Internal correspondence analysis of codon and amino-acid usage in thermophilic bacteria

Jean R. LOBRY, Daniel CHESSEL

Laboratoire de Biométrie, Biologie Evolutive, Centre National de la Recherche Scientifique -UMR-5558, Université Claude Bernard - LYON I, Villeurbanne Cedex, France

Abstract. Starting from two datasets of codon usage in coding sequences from mesophilic and thermophilic bacteria, we used internal correspondence analysis to study the variability of codon usage within and between species, and within and between amino acids. The first dataset included 18,958,458 codons from 58,482 coding sequences from completely sequenced genomes of 25 species, along with 6,793,581 dinucleotides from 21,876 intergenic spaces. The second dataset, with partially sequenced genomes, included 97,095,873 codons from 293 bacterial species. Results were consistent between the two datasets. The trend for the amino-acid composition of thermophilic proteins was found to be under the control of a pressure at the nucleic acid level, not a selection at the protein level. This effect was not present in intergenic spaces, ruling out a pressure at the DNA level. The pattern at the mRNA level was more complex than a simple purine enrichment of the sense strand of coding sequences. Outliers in the partial genome dataset introduced a note of caution about the interpretation of temperature as the direct determinant of the trend observed in thermophiles. The surprising lack of selection on the amino-acid content of thermophilic proteins suggests that the amino-acid repertoire was set up in a hot environment.

Key words: amino-acid usage, codon usage, correspondence analysis, thermophiles.
**Introduction**

**Temperature and bacterial growth**

Microbial growth is characterized during the exponential phase of growth (BUCHANAN 1918) by the specific growth rate, $\mu$ of the biomass $x$:

$$\frac{dx}{dt} = \mu x$$

The specific growth rate is a function of many environmental factors (e.g. pH, water activity, salinity, nutrient concentration). Here, we are interested in the effect of temperature. A typical temperature response is illustrated below with data for *Escherichia coli* (BARBER 1908).

![Influence of temperature on E. coli growth](image-url)  

The three cardinal temperatures are: $T_{\text{min}}$, the temperature below which growth is no longer observed, $T_{\text{opt}}$, the temperature at which the maximum specific growth rate reaches its maximal value, $T_{\text{max}}$, the temperature above which no growth occurs. Between-species comparisons have shown that these parameters are highly correlated (ROSSO et al. 1993) so that $T_{\text{opt}}$ alone is a good summary of the temperature effect on one species. The precision and accuracy of $T_{\text{opt}}$ estimates are typically ±1°C, at least in mesophilic species. Observed values are: $T_{\text{opt}} \approx 20^\circ\text{C}$ in psychrophilic, $T_{\text{opt}} \approx 40^\circ\text{C}$ in mesophilic, $T_{\text{opt}} \approx 60^\circ\text{C}$ in thermophilic, and $T_{\text{opt}} \approx 80^\circ\text{C}$ in hyperthermophilic bacteria (this discretization of a continuous variable being admittedly somewhat arbitrary). The currently known organism with the highest $T_{\text{max}}$ of 113°C is *Pyrolobus fumarii* (BLÖCHL et al. 1997). The preliminary report on bacteria growing at 250°C (BAROSS, DEMING 1983) is most likely an artifact (TRENT et al. 1984, WHITE 1984, BERNHARDT et al. 1984).
The optimal growth temperature is a selectable phenotypic life trait (BENNETT et al. 1992), corresponding to the temperature a bacterial population can afford in its ecological niche, but note that psychrotrophic bacteria have been isolated from constantly, on an evolutionary time scale, warm environments (ASTWOOD, WAIS 1998), suggesting a complex and pleiotropic nature of the underlying genotypic background for temperature adaptation: there is no one-to-one map between one gene and $T_{\text{opt}}$, and the genes involved are also involved in the control of other phenotypic traits. In this context, it is sensible to look for genome-wide adaptations for temperature, for instance one may expect chromosomes from thermophilic species to be enriched in G•C pairs to increase their stability, but this is not the case (GALTIER, LOBRY 1997). This is not surprising since the major problem faced by covalently closed plasmids at high temperatures is not thermodenaturation, but thermodegradation (MARGUET, FORTERRE 1994).

