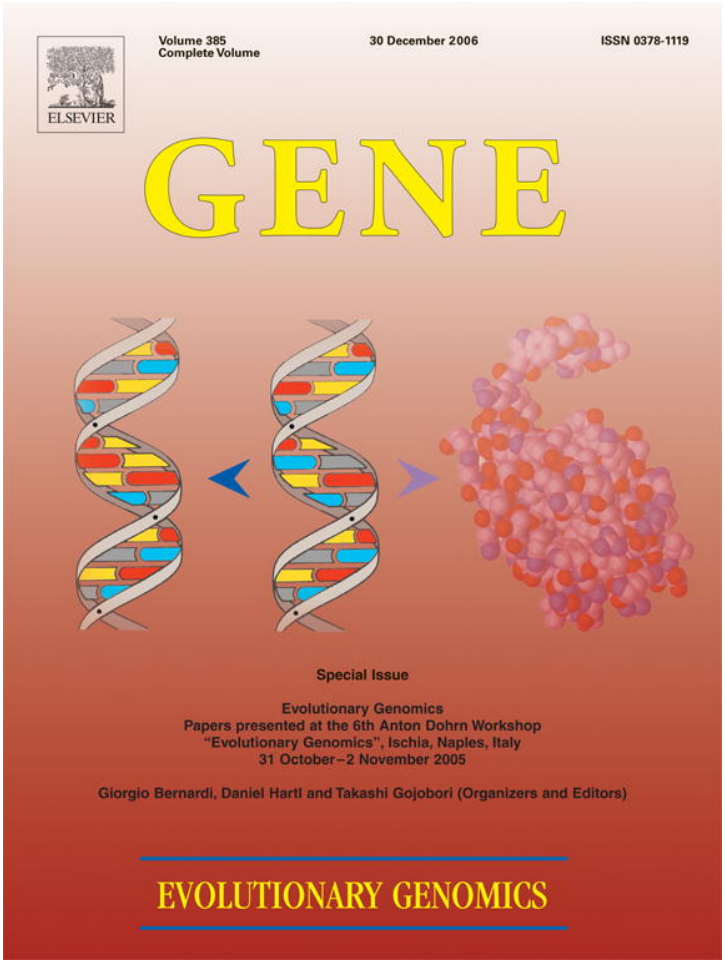

Available online at www.sciencedirect.com
ScienceDirect
Gene 385 (2006) 128–136
GENE
SECTION
EVOLUTIONARY GENOMICS
www.elsevier.com/locate/gene

Synonymous codon usage and its potential link with optimal growth temperature in prokaryotes

J.R. Lobry *, A. Necşulea

Laboratoire de Biométrie et Biologie Evolutive (UMR 5558), France
CNRS, France
Univ. Lyon 1, 43 bd 11 nov, 69622, Villeurbanne Cedex, France
HELIX, Unité de recherche Inria, France

Received 10 January 2006; accepted 29 May 2006
Available online 22 August 2006

Abstract

The relationship between codon usage in prokaryotes and their ability to grow at extreme temperatures has been given much attention over the past years. Previous studies have suggested that the difference in synonymous codon usage between (hyper)thermophiles and mesophiles is a consequence of a selective pressure linked to growth temperature.

Here, we performed an updated analysis of the variation in synonymous codon usage with growth temperature; our study includes a large number of species from a wide taxonomic and growth temperature range. The presence of psychrophilic species in our study allowed us to test whether the same selective pressure acts on synonymous codon usage at very low growth temperature.

Our results show that the synonymous codon usage for Arg (through the AGG, AGA and CGT codons) is the most discriminating factor between (hyper)thermophilic and non-thermophilic species, thus confirming previous studies. We report the unusual clustering of an Archaeal psychrophile with the thermophilic and hyperthermophilic species on the synonymous codon usage factorial map; the other psychrophiles in our study cluster with the mesophilic species.

Our conclusion is that the difference in synonymous codon usage between (hyper)thermophilic and non-thermophilic species cannot be clearly attributed to a selective pressure linked to growth at high temperatures.

Keywords: Synonymous codon usage; Optimum growth temperature; Thermophiles; Psychrophiles; AGG

1. Introduction

Prokaryotic organisms can occupy ecological niches best characterized as extreme, with respect to environmental factors such as temperature, acidity, salt concentration, hydrostatic pressure etc. It has long been known that some prokaryotes are able to live at very high temperatures (Brock, 1967); how these organisms have adapted to such extreme environments has been an equally hot topic ever since.

With the increasing availability of genomic sequences, it has become possible to investigate the existence of adaptive traits in the genomes and proteomes of these extremophiles. At the nucleotide level, it has been established that thermophilic genomes are not particularly enriched in guanine and cytosine, although G:C pairs are more thermally stable (Galtier and Lobry, 1997; Hurst and Merchant, 2001). This enrichment in guanine and cytosine is nevertheless observed for transfer and ribosomal RNAs of thermophilic and hyperthermophilic organisms (Galtier and Lobry, 1997). A recent study has reported the existence of correlations between genomic G+C content and optimal growth temperature when families of prokaryotes are considered (Musto et al., 2004), however, this aspect remains very controversial (Marashi and Ghalanbor, 2004; Basak et al., 2005a; Musto et al., 2005). The optimum growth temperature of prokaryotes has been shown to correlate with another feature of nucleotide composition, namely purine content. Compared to mesophilic species, thermophilic genomes are significantly enriched in purines and in purine-clusters, a possible explanation

Abbreviations: T_{opt}; optimal growth temperature.

* Corresponding author. Tel.: +33 472431287; fax: +33 472431388.
E-mail address: lobry@biomserv.univ-lyon1.fr (J.R. Lobry).

0378-1119/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2006.05.033

being the existence of a selective pressure to avoid undesirable RNA–RNA interactions (Lao and Forsdyke, 2000; Lobry and Chessel, 2003; Lambros et al., 2003; Paz et al., 2004).

Unlike their mesophilic homologues, the proteins of thermophilic organisms are stable and functional at very high temperatures; numerous studies have focused on understanding the mechanisms of protein thermostabilization in these species. It has been reported that the protein sequences of thermophilic species are shorter than the mesophilic ones, due to a loop-deletion mechanism that increases stability (Thompson and Eisenberg, 1999). Analyses of the amino-acid composition in thermophiles have shown that there is a slight increase of charged residues (Glu, Arg, Lys) and a decrease of polar uncharged residues with growth temperature (Asn, Gln, Ser, Thr) (Vieille and Zeikus, 2001; Kreil and Ouzounis, 2001; Tekaia et al., 2002), a reduction in the frequency of thermo-labile amino-acids (His, Gln and Thr) (Singer and Hickey, 2003; Hickey and Singer, 2004), and an increase in the (Glu + Lys)/(Gln + His) ratio (Farias and Bonato, 2003); the amino-acid usage is therefore a discriminant factor between thermophilic and mesophilic species (Kreil and Ouzounis, 2001). At the structural level, it appears that the presence of salt bridges and the ionic interactions are key factors for protein thermostability (Kumar and Nussinov, 2001; Vetriani and Maeder, 2005).

Another important aspect of the genomic composition of thermophilic species is the pattern of synonymous codon usage in protein-coding sequences. The relative frequencies of synonymous codons vary among different genes of the same genome, as well as among genomes, as a result of the interference between mutational bias and selective pressures (Grantham et al., 1980; Bulmer, 1991). The pattern of synonymous codon usage is different between thermophilic and mesophilic species, the strongest effect being produced by Arginine and Isoleucine codons; (hyper)thermophiles use more frequently the AGG, ATA and AGA codons and avoid CGT and CGA (Lynn et al., 2002; Lobry and Chessel, 2003; Singer and Hickey, 2003). The most likely explanation for this difference in synonymous codon usage has been considered to be the existence of a selective pressure linked to growth temperature, although its nature could not be clearly identified (Lynn et al., 2002; Lobry and Chessel, 2003). For amino-acids encoded by two codons, (hyper)thermophiles use more frequently the G/C-ending codon (Lobry and Chessel, 2003). A recent study (Basak et al., 2005b) has suggested as a possible explanation the "right choice" hypothesis (Grosjean and Fiers, 1982; Gouy and Gautier, 1982), that states the existence of a selective pressure in favor of a codon–anticodon pairing energy of intermediate strength.

Among the statistical methodologies used to determine genomic characteristics linked to growth temperature, multivariate approaches (such as principal component analysis (Thompson and Eisenberg, 1999; Kreil and Ouzounis, 2001) or correspondence analysis (Tekaia et al, 2002; Lobry and Chessel, 2003; Singer and Hickey, 2003)) have proven to be very effective. Indeed, a major issue when analyzing compositional genomic features (such as amino-acid and synonymous codon usage) is the confounding effect of the guanine and cytosine content (Lobry, 1997); multivariate analyses allow the removal of this influence by a projection on the orthogonal space. Over the years, the number of species analyzed in such studies has strongly increased, from 20 in Thompson and Eisenberg (1999) to 293 in Lobry and Chessel (2003), thus making necessary the use of multivariate techniques.

The amount of available genomic data increases rapidly, and so does our knowledge on the possible growth temperatures of prokaryotes; currently, the known range for Bacterial and Archaeal species is very wide, starting at less than −10 °C (Methé et al., 2005) and ending at more than 120 °C (Kashefi and Lovley, 2003). Our main purpose here is to perform an updated analysis of the variation of the patterns of synonymous codon usage in prokaryotes, with respect to their optimal growth temperature, by taking advantage of the large number of genomic sequences currently available.

We were able to include in our analysis species from a wide taxonomic and growth temperature range, such as psychrophilic species, more mesophilic Archaeal species and more thermophilic Bacterial species than used in previous studies.

We focused on the pattern of synonymous codon usage, by specifically analyzing the frequency variation of the salient codons with growth temperature and base composition. The significance of the observed patterns is discussed, and we propose an alternative explanation for the differences in codon usage between thermophilic and non-thermophilic species.

2. Materials and methods

2.1. Codon usage data

Our dataset consisted of prokaryotic species for which at least 50 kb of complete coding sequences was publicly available. The data were extracted from the international nucleotide sequence databases (DDBJ (Tateno et al., 2005), EMBL (Kanz et al., 2005), GenBank (Benson et al., 2005)), structured under the ACNUC system (Gouy et al., 1985), on January 2006.

Coding sequences were extracted and analyzed with the R (R Development Core Team, 2005) package seqinR (Charif and Lobry, in press). All sequences annotated as partial were removed. Coding sequences with less than 300 bp were removed in an attempt to decrease the proportion of ELFs (Ochman, 2002), since the expected average reading-frame lengths in random sequences are about 40 bp and 200 bp for 25% and 75% G+C content, respectively (Oliver and Marin, 1996). Species using non-standard genetic codes (*viz.* *Mesoplasma*, *Mycoplasma*, *Spiroplasma*, *Ureaplasma*) were removed because they cannot be analyzed simultaneously with within- and between-correspondence analyses, this would require further methodological developments (Anne Dufour, personal communication). Stop codons (*viz.* TAA, TAG, TGA) were removed, but not the start codon for translation.

It has been estimated (Zavala et al., 2005) that a sample of 10 genes (i.e. roughly 10 kb) is sufficient to estimate prokaryotic genomic G+C content. We are unaware of such a systematic study for codon and amino-acid usage estimations; we have therefore decided to keep only species with a total of at least

50 kb of coding sequences. We made an exception to incorporate *Cenarchaeum symbiosum* with 47.649 kb of available data, because it was the only psychrophilic Archaeum available at that time. We checked that results remained similar when considering only the 294 complete prokaryotic genomic sequences available from the NCBI repository on the 6th of January 2006, and that they were consistent with previously published multivariate analyses for codon usage (Lobry and Chessel, 2003; Carbone et al., 2005), synonymous codon usage (Lynn et al., 2002; Lobry and Chessel, 2003; Singer and Hickey, 2003; Chen et al., 2004) and non-synonymous codon (*i.e.* amino-acid) usage (Thompson and Eisenberg, 1999; Kreil and Ouzounis, 2001; Tekaia et al., 2002; Dumontier et al., 2002; Lobry and Chessel, 2003; Singer and Hickey, 2003) between prokaryotic species.

We have also removed one already described (Lobry and Chessel, 2003) outlier, *Candidatus* Carsonella ruddii, caused by a gene sampling bias. No other systematic gene sampling bias was found.

The analyzed dataset was composed of 559,514,732 codons, 739 strains, 457 species and 219 genera. As compared with a previous study with a June 2002 similar dataset (Lobry and Chessel, 2003), this was about 5.5 times more data in terms of codon counts, but only about 2 times more in terms of strain sampling, because of a larger proportion of complete genome sequences in the most recent data.

In order to verify the consistence of the results when only highly expressed genes are considered, we built a second codon usage dataset by querying the nucleotide sequence databases for ribosomal proteins. 249 species (481 strains) were represented in this dataset, with a total of 4,382,865 codons (TAA, TAG and TGA excluded).