**Genetic codes and the genome hypothesis**

Let $\mathbf{C}$ be the set of the 64 possible codons,

$$\mathbf{C} = \{ \text{AAA, AAC, AAG, AAT, ..., TTT} \},$$

and $\mathbf{A}$ the set of the empty set plus the 20 possible amino acids in proteins,

$$\mathbf{A} = \{ \emptyset, \text{Ala, Arg, ..., Val} \},$$

where the empty set represents stop codons and unassigned codons. A genetic code is a subjective function from $\mathbf{C}$ onto $\mathbf{A}$: every element of $\mathbf{C}$ maps to one element in $\mathbf{A}$, and every element of $\mathbf{A}$ is mapped to by some element of $\mathbf{C}$, as in the example below, corresponding to the so-called universal genetic code.
Codons for the same amino acid are termed synonymous. The first evidence that the genetic code is the same in all organisms, with few secondary and minor alterations (reviewed by Knight et al. 2001a), was the correlation between the genomic G+C content and the average amino-acid composition of proteins (Sueoka 1961): G+C-rich genomes are enriched in amino acids encoded by G+C-rich codons. This correlation would of course vanish if genetic codes differed between species.

While the genetic code is almost universal, there is some freedom left, thanks to its subjective (degenerate) nature, and correspondence analysis of the first available coding sequences showed that different species use different dialects; this is the genome hypothesis (GrantHam et al. 1980). In bacteria, the most important factor of between-species variability is the genomic G+C content (Lee et al. 1956, Belozersky, SpirIn 1958), in contrast to a low variability of G+C content within species (Sueoka et al. 1959, Rolfe, Meselson 1959). The wide between-species variation and narrow within-genome heterogeneity of the DNA G+C content was interpreted as the result of bi-directional mutation rates between A•T and G•C pairs (Sueoka 1962). While selection for translation optimization was evidenced early as a major factor for the within-species variability of codon usage in bacteria (Ikemura 1981, Gouy, Gautier 1982, Kanaya et al. 1999, Gautier 2000), only recently was a link found between bacterial genomic G+C content and an environmental factor (Naya et al. 2002) raising the interesting possibility of a non-zero impact of the genomic G+C content on cell fitness in bacteria.

Amino-acid and codon usage in thermophilic bacteria

Despite a rich literature (Russell, Taylor 1995, Kumar, NuSSInov 2001) concerning the search for thermal adaptations in proteins from thermophilic species, few authors have tried to cope with the problem of the genomic G+C content as a confusing factor. The dramatic impact of the genomic G+C content on protein composition is well documented in bacteria (Sueoka 1961, Lobry 1997, Gu et al. 1998, D’Onofrio et al. 1999, Wilquet, Van De Casteele 1999, Singer, Hickey 2000, Knight et al. 2001b, Radomski, Slonimski 2001). This is not so easy, and using empirical approaches, such as grouping Lys + Arg to overcome the genomic G+C content effect (Cambillau, Claverie 2000), is not satisfactory since their observed individual dependence on G+C content does not cancel out (Lobry 1997, Knight et al. 2001b).

One approach is to look for asymmetrical replacement patterns in orthologous proteins from organisms living at different temperatures and sharing a similar genomic G+C content. This was first done independently by two research teams (Haney et al. 1999, McDonalD et al. 1999) within the Methanococcus genus thanks to the availability of the complete genome of the thermophilic MethanococcuS jannaschii. The general pattern was an increase in Ile and charged...
amino acids (Glu, Arg, Lys) and a decrease in Met and polar uncharged amino acids (Asn, Gln, Ser, Thr) with increasing thermophily. Although it was tempting to generalize and interpret these results as a universal adaptation to high temperatures, because an increase in charged residues leads to a stronger stabilizing electrostatic effect in proteins (PERUTZ 1978, KUMAR, NUSSINOV 2001 and references therein), a comparison of orthologous proteins from *Deinococcus radiodurans* with *Thermus thermophilus* showed (MCDONALD 2001) that the situation is not so simple with a different behavior for some amino acids (Asn, Asp, Cys, Ile). Clearly, as concluded by MCDONALD (2001), selection due to environmental factors other than temperature is not unlikely, and it will be necessary to compare a large number of mesophile/thermophile pairs to detect universal asymmetrical replacement patterns. For instance, KENNEDY et al. (2001) showed that proteins from the mesophilic halophilic *Halobacterium* species NRC-1 are enriched in negatively charged amino acids, probably to enhance solubility and maintain function at high salinity (LANYI 1978, EISENBERG 1995), and the same trend was reported for the hyperthermophilic halophilic *Methanopyrus kandleri* (SLESAREV et al. 2002).