2.2. Optimum growth temperature data

The information concerning optimum growth temperatures (T_{opt}) of prokaryotes was compiled from three distinct sources: the German National Resource Centre for Biological Material (DSMZ, http://www.dsmz.de/), the Prokaryotic Growth Temperature Database (PGTdb, http://pgtdb.csie.ncu.edu.tw/), and the data from Galtier and Lobry (1997), extracted from the Bergey's (1984–1989) Manual of Systematic Bacteriology (ftp://pbil.univ-lyon1.fr/pub/datasets/JME97/species). For T_{opt} values where there was a general agreement between the sources, we used the average of the available values. When neither of these sources could provide the required information for a certain species, or when the difference in T_{opt} between the sources was larger than 5 °C (see Section 3.1 for details), we referred to the primary literature (Table 1). If a search in the literature could not provide an unambiguous answer, the information given by the DSMZ was used.

We assigned species into temperature classes according to the following rules: psychrophiles are characterized by a $T_{opt} \leq 20$, for mesophiles $T_{opt} > 20$ and < 59, for thermophiles $T_{opt} \geq 59$ and < 80 and for hyperthermophiles $T_{opt} \geq 80$.

With this convention, our dataset was composed of 6 psychrophilic species (8 strains), 415 mesophilic species (676 strains), 16 thermophilic species (23 strains) and 20 hyperthermophilic species (32 strains). This temperature class distribution is not independent of our choices for the growth temperature data; for example, if the data had been extracted solely from PGTdb, the number of species termed psychrophiles would have been equal to 9, with only 3 species in common with our current psychrophilic dataset.

Table 1
List of discrepancies between the growth temperature data sources (difference of at least 10 °C)

Species	T_{opt} DSMZ	T_{opt} PGTdb	T_{opt} Galtier and Lobry (1997)
1. *Bacillus halodurans*	30	45	NA
2. *Bacillus megaterium*	30	10	50.3
3. *Bacillus pumilus*	30	10	NA
4. *Campylobacter fetus*	37	25	37
5. *Haloarcula marismortui*	37	50	NA
6. *Halobacterium salinarum*	37	50	50
7. *Klebsiella oxytoca*	37	10	NA
8. *Lactobacillus acidophilus*	34	45	NA
9. *Lactobacillus casei*	30	15	NA
10. *Lactobacillus plantarum*	30	15	NA
11. *Listeria innocua*	37	22	33.5
12. *Streptomyces antibioticus*	28	45	NA
13. *Streptomyces aureofaciens*	28	45	NA
14. *Streptomyces rochei*	28	45	NA
15. *Streptomyces viridochromogenes*	28	45	NA
16. *Xanthomonas campestris*	26	37	NA

The taxonomic distribution of thermophilic and hyperthermophilic species is biased in favor of the Archaeal domain; there are 15 hyperthermophiles and 7 thermophiles among the Archaea included in our dataset, compared to 5 hyperthermophiles and 9 thermophiles in Bacteria. There is however a slight increase in the number of available mesophilic Archaea — 10 species (15 strains) in the present study as compared to 6 species in Lobry and Chessel (2003). The psychrophilic dataset is composed of 5 Bacterial species and 1 Archaeal species.

The list of available thermophilic and hyperthermophilic species is given in Table 2, the list of available psychrophilic species is given in Table 3.

2.3. Multivariate analyses

The patterns of codon usage were analyzed using the within–between-correspondence analysis method, as described elsewhere (Lobry and Chessel, 2003). This method allows the decomposition of global codon usage variability into effects at the synonymous level (within-classes correspondence analysis) and at the non-synonymous level (between-classes correspondence analysis).

The codon usage dataset consisted of a contingency table with 739 rows (corresponding to different organisms) and 61 columns (corresponding to sense codons). The correspondence analyses were performed on the absolute codon frequencies. The codon usage variability was analyzed solely at the between-genomes level; the variability at the within-genome level was ignored, since it is irrelevant for our present concern.

The statistical computations were performed using the ade4 (Thioulouse et al., 1997) package in R (R Development Core Team, 2005).

2.4. Reproducibility

The results presented in this paper are reproducible online at http://pbil.univ-lyon1.fr/members/lobry/repro/gene06/.

3. Results

3.1. Optimal growth temperatures

There was a general agreement between the three sources of data for T_{opt}. Out of the 127 T_{opt} values documented in both the DSMZ and PGTdb databases, 59 were found to be equal, 87 to differ by at most 5 °C, and 106 to differ by at most 10 °C. Out of the 69 T_{opt} values documented in both the DSMZ database and in Galtier and Lobry (1997), 18 were found to be equal, 58 to differ by at most 5 °C, and 67 to differ by at most 10 °C. Out of the 24 T_{opt} values documented in both the PGTdb database and in Galtier and Lobry (1997), 14 were found to be equal, 21 to differ by at most 5 °C, and 21 to differ by at most 10 °C.

However, the optimum temperatures were not always consistent between the three sources; for several species the differences between the three sources exceeded 10 °C (see Table 1). Discrepancies appeared to be concentrated in few genera (*viz.* *Bacillus*, *Lactobacillus*, *Streptomyces*, *Xanthomonas*) accounting for 12 out of the 17 observed cases. In some cases, the difference between the three sources was large enough to affect the predicted temperature class of the organism (i.e. psychrophile, mesophile, thermophile or hyperthermophile). An extreme case is *Bacillus megaterium*, which is a mesophile according to DSMZ, a psychrophile according to PGTdb and a moderate thermophile according to Galtier and Lobry (1997). In these cases, if a search in the primary literature could not provide an unambiguous answer, the information given by the DSMZ was used.

Table 2
List of thermophilic species ($T_{opt} \geq 59.0$ °C) in the analyzed dataset

Species	T_{opt}	Kingdom
1. *Acidianus ambivalens*	80.0	Archaea
2. *Aeropyrum pernix*	92.5	Archaea
3. *Aquifex aeolicus*	95.0	Bacteria
4. *Aquifex pyrophilus*	85.0	Bacteria
5. *Archaeoglobus fulgidus*	85.0	Archaea
6. *Caldicellulosiruptor saccharolyticus*	65.0	Bacteria
7. *Carboxydothermus hydrogenoformans*	65.0	Bacteria
8. *Chloroflexus aurantiacus*	60.0	Bacteria
9. *Clostridium thermocellum*	61.0	Bacteria
10. *Methanobacterium thermoautotrophicum*	65.0	Archaea
11. *Methanocaldococcus jannaschii*	85.0	Archaea
12. *Methanopyrus kandleri*	98.0	Archaea
13. *Methanothermobacter thermautotrophicus*	65.0	Archaea
14. *Nanoarchaeum equitans*	90.0	Archaea
15. *Picrophilus torridus*	60.0	Archaea
16. *Pyrobaculum aerophilum*	98.0	Archaea
17. *Pyrococcus abyssi*	98.0	Archaea
18. *Pyrococcus furiosus*	98.5	Archaea
19. *Pyrococcus horikoshii*	96.5	Archaea
20. *Rhodothermus marinus*	65.0	Bacteria
21. *Sulfolobus acidocaldarius*	72.5	Archaea
22. *Sulfolobus islandicus*	80.0	Archaea
23. *Sulfolobus solfataricus*	87.0	Archaea
24. *Sulfolobus tokodaii*	77.5	Archaea
25. *Symbiobacterium thermophilum*	60.0	Bacteria
26. *Thermoanaerobacter tengcongensis*	75.0	Bacteria
27. *Thermococcus kodakarensis*	95.0	Archaea
28. *Thermococcus litoralis*	85.0	Archaea
29. *Thermoplasma acidophilum*	59.5	Archaea
30. *Thermoplasma volcanium*	60.0	Archaea
31. *Thermoproteus tenax*	85.0	Archaea
32. *Thermotoga maritima*	80.0	Bacteria
33. *Thermotoga neapolitana*	85.0	Bacteria
34. *Thermotoga* sp.	80.0	Bacteria
35. *Thermus aquaticus*	70.0	Bacteria
36. *Thermus thermophilus*	75.0	Bacteria

Table 3
List of psychrophilic species ($T_{opt} \leq 20.0$ °C) in the analyzed dataset

Species	T_{opt}	Kingdom
1. *Cenarchaeum symbiosum*	10.0	Archaea
2. *Colwellia psychrerythraea*	10.0	Bacteria
3. *Desulfotalea psychrophila*	7.0	Bacteria
4. *Photobacterium profundum*	10.0	Bacteria
5. *Planktothrix agardhii*	17.0	Bacteria
6. *Psychrobacter arcticus*	10.0	Bacteria

3.2. Synonymous codon usage

Most (87.8%) of the variability for codon usage was found at the synonymous level, as expected. At the synonymous level, the first and second factor accounted for 71.1% and 8.1% of the total variability, respectively. These values were close to the 68.3% and 11.0% reported from a smaller dataset (Lobry and Chessel, 2003). As expected for a between Bacterial genomes

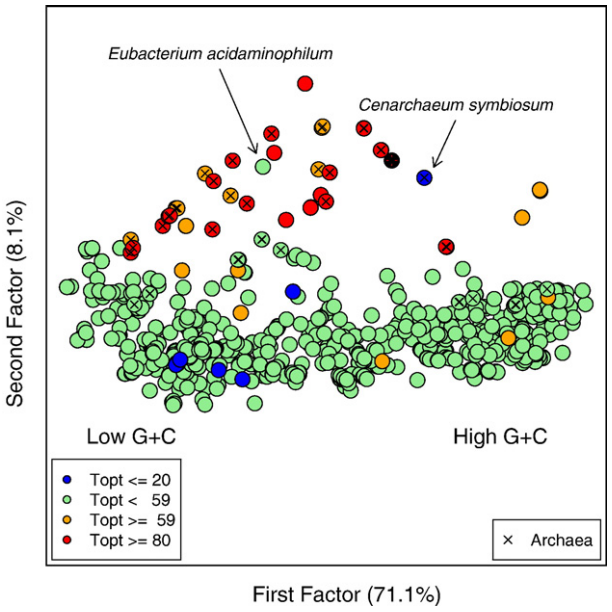

Fig. 1. First factorial map for synonymous codon usage analysis.

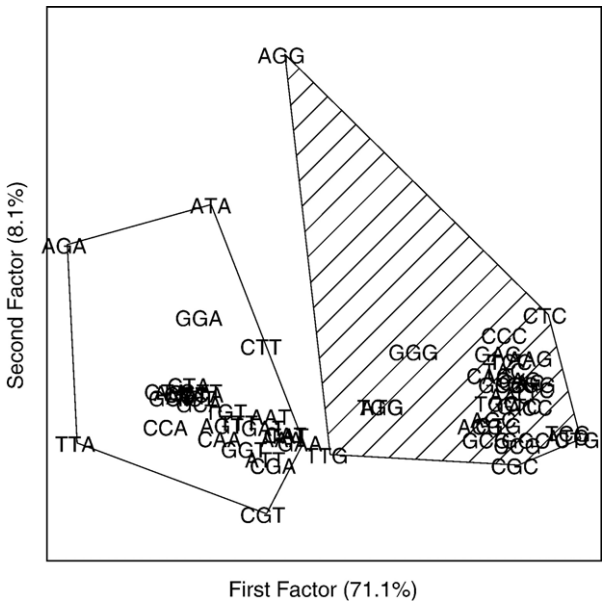

Fig. 2. First factorial map in codon space for synonymous codon usage analysis. The white area encloses all AT-ending codons while the dashed area encloses all G/C-ending codons.

study, the first factor was found to be the G+C content (r^2=0.985; the G+C content was estimated from the available sequence data for each organism).

The distribution of species on the first synonymous factorial map (Fig. 1) had a general shape similar to what has been published previously (Lynn et al., 2002; Lobry and Chessel, 2003; Singer and Hickey, 2003; Chen et al., 2004) with some species discriminated by the second factor. Hyperthermophilic species ($T_{opt} \geq 80$ °C) were found to be well discriminated by the second factor. Thermophiles were found to be less well discriminated by this factor because of four species (*Carboxydothermus hydrogenoformans*, T_{opt}=65 °C; *Chloroflexus auriantiacus*, T_{opt}=60 °C; *Rhodothermus marinus*, T_{opt}=65 °C and *Symbiobacterium thermophilum*, T_{opt}=60 °C) that cluster with the mesophilic group. An already mentioned mesophilic outlier (Lobry and Chessel, 2003) is *Eubacterium acidaminophilum*, which is situated within the thermophilic group (top-center in Fig. 1). Although the data for *C. symbiosum* are scarce (*cf* Materials and methods), it is worth mentioning that this psychrophilic Archaeum clustered clearly with the thermophilic group. Bacterial psychrophilic species do not have extremely low coordinates on the second factor but rather cluster with the mesophilic group (at the bottom of Fig. 1, some mesophilic species have lower coordinates than the psychrophilic species).