Another approach is the comparison of the global amino-acid composition, as deduced from genome sequences. Using principal component analysis (PCA) with uniform column weighting of global amino-acid composition from 8 thermophilic and 17 mesophilic bacteria, KREIL and OUZOUNIS (2001) found a clear separation of thermophiles and mesophiles along the second factor (20% of total variability), the first factor being the genomic G+C content (45% of total variability). Since in PCA successive factors are under orthogonality constraints, this clearly demonstrated that once the major G+C content effect is removed, there is a universal trend for the composition of thermophilic proteins with an enrichment in Glu and Val and a depletion in Gln. Again, it would be tempting to interpret these results as evidence for selection on the amino-acid content in thermophilic proteins. However, LAO and FORSDYKE (2000) showed that some synonymous codons ending with a purine are favored in thermophilic species. Moreover, correspondence analysis of relative synonymous codon usage (CA/RSCU, SHARP et al. 1986) in 83,985 coding sequences has shown (LYNN et al. 2002, Hector MUSTO personal communication) that coding sequences from thermophiles are separated along the second factor (10.5% of total variability), the first factor being the genomic G+C content (28.0% of total variability). Therefore, as pointed by LAO and FORSDYKE (2000), some features of the amino-acid composition of thermophiles might reflect mutational or selective pressure at the nucleic acid level rather than direct constraints on protein structure and function, in a way similar to the influence of G+C directional mutation pressure on the average composition of proteins.

In summary, thanks to the availability of many complete genomes from thermophilic and hyperthermophilic bacteria, a new exciting correlation between a structure at the molecular level and a life-trait has emerged. Whether the struc-
ture is coherent between the protein and nucleic acid level is unclear, because previous multivariate analyses, using different weighting schemes, are not directly comparable. Moreover, CA/RSCU is not without methodological pitfalls when analysing the within-amino-acid variability of rare amino acids (PERRIÈRE, THIOULOUSE 2002). Taking advantage of the additivity of the variability decomposition scheme associated with internal correspondence analysis (ICA), we have tried to clarify the nature of the correlation between the structure at the molecular level and thermophily.

Material and methods

Datasets

Complete genomes: data source, organization, selection and structure

Complete genomes in GenBank format (BENSON et al. 2002) were downloaded on 19 June 2002 from the NCBI repository site at the URL: ftp://ftp.ncbi.nih.gov/genomes/Bacteria. Sequences were structured under the ACNUC model (GOUY et al. 1984) to allow an easy manipulation of sequence features. Coding sequences (CDS) annotated as partial or containing less than 300 bp were discarded to remove dubious ELFns, an acronym for Evil Little F...ellows (OCHMAN 2002). Codon counts were done with simple home-made C programs (available upon request) interfaced with the ACNUC library. The table of observed frequencies had 58,482 rows, each corresponding to a CDS, and 64 columns corresponding to the 64 possible codons. The row-block structure was the species of origin: a CDS belongs to one and only one species. The column-block structure was the encoded amino acid: a codon codes for one and only one amino acid. To ensure a homogeneous column-block structure among all CDS, species with a non standard genetic code were discarded. The total number of observed codons was 18,958,458.