The kingdom (*i.e.* Archaea *vs.* Bacteria) was only partly discriminated by the second factor. Most Archaeal species are situated on the top of the map, and most Bacterial species at the bottom, but there are exceptions to this trend: some Archaea are placed at the bottom of Fig. 1 (*viz. Haloarcula marismortui*,

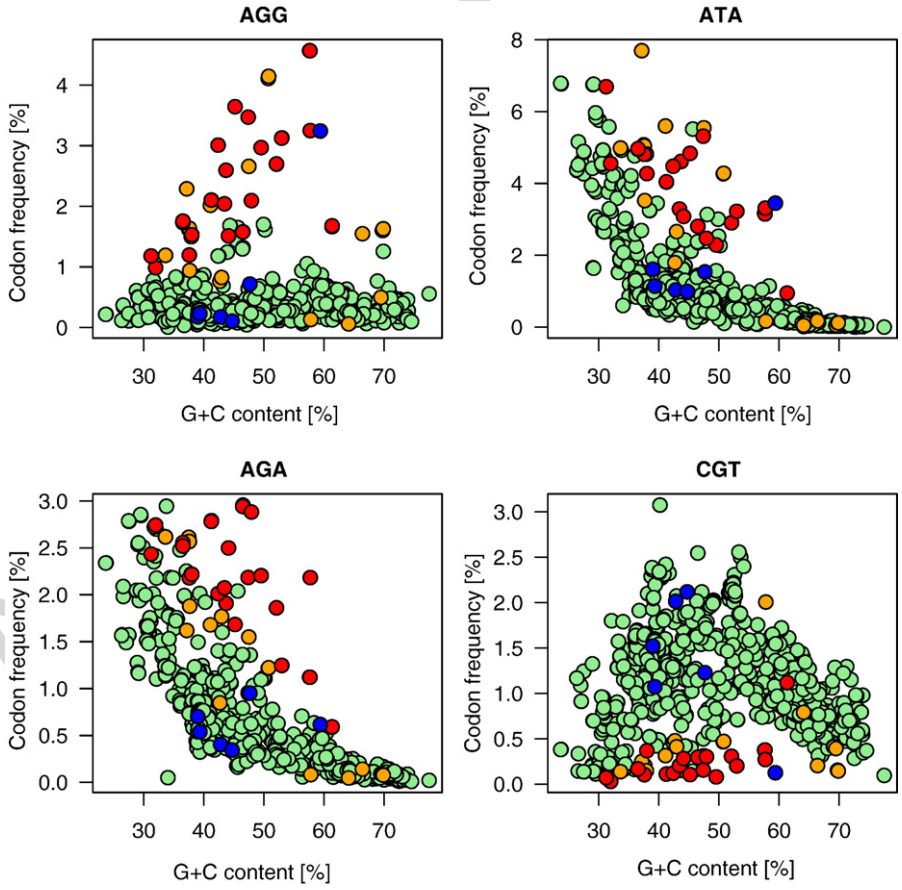

Fig. 3. Relationship between codon frequencies and G+C content for the most salient codons on second factor in Fig. 2. The color code is the same as in Fig. 1.

Halobacterium salinarum, *Haloferax volcanii*, *Methanococcus maripaludis*, *M. voltae*, *Natronomonas pharaonis*) and conversely some Bacteria at the top (*viz. Aquifex aeolicus*, *A. pyrophilus*, *E. acidaminophilum*, *Thermotoga maritima*, *T. neapolitana*, *Thermus thermophilus*). The second factor for synonymous codon usage variability among prokaryotes was therefore more related to the optimal growth temperature than to the kingdom in the analyzed dataset.

The distribution of codons on the first synonymous factorial map is represented in Fig. 2. The first factor was found to discriminate between G+C-poor and G+C-rich codons as expected, and the pattern was found to be stable as compared to a previous study on a smaller dataset (Lobry and Chessel, 2003). On the second factor, the AGG codon for Arg was found to be the most specific of thermophiles (on the top of the dashed area in Fig. 2). This codon had also the smallest coordinate of all G+C-rich codons on the first factor, meaning that its behavior with respect to G+C content was atypical. Three other salient codons on the second factor were ATA (Ile), AGA (Arg) and CGT (Arg).

3.3. Discriminating codons and G+C content

Fig. 3 displays the relationship between codon frequencies and G+C content for the salient codons mentioned above. The G+C-rich codon AGG was found to be systematically avoided in most psychrophilic and mesophilic species with frequencies typically below 1%, and a noticeable absence of frequency increase with G+C content (with the exception of *C. symbiosum*). In contrast, the frequency of codon AGG in thermophilic and hyperthermophilic species was found to be much higher, typically exceeding 1% and reaching values close to 5%, and increasing with G+C content. The behavior of the AT-rich codons ATA and AGA was more regular with a rarefaction with G+C content up to an almost complete absence in extremely G+C-rich organisms. Both ATA and AGA codons were found to be more frequent, for a given G+C content, in thermophilic and hyperthermophilic species. The G+C-rich CGT codon appeared to have an atypical behavior in mesophilic and psychrophilic species, because it does not increase with G+C content but tends to be avoided at extreme G+C contents, with frequencies up to 3% in organisms with average G+C content. Furthermore, there was a trend for CGT to be avoided in thermophilic and hyperthermophilic organisms, with frequencies typically below 0.5%.

3.4. Right choice hypothesis

According to the right choice hypothesis (Grosjean and Fiers, 1982; Gouy and Gautier, 1982), for codons recognized by the

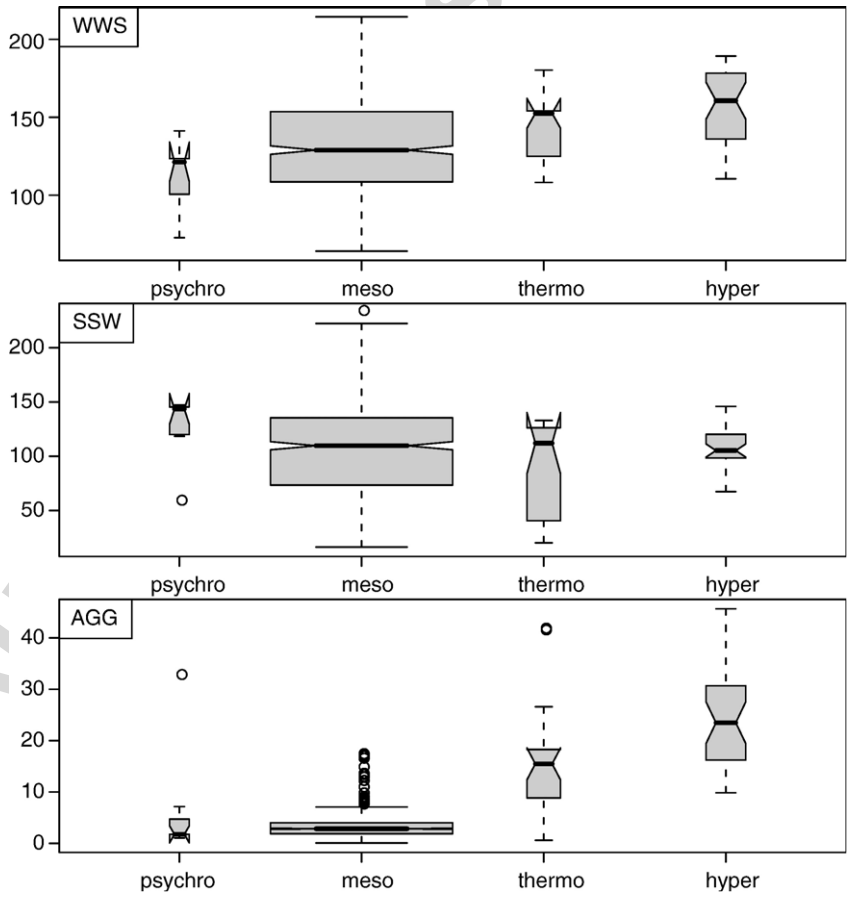

Fig. 4. Relationship between codon frequencies (in per thousand) and optimal growth temperature for WWS codons (top), SSW codons (middle) and AGG codons (bottom). The widths of the boxes are proportional to the square-roots of the number of observations in each group. The notches drawn on the boxes represent confidence intervals for the median value of each group.

same tRNA, the optimal codon is such that the codon–anticodon pairing energy is of intermediate strength. This rule applies to codons that have in the first positions either two "strong" bases (S, *i.e.* G or C), or two "weak" bases (W, *i.e.* A or T), and that end in C or T; the preferred codons are of the form WWS and SSW. A recent paper by Basak et al. (2005b) has suggested that this hypothesis could provide an explanation for the pattern of synonymous codon usage in thermophilic prokaryotes; according to the results of Basak et al. (2005b), there are significant correlations between the frequency of WWS codons and growth temperature. Here, we wanted to verify whether these relations remain valid when a larger dataset of prokaryotic species is analyzed, and whether the "right choice" hypothesis can provide an accurate explanation for the observed patterns.

The frequencies of WWS codons were found to increase slightly but significantly with temperature (top of Fig. 4), and no clear trend was found for SSW codons (middle of Fig. 4). However, the increase in frequency for WWS codons with growth temperature was relatively small as compared to what is observed for AGG codons (bottom of Fig. 4). As stated previously, the CGT codon for Arginine is avoided in thermophilic and hyperthermophilic species; this is an argument against the "right choice" hypothesis as an explanation for the pattern of synonymous codon usage at high growth temperature. Indeed, CGT is a "right choice" codon, yet it is more rarely used than the corresponding "wrong choice" codon (CGC) in thermophiles and hyperthermophiles.

3.5. Gene expressivity effect

Consistent with results reported by Lynn et al. (2002) we found that the discriminating power of the second factor for synonymous codon usage was increased when only ribosomal proteins were taken into account (data not shown), for instance the *y*-scale upper limit in Fig. 3 had to be doubled to allow for the plot of AGG codon frequencies in ribosomal proteins.

4. Discussion

Consistent with previous studies on synonymous codon usage in prokaryotes, our results show that the second factor, orthogonal to the G+C content, discriminates a small subset of species (on the top of Fig. 1). Because of the high proportion of thermophilic species in this group, this factor has been consensually interpreted in the past as a temperature effect. We consider this interpretation as an overstatement with respect to current available evidence, or at least a point that should be discussed.

The supposedly predictive variable, T_{opt}, proved to be highly inconsistent between sources in some cases (Table 1). The high proportion of conflicting measurements within spore-forming bacteria (*Bacillus*, *Lactobacillus*) and antibiotic-producing bacteria (*Streptomyces*) suggests that there could be an anthropocentric bias here: for example the optimal growth temperature is not the same as the optimal temperature for antibiotic production. The typical between-sources differences in the ±5 °C range suggest that T_{opt} should be used to build an ordered categorical variable instead of a pure continuous one, by defining broad classes of optimal growth temperature range.

The large number of available species should not hide the fact that classes of interest (psychrophiles, thermophiles and hyperthermophiles) are still poorly documented. What is especially missing is thermophilic species with a high G+C content. Because two discriminating codons (ATA and AGA) are decreasing with the G+C content, it may turn out that the discriminating power of the second factor would not be as good when more such species are incorporated in the analysis.

When analyzing the relationship between codon usage and optimal growth temperature, there are (at least) two possibly confounding factors that need to be eliminated. The first factor is the influence of G+C content on codon usage; the correspondence analysis method applied here allows an efficient removal of this confounding effect, because the link with growth temperature is observed on the second principal factor of the analysis, which is orthogonal to the one linked to G+C content. The second possible confounding factor is represented by the phylogenetic relationships among the analyzed species. For example, since a larger proportion of hyperthermophiles are Archaea rather than Bacteria, and a larger proportion of mesophiles are Bacteria rather than Archaea, one may wonder if the apparent relationship between synonymous codon usage and growth temperature is not an artefact caused by the distribution of the species between the two taxonomic domains. Our results suggest that this is probably not the case: most mesophilic Archaea included in our dataset are clustered with the mesophilic Bacteria, and not with the thermophilic or hyperthermophilic Archaea, and most of the thermophilic Bacteria are placed closer to the thermophilic Archaea than to the mesophilic Bacteria. However, this evidence is based on a very small number of cases, so that this should be understood as a preliminary result, waiting for more mesophilic Archaea and more thermophilic Bacteria to be documented.

In addition to the already noticed outlier *E. acidaminophilum*, our results show that the single Archaeal psychrophile included in our dataset (*C. symbiosum*) clusters with the thermophiles and hyperthermophiles on the first factorial map for synonymous codon usage. The presence of these two outliers introduces a note of caution with respect to the interpretation for T_{opt} as the unique determinant of the second factor. Since all the bacterial psychrophiles analyzed are placed on the bottom of the factorial map, with the mesophiles, we wanted to check whether the unusual position of *C. symbiosum* was specific of Archaeal psychrophiles. However, the DDBJ/EMBL/GenBank databases did not include at the time any other Archaeal psychrophilic species with enough coding sequence data. We have used instead the draft sequence data for *Methanogenium frigidum* (T_{opt}=15 °C) and *Methanococcoides burtonii* (T_{opt}=20 °C) described in Saunders et al. (2003) and available on the authors' website (http://psychro.bioinformatics.unsw.edu.au). When included in the analysis, these species were clustered with the mesophiles and close to the other methanogens on the factorial map (not shown). The hypothesis that the unusual position of *C. symbiosum* is common to all Archaeal psychrophiles can therefore be rejected. In any case, since there are currently only 42 protein-coding sequences available for this organism, this situation needs to be considered with caution.

The interesting hypothesis that codon–anticodon pairs with an intermediate interaction strength are favored in (hyper)thermophilic species was previously tested with a dataset consisting of 16 species (Basak et al., 2005b). Our results do not allow us to reject clearly this hypothesis (*cf* Fig. 4) but show that its contribution to the second factor is marginal in terms of explained variability (*cf* Fig. 2). As noted by Lynn et al. (2002) the second factor is pervasive in the sense that it is still present if the most discriminating codons are removed from the analysis. Phenomena of different nature and intensities are therefore contributing to the definition of the second factor.