All available sequences of genomes of thermophilic and hyperthermophilic bacteria were used. There were 16 thermophilic species, with an over representation of archaea, including 13 species, as compared to only 3 eubacteria (*Thermoanaerobacter tengcongensis*, *Aquifex aeolicus*, *Thermotoga maritima*). There was a sampling bias towards low genomic G+C contents, with a mean value of 44.2% and a maximum of 61.2% for *Methanopyrus kandleri*. A pseudo-control set of mesophilic species to ensure comparison was made as follows. (i) We wanted to cover a wide range of genomic G+C contents with a minimum of species. We selected for this purpose *Rickettsia prowazekii* (G+C 29.1%), *Staphylococcus aureus* (32.8%), *Escherichia coli* (50.8%), *Pseudomonas aeruginosa* (66.6%), and *Streptomyces coelicolor* (72.1%). (ii) As marine biotopes of hyperthermophiles are characterized by high concentrations of salt (HUBER et al. 2000), we added two halophilic mesophilic bacteria (*Halobacterium sp.* and *Bacillus*)
halodurans). (iii) Thermophilic archaea are often deficient in mismatch repair (FITZ-GIBBON et al. 2002), yielding a mutator phenotype with a high length polymorphism at repeat-tracts of Gs (or Cs). We therefore added a mesophilic bacterium without MutL homolog and a similar mutator phenotype, Campylobacter jejuni. (iv) Archaea are also deficient in glutaminyl-tRNA synthetases, Gln-tRNA are probably amino-acylated with Glu, followed by a transamidation to yield Gln-tRNA (CURNOW et al. 1997). Since gram-positive bacteria are also deficient in this enzyme, and Staphylococcus aureus and Bacillus halodurans were already present in the dataset, no more species were added. (v) The mesophilic Deinococcus radiodurans, as a relative of the thermophilic Thermus thermophilus, was added to facilitate comparisons with previous works (MC DONALD 2001). (vi) No complete genome of psychrophilic bacteria was available.

Coding sequences are known to be enriched in purine clusters (i.e. runs of RpRpRpR...) in their sense strand (SZYBALSKI et al. 1966), and this trend was found to be exacerbated in thermophilic species, suggesting an adaptive function at the mRNA level (LAO, FORSYDYE 2000). However, spectacular mono-nucleotide runs, up to 16 bp, have been reported in the genome of the hyperthermophilic Pyrobaculum aerophilum in both CDS and tolerant regions, such as suspected pseudogenes and long intergenic regions, introducing a note of caution for an adaptive interpretation of purine clusters in thermophilic bacteria (FITZ-GIBBON et al. 2002). We computed for each CDS the number of purine clusters, and used the number of runs test (WALD, WOLFOWITZ 1940) to evaluate the significance of observed values. At a critical level of 5%, a normalized z-score of less than −1.96 means that the observed number of purine runs is significantly lower than expected under a random repartition of purines along the sequence; in other words, there is a significant excess of purine clusters.

Assuming correct annotations, complete genomes are very interesting to look for base frequency features in intergenic spacers. As these regions are not transcribed, translation- and transcription-induced effects are removed, allowing to detect factors at the DNA level. For example, a predominance of YpY and RpR dinucleotides over YpR and RpY dinucleotides has been reported, although preliminarily, in non-coding regions of thermophilic species (KAWASHIMA et al. 2000). Intergenic spacers were defined here as continuous stretches of DNA between two annotated genes with a minimum length of 100 bp. The strand used for computations was the published strand in GenBank. The analysed table had 21,878 lines corresponding to the intergenic spacers and 16 columns corresponding to the dinucleotide frequencies (TpT, TpC, TpA, TpG, CpT, ..., GpG) for a total of 6,793,581 observations. The row-block structure was the species of origin: an intergenic spacer belongs to one and only one species.

Partial genomes: data source, organization, selection and structure

GenBank (BENSON et al. 2002) release 130 (15-JUN-2002), including daily updates on the date of 13 July 2002 and structured under the ACNUC model (GOUY
et al. 1984), was scanned with simple home-made C programs (available upon request) interfaced with the ACNUC library for the bacterial (archaeal or eubacterial) genomes with a standard genetic code and with a total of more than 50 kb of CDS available, each complete and with more than 300 bp in length. We found 293 bacterial genomes that satisfy this criterion. Since the analysis of the complete genome dataset showed (cf. result section) that the within-species variability was negligible for our present concern, comparisons were made directly on a between-species basis by summing all data within each genome. The table of observed frequencies had 293 rows, each corresponding to a bacterial species, and 64 columns corresponding to the 64 possible codons. The entry \((i,j)\) in this table is the number of \(j\)th codons in \(i\)th species. The column-block structure was, as previously, the encoded amino acid, and there was no more row-block structure because of the within-species summing. The total number of codons was 97,095,873, that is only 5 times the number available for the complete genome dataset, while the number of species was 12 times higher. The trade-off to pay for the increase in taxonomic coverage is the possibility of within-genome sampling bias. We found a clear example of sampling bias for \textit{Candidatus Carsonella ruddii} (THAO et al. 2000), out of 75 available CDS, 31 (41\%) were coding for the ATP synthase gamma subunit, yielding a biased base composition.

**Internal correspondence analysis**

Correspondence analysis (CA) is a multivariate method and, as such, its aim is to summarize data structures in high-dimension space by projection onto low-dimension subspaces, while losing as little information as possible. The principal factors are therefore along the directions of maximum variability in the dataset. CA is based on the principe d’équivalence distributionnelle (BENZÉCRI 1973) that can be illustrated as follows in our case. Let us assume that there is a completely uniform usage of codons within an amino acid. If all the corresponding columns in the original table are merged by summing them and the CA is repeated with the merged table, the result will be exactly the same: same total variability, same coordinates of CDS and codons on factorial maps. Therefore, if the within-amino-acid variability of synonymous codon usage was null for all amino acids, CA of the initial table would be strictly equivalent to CA analysis of the amino-acid table usage. In a symmetric way, let us suppose that codon usage within a species is without variability, that is all CDS are using codons with same relative frequencies. Then, all the corresponding rows can be merged by summing within species, and the repeated CA on the merged table will give exactly the same result. If the within-species variability was null, it would be strictly equivalent to run CA on the average species codon usage. This leads in a natural way to within-CA and between-CA (BENZÉCRI 1983), in which total variability is decomposed in its between-block component as we have just seen, and its within-block component by centering blocks at their average block value. In the present context, we want to compare the within-amino-acid and be-
tween-amino-acid variability, the within-species and between-species variability, so that it is necessary to use a common weighting scheme to make sensible comparisons. This is ensured by internal correspondence analysis (ICA), which is basically a double within-between-CA. ICA has already been unambiguously described (CAZES et al. 1988) in a vernacular background, but formulas are universal, so that the access to the primary literature should not be a problem (e.g. TAKAHE et al. 1991).

ICA and associated analyses were run through their implementations in the ADE4 library (THIOULOUSE et al. 1997), which is available as a free package in the R language for statistics (IHAKA, GENTLEMAN 1996). The ADE4 package is available in the “contributed packages” area of all the servers of the CRAN archives network. Please use the nearest mirror listed at the URL: http://cran.r-project.org/mirrors.html. ADE4 is a libre software (we do not use the term “free”, which has been proved to be extremely confusing because of its polysemic nature in English: free speech, no free beer) available under the terms of the Free Software Foundation’s GNU General Public License in source code form. The hardware was a Sun Fire-880 having two sparc processors with 2 Gio of RAM each. As a rule of thumb, we found it efficient to have 1 Mio of RAM available per cent lines of the analysed table, although this requirement could be reduced at the expense of periodic dumping on non-volatile memory.

Results

Complete genome dataset

Global analysis of codon usage

The eigenvalue graph (Figure 1) showed that there was a strong structure in the dataset with two major factors accounting for 49.9% of total variability, so that the first factorial map is a good summary of the dataset with about half of the variability extracted. The first factorial map for individuals (Figure 2) showed that CDS from the same species tend to be clustered together, which is an illustration of the genome hypothesis (GRANTHAM et al. 1980) stating that codon usage is different between species. The first axis (41.9%) was linked to the genomic G+C content, with GC-poor genomes on the left (e.g. Rickettsia prowazekii 29.1%), G+C-rich genomes on the right (e.g. Streptomyces coelicolor 72.1%), and G+C-intermediate genomes in the middle (e.g. Escherichia coli 50.8%). The second axis (8%) was linked to temperature with all thermophilic bacteria including eubacteria and archaea at the top (e.g. Aeropyrum pernix $T_{opt} \approx 95^\circ C$) and all mesophilic bacteria at the bottom (e.g. Escherichia coli $T_{opt} \approx 41^\circ C$). The two mesophilic and halophilic bacteria (Halobacterium sp. and Bacillus subtilis) were not close to the thermophilic bacteria, so that there was no evidence for halophily as a confounding factor. Campylobacter jejuni was not close
to thermophilic bacteria, so that there was no evidence for mismatch repair deficiency as a confounding factor. Mesophilic Gram-positive bacteria (e.g. *Staphylococcus aureus*) were not close to thermophilic bacteria, so that there was no evidence for glutaminyl-tRNA synthetase deficiency as a confounding factor.