Consistent with the results from Lynn et al. (2002) we found that the discriminating power of the second factor was enhanced when restricting the analysis to ribosomal proteins, suggesting a relationship with gene expressivity. However, we do not claim this as an evidence for selection on synonymous codon usage in (hyper)thermophilic prokaryotes, because the same result would be obtained with a transcription-induced mutational bias. Our results do not allow to discriminate between a selective or a mutational pressure.

As stated earlier, the AGG, AGA and CGT codons for Arginine are among the most discriminating between (hyper) thermophilic and non-thermophilic lifestyles. It has been reported earlier that the AGG codon is rare in some mesophilic bacterial species and that its presence has a strong negative effect on translation efficiency and speed; this codon is translated slowly in *Escherichia coli* (Robinson et al., 1984; Bonekamp and Jensen, 1988; Chen and Inouye, 1990) and it may be a cause for ribosomal frameshift (Spanjaard and van Duin, 1988). However, it has also been shown that AGA and AGG are the major Arginine codons in several eukaryotic (mesophilic) parasites such as *Plasmodium falciparum*, *Babesia bovis*, *Entamoeba hystolica* or *Theileria parva* (Sayers et al., 1995). Interestingly, these two codons have been reassigned to Stop or Serine in many mitochondrial genomes (Andersson and Kurland, 1991), another indication of their very particular nature.

The analysis of the relationship between AGG frequency and G+C content has revealed that there are two distinct groups of species: the first one shows almost no variation of AGG frequency with G+C content and is composed mainly of mesophilic species, with the addition of the 5 Bacterial psychrophiles and of a few thermophilic outliers. The second group of species includes mostly thermophiles and hyperthermophiles, and the exceptional psychrophile *C. symbiosum*. In this group the frequency of AGG is relatively high, and it seems to increase with the G+C content, as expected under the hypothesis of neutrality, since AGG is a rather G+C-rich codon (Knight et al., 2001) (see Fig. 3). A likely interpretation of the low AGG frequencies encountered in the first group of species is that this codon is under negative selection, due to its effects on the efficiency and speed of translation. As far as the second group is concerned, the apparent increase in frequency with G+C content suggests that the usage of this codon could be at least partially neutral. However, the relatively high AGG frequencies prevent us from excluding the hypothesis of a selective pressure in favor of this codon in the second group of species. Since AGG is an pure-purinic codon, the latter hypothesis may be linked to the observed purine enrichment in thermophilic and hyperthermophilic species (Lao and Forsdyke, 2000; Lobry and Chessel, 2003; Lambros et al., 2003; Paz et al., 2004), but it should be noted that the other pure-purinic codons do not show the same trend as AGG.

The subsequent question is what is the biological significance of this separation in two distinct trends for AGG usage. A very appealing hypothesis is of course the existence of a link with growth temperature, since the first group is composed mostly of species living at relatively low temperatures, and the second one consists mostly of thermophilic and hyperthermophilic species. However, the presence of the psychrophilic *C. symbiosum* in the second group is a very strong counter-argument.

Lastly, we would like to stress that even if the second factor of the correspondence analysis is linked to temperature, we are perhaps interpreting this in the wrong direction: if the available data were not biased towards mesophilic species, wouldn't we interpret this factor as evidence for mutation or selection bias at *low* temperature? What would be the point of view of a thermophilic researcher? We are all assuming implicitly that high temperatures are extreme environmental conditions, but this is not necessarily relevant from an evolutionary perspective if the genetic code was fixed in a hot environment. The behavior of the most discriminating codon AGG with respect to G+C content is especially puzzling: its increase in frequency in thermophilic species is what regular G+C-rich codons are supposed to do, its systematic avoidance in mesophilic species is not the regular pattern. The anomaly is in the mesophilic group, not in thermophilic group, as if it was the result of a secondary adaptation. Again, more thermophilic G+C-rich species are needed to check if this interesting trend is still supported, but it's worth mentioning that from an anticodon point of view, AGG is already a noticeable exception (Rocha, 2004).

Acknowledgement

The authors would like to thank Dr. Manolo Gouy for critically reading the manuscript and for his valuable suggestions.

References

Andersson, S., Kurland, C., 1991. An extreme codon preference strategy: reassignment. Mol. Biol. Evol. 8, 530–544.

Basak, S., Mandal, S., Ghosh, T., 2005a. Correlations between genomic GC levels and optimal growth temperatures: some comments. Biochem. Biophys. Res. Commun. 327, 969–970.

Basak, S., Mandal, S., Ghosh, T., 2005b. On the origin of genomic adaptation at high temperature for prokaryotic organisms. Biochem. Biophys. Res. Commun. 330, 629–632.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., Wheeler, D.L., 2005. GenBank. Nucleic Acids Res. 33, 34–38.

Bonekamp, F., Jensen, K., 1988. The AGG codon is translated slowly in *E. coli* even at very low expression levels. Nucleic Acids Res. 16, 3013–3024.

Brock, T., 1967. Life at high temperatures. Evolutionary, ecological, and biochemical significance of organisms living in hot springs is discussed. Science 158, 1012–2269.

Bulmer, M., 1991. The selection-mutation-drift theory of synonymous codon usage. Genetics 129, 897–907.

Carbone, A., Kepes, F., Zinovyev, A., 2005. Codon bias signatures, organization of microorganisms in codon space, and lifestyle. Mol. Biol. Evol. 22, 547–561.

Charif, D., Lobry, J., in press. SeqinR 1.0-2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis. In: U. Bastolla, M. Porto, 22 H.R., Vendruscolo, M. (Eds.), Structural Approaches to Sequence Evolution: Molecules, Networks, Populations. Vol. NA of Biological and Medical Physics, Biomedical Engineering. Springer Verlag, New York, p. NA.

Chen, G., Inouye, M., 1990. Suppression of the negative effect of minor arginine codons on gene expression: preferential usage of minor codons within the first 25 codons of the *Escherichia coli* genes. Nucleic Acids Res. 18, 1465–1473.

Chen, S.L., Lee, W., Hottes, A.K., Shapiro, L., McAdams, H.H., 2004. Codon usage between genomes is constrained by genome-wide mutational processes. Proc. Natl. Acad. Sci. U. S. A. 101, 3480–3485.

Dumontier, M., Michalickova, K., Hogue, C.W.V., 2002. Species-specific protein sequence and fold optimizations. BMC Bioinformatics 3, 39.

Farias, S., Bonato, M., 2003. Preferred amino-acids and thermostability. Genet. Mol. Res. 2, 383–393.

Galtier, N., Lobry, J., 1997. Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. J. Mol. Evol. 44, 632–635.