Inspection of codon coordinates on the first factorial map (Figure 3) showed that AT-rich codons (e.g. TTA, AAT) were on the left of the first axis, while GC-rich codons (e.g. CGG, CGC) were on the right of the first axis, illustrating the correspondence between the genomic G+C content of a genome and its codon choice strategy. The second axis showed that thermophilic bacteria, as compared to mesophilic bacteria, tended to favor codons at the top of the map (e.g. AGG, ATA, AGA, AAG) and to avoid codons at the bottom (e.g. CAA, CGT).

Figure 1. Eigenvalue graphs for internal correspondence analysis and associated analyses of codon usage in 58,482 bacterial coding sequences from 25 complete genomes. The eigenvalue for a given factor is proportional to the variance in the table under analysis that is accounted for by that factor. Conceptually assigned factors were colored as follows: black for the G+C content effect, dark gray for temperature effect, and light gray for the sub-cellular location effect. The relative contribution of one factor within an analysis is also indicated in percentage for assigned factors. All the graphs are on the same scale (only the first 10 eigenvalues represented) to allow a direct visual comparison.
The genomic G+C content was found to be the major factor of variability at all levels (Figure 1) except for the between-amino-acid within-species variability. In this analysis the between-species variability in amino-acid usage in proteins was removed. The distribution of protein coordinates on the first factor was bi-modal (Figure 4), with the minor group (13.28%) corresponding to integral membrane proteins and the major group (86.72%) to cytoplasmic proteins.

Variability decomposition

The genomic G+C content was found to be the major factor of variability at all levels (Figure 1) except for the between-amino-acid within-species variability. In this analysis the between-species variability in amino-acid usage in proteins was removed. The distribution of protein coordinates on the first factor was bi-modal (Figure 4), with the minor group (13.28%) corresponding to integral membrane proteins and the major group (86.72%) to cytoplasmic proteins.
The maximum likelihood estimate of the proportion of integral membrane proteins (13.28%) was close to the 11% proportion found in a subset of *Escherichia coli* proteins (LOBRY, GAUTIER 1994). Integral membrane proteins were found to avoid charged amino acids (e.g. Glu, Lys, Asp) and to be enriched in hydrophobic ones (e.g. Phe, Trp), as expected. The sub-cellular location was therefore found to be a universal factor of within-species variability of amino-acid usage between proteins. As expected from a selective pressure at the protein level, this factor is no more present at the within-amino-acid level (Figure 1). Residual factors for the within-species synonymous codon usage variability were found to be: the G+C content and the strand location effect (PERRIÈRE et al. 1996, LOBRY
As expected from directional mutation pressure (Sueoka 1962) the within-species synonymous G+C content variability was found to be extremely low (ratio 1:11), as compared to its between-species component (Figure 1).

The results for the between-species between-amino-acid variability were found to be consistent with those previously published (Kreil, Ouzounis 2001), with an average protein composition between species dominated by the genomic G+C content effect (75.2%), as expected (cf. introduction), and a second factor linked to thermophily (11.4%). We noted, however, that this effect was masked when the within-species variability was restored (Figure 1), so that this thermophilic factor at the protein level was less important than the variability due to the selective pressure for the sub-cellular location of proteins. In other words, the average protein composition differed between thermophilic and mesophilic bacteria, but this difference was quantitatively less important than the difference between cytoplasmic and integral membrane proteins.

The results for the within-amino-acid variability (Figure 1) were found to be consistent with those previously published (Lynn et al. 2002), with a synonymous codon usage in thermophiles.
variability dominated (48.2%) by the genomic G+C content effect as the first factor and thermophily as the second factor (9.5%). Because a selective pressure at the protein level does not affect synonymous codon usage, the results for this analysis were in contradiction to the hypothesis of a selective pressure on the average protein content in thermophilic bacteria as the unique source of their peculiar global codon usage.

At the between-species level, the comparisons of the structures for synonymous and non-synonymous codon usage showed that the first two factors genomic G+C content and thermophily were highly correlated with similar response slopes (Figure 5). Since the first factor of variability is the G+C content, Figure 5 is very similar to Sueoka’s neutrality plots in which $P_{12}$, the corrected G+C content in first and second positions, is plotted against $P_3$, the corrected G+C content in third codon positions (SUEOKA 1988). The value of the slope, or degree of neutrality of about 31%, was found consistent with previous estimates in bacteria (SUEOKA 1988, LOBRY, SUEOKA 2002). The second factor response was found to be almost parallel to the first one (Figure 5). Therefore, these results were (i) in contradiction to the hypothesis that the average thermophilic proteome com-
position is dictated by direct constraints on protein structure and function, and (ii) not in contradiction to the hypothesis that some features of the amino-acid composition of thermophiles might reflect mutational or selective pressure at the nucleic acid level, in a way similar to the influence of G+C directional mutation pressure on the average composition of protein, as originally suggested by LAO and FORSDYKE (2000).

**Synonymous codon usage in thermophilic bacteria**

The second factor (12.6%) for the between-species variability for synonymous codon usage was linked to thermophily, after the effect of the genomic G+C content (71.6%). Observed trends in thermophiles were consistent with previously reported results (LYNN et al. 2002), with a strong preference for AGG, ATA, AGA and strong avoidance of CGT, CGA codons in thermophiles. The pattern of synonymous codon usage was more complex than a simple purine enrichment in thermophiles (Table 1). When in the third codon position there is only a synonymous choice between T and C (Asn, Asp, Cys, His, Phe, Tyr), the C-ending coding was systematically favored in thermophiles. When in the third codon position there is only a synonymous choice between A and G (Gln, Glu, Lys), the G-ending codon was systematically favored in thermophiles. This last trend was also present in stop codons with TAG favored over TAA in thermophiles. This regular pattern, however, was no more visible for amino acids encoded by more than two codons. Ile is encoded by ATT, ATC and ATA codons, so from extrapolation of the trend in 2-boxed amino acids, we would expect the C-ending to be favored, but codon ATA was found to be strongly favored in thermophiles. Thr is encoded by the four ACN codons, and no clear preference was found in thermophiles. Ala (GCN) and Val (GTN) are encoded by the four codons and the T-ending tended to be favored in thermophiles. Gly is encoded by the four GGN codons and the A-ending was favored in thermophiles. Among the 4-boxed amino acids, only Pro (CCN) would follow the 2-boxed trend with the C-ending codon favored in thermophiles. For the 6-boxed amino acids (Arg, Leu, Ser) no clear tendency was obvious, with a strong preference for AGA and AGG (Arg), a moderate preference for CTT, CTC and CTA (Leu), and a weak preference for TCC (Ser) in thermophiles. These results were in contradiction to a transcription-induced, context-independent, mutation bias increase due to temperature to explain the pattern of synonymous codon usage in thermophiles. There was no obvious dinucleotide (such as CpG) avoidance in thermophiles thanks to the flexibility of the third codon position (Table 1), so that a transcription-induced, short-range context-dependent, mutation bias increase was also unlikely.

**Purine cluster analysis**

There was a general clear tendency for enrichment in purine clusters with thermophily. The results were consistent with a previous report for the 9 species that were documented (KAWASHIMA et al. 2000). However, the majority of the CDS from 4 thermophilic species (viz Thermoplasma acidophilum,
Table 1. Codon coordinates on the first factorial map for the between-species synonymous codon usage in the complete genome dataset. The first axis is the genomic G+C content with GC-rich codons having positive values. The second factor is linked to thermostability, with positive values corresponding to codons favored in thermophilic species, irrespective of the G+C content effect.