Gouy, M., Gautier, C., 1982. Codon usage in bacteria: correlation with gene expressivity. Nucleic Acids Res. 10, 7055–7074.

Gouy, M., Gautier, C., Attimonelli, M., Lanave, C., di Paola, G., 1985. ACNUC — a portable retrieval system for nucleic acid sequence databases: logical and physical designs and usage. Comput. Appl. Biosci. 1, 167–172.

Grantham, R., Gautier, C., Gouy, M., 1980. Codon frequencies in 119 individual genes confirm consistent choices of degenerate base according to genome type. Nucleic Acids Res. 8, 1892–1912.

Grosjean, H., Fiers, W., 1982. Preferential codon usage in prokaryotic genes: the optimal codon–anticodon interaction energy and the selective codon usage in efficiently expressed genes. Gene 18, 199–209.

Hickey, D., Singer, G., 2004. Genomic and proteomic adaptations to growth at high temperature. Genome Biol. 5, 117.1–117.7.

Hurst, L., Merchant, A., 2001. High guanine-cytosine content is not an adaptation to high temperature: a comparative analysis among prokaryotes. Proc. R. Soc. Lond., B Biol. sci. 268, 493–497.

Kanz, C., et al., 2005. The EMBL nucleotide sequence database. Nucleic Acids Res. 33, 29–33.

Kashefi, K., Lovley, D., 2003. Extending the upper temperature limit for life. Science 301, 934.

Knight, R., Freeland, S., Landweber, R., 2001. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC-composition within and across genomes. Genome Biol. 2, 1–13.

Kreil, D.P., Ouzounis, C.A., 2001. Identification of thermophilic species by the amino acid compositions deduced from their genomes. Nucleic Acids Res. 29, 1608–1615.

Kumar, S., Nussinov, R., 2001. Fluctuations in ion pairs and their stabilities in proteins. Proteins 43, 433–454.

Lambros, R., Mortimer, J., Forsdyke, D.R., 2003. Optimum growth temperature and the base composition of open reading frames in prokaryotes. Extremophiles 7, 443–450.

Lao, P., Forsdyke, D.R., 2000. Thermophilic bacteria strictly obey Szybalski's transcription direction rule and politely purine-load RNAs with both adenine and guanine. Genome Res. 10, 228–236.

Lobry, J.R., 1997. Influence of genomic G+C content on average amino-acid composition of proteins from 59 Bacterial species. Gene 205, 309–316.

Lobry, J., Chessel, D., 2003. Internal correspondence analysis of codon and amino-acid usage in thermophilic bacteria. J. Appl. Genet. 44, 235–261.

Lynn, D.J., Singer, G.A.C., Hickey, D.A., 2002. Synonymous codon usage is subject to selection in thermophilic bacteria. Nucleic Acids Res. 30, 4272–4277.

Marashi, S., Ghalanbor, S., 2004. Correlations between genomic GC levels and optimal growth temperatures are not robust. Biochem. Biophys. Res. Commun. 325, 381–383.

Methé, B., et al., 2005. The psychrophilic life style as revealed by the genome sequence of *Colwellia psychrerythrea* 34H through genomic and proteomic analyses. Proc. Natl. Acad. Sci. U. S. A. 102, 10913–10918.

Musto, H., Naya, H., Zavala, A., Romero, H., Alvarez-Valin, F., Bernardi, G., 2004. Correlations between genomic GC levels and optimal growth temperatures in prokaryotes. FEBS Lett. 573, 73–77.

Musto, H., Naya, H., Zavala, A., Romero, H., Alvarez-Valin, F., Bernardi, G., 2005. The correlation between genomic G+C and optimal growth temperature of prokaryotes is robust: a reply to Marashi and Ghalanbor. Biochem. Biophys. Res. Commun. 330, 357–360.

Ochman, H., 2002. Distinguishing the ORFs from the ELFs: short bacterial genes and the annotation of genomes. Trends Genet. 18, 335–337.

Oliver, J.L., Marin, A., 1996. A relationship between GC content and coding-sequence length. J. Mol. Evol. 43, 216–223.

Paz, A., Mester, D., Baca, I., Nevo, E.A.K., 2004. Adaptive role of increased frequency of polypurine tracts in mRNA sequences of thermophilic prokaryotes. Proc. Natl. Acad. Sci. U. S. A. 101, 2951–2956.

R Development Core Team, 2005. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria3-900051-07-0. http://www.R-project.org.

Robinson, M., et al., 1984. Codon usage can affect efficiency of translation of genes in *Escherichia coli*. Nucleic Acids Res. 12, 6663–6671.

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

Saunders, N., et al., 2003. Mechanisms of thermal adaptation revealed from the genomes of the Antarctic Archaea *Methanogenium frigidum* and *Methanococcoides burtonii*. Genome Res. 13, 1580–1588.

Sayers, J., Price, H., Fallon, P., Doenhoff, M., 1995. AGA/AGG codon usage in parasites: implications for gene expression in *Escherichia coli*. Parasitol. Today 11 (9), 345–346.

Singer, G.A.C., Hickey, D.A., 2003. Thermophilic prokaryotes have characteristic patterns of codon usage, amino acid composition and nucleotide content. Gene 317, 39–47.

Spanjaard, R., van Duin, J., 1988. Translation of the sequence AGA-AGG yields 50% ribosomal frameshift. Proc. Natl. Acad. Sci. U. S. A. 85, 7967–7971.

Tateno, Y., Saitou, N., Okubo, K., Sugawara, H., Gojobori, T., 2005. DDBJ in collaboration with mass-sequencing teams on annotation. Nucleic Acids Res. 33, 25–28.

Tekaia, F., Yeramian, E., Dujon, B., 2002. Amino acid composition of genomes, lifestyles of organisms, and evolutionary trends: a global picture with correspondence analysis. Gene 297, 51–60.

Thioulouse, J., Chessel, D., Dolédec, S., Olivier, J., 1997. ADE-4: a multivariate analysis and graphical display software. Stat. Comput. 7, 75–83.

Thompson, M.J., Eisenberg, D., 1999. Transproteomic evidence of a loop-deletion mechanism for enhancing protein thermostability. J. Mol. Biol. 290, 595–604.

Vetriani, C., Maeder, D., 2005. Protein thermostability above 100°C: a key role for ionic interactions. Proc. Natl. Acad. Sci. U. S. A. 95, 12300–12305.

Vieille, C., Zeikus, G., 2001. Hyperthermophilic enzymes: sources, uses and molecular mechanisms for thermostability. Microbiol. Mol. Biol. Rev. 65, 1–43.

Zavala, A., Naya, H., Romero, H., Sabbia, V., Piovani, R., Musto, H., 2005. Genomic GC content prediction in prokaryotes from a sample of genes. Gene 357, 137–143.