<table>
<thead>
<tr>
<th>AA</th>
<th>Codon</th>
<th>Axis 1</th>
<th>Axis 2</th>
<th>AA</th>
<th>Codon</th>
<th>Axis 1</th>
<th>Axis 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ala</td>
<td>GCT</td>
<td>−0.819</td>
<td>0.076</td>
<td>CTC</td>
<td>0.449</td>
<td>0.256</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GCC</td>
<td>0.467</td>
<td>0.015</td>
<td>CTA</td>
<td>−0.643</td>
<td>0.305</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GCA</td>
<td>−0.757</td>
<td>0.053</td>
<td>CTG</td>
<td>0.738</td>
<td>−0.131</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GCG</td>
<td>0.369</td>
<td>−0.101</td>
<td>Lys</td>
<td>AAA</td>
<td>−0.288</td>
<td>−0.268</td>
</tr>
<tr>
<td>Arg</td>
<td>CGT</td>
<td>−0.119</td>
<td>−0.700</td>
<td>AAG</td>
<td>0.303</td>
<td>0.282</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CGC</td>
<td>0.739</td>
<td>−0.412</td>
<td>Met</td>
<td>ATG</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>CGA</td>
<td>−0.189</td>
<td>−0.423</td>
<td>Phe</td>
<td>TTT</td>
<td>−0.520</td>
<td>−0.198</td>
</tr>
<tr>
<td></td>
<td>CGG</td>
<td>0.776</td>
<td>−0.224</td>
<td>TTC</td>
<td>0.448</td>
<td>0.171</td>
<td></td>
</tr>
<tr>
<td></td>
<td>AGA</td>
<td>−0.931</td>
<td>0.313</td>
<td>Pro</td>
<td>CCT</td>
<td>−0.718</td>
<td>0.029</td>
</tr>
<tr>
<td></td>
<td>AGG</td>
<td>−0.573</td>
<td>0.836</td>
<td>CCC</td>
<td>0.389</td>
<td>0.196</td>
<td></td>
</tr>
<tr>
<td>Asn</td>
<td>AAT</td>
<td>−0.463</td>
<td>−0.213</td>
<td>Ser</td>
<td>TCT</td>
<td>−0.613</td>
<td>−0.039</td>
</tr>
<tr>
<td></td>
<td>AAC</td>
<td>0.413</td>
<td>0.190</td>
<td>TCC</td>
<td>0.461</td>
<td>0.156</td>
<td></td>
</tr>
<tr>
<td>Asp</td>
<td>GAT</td>
<td>−0.619</td>
<td>−0.087</td>
<td>TCA</td>
<td>−0.589</td>
<td>0.034</td>
<td></td>
</tr>
<tr>
<td></td>
<td>GAC</td>
<td>0.505</td>
<td>0.071</td>
<td>TCG</td>
<td>0.551</td>
<td>−0.049</td>
<td></td>
</tr>
<tr>
<td>Cys</td>
<td>TGT</td>
<td>−0.515</td>
<td>−0.129</td>
<td>Glu</td>
<td>GAA</td>
<td>−0.309</td>
<td>−0.187</td>
</tr>
<tr>
<td></td>
<td>TGC</td>
<td>0.342</td>
<td>0.086</td>
<td>GAG</td>
<td>0.288</td>
<td>0.174</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Gln</td>
<td>CAA</td>
<td>−0.671</td>
<td>−0.296</td>
<td>TAG</td>
<td>−0.016</td>
<td>0.234</td>
</tr>
<tr>
<td></td>
<td>CAG</td>
<td>0.400</td>
<td>0.176</td>
<td>AGT</td>
<td>−0.549</td>
<td>−0.202</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Glu</td>
<td>GAA</td>
<td>−0.309</td>
<td>−0.187</td>
<td>ACC</td>
<td>0.607</td>
<td>−0.031</td>
</tr>
<tr>
<td></td>
<td>GAG</td>
<td>0.288</td>
<td>0.174</td>
<td>ACA</td>
<td>−0.706</td>
<td>0.014</td>
<td></td>
</tr>
<tr>
<td>Gly</td>
<td>GGT</td>
<td>−0.442</td>
<td>−0.145</td>
<td>Thr</td>
<td>TGT</td>
<td>−0.628</td>
<td>0.152</td>
</tr>
<tr>
<td></td>
<td>GGC</td>
<td>0.587</td>
<td>−0.140</td>
<td>Trp</td>
<td>TGG</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>GGA</td>
<td>−0.609</td>
<td>0.246</td>
<td>Val</td>
<td>GTC</td>
<td>0.557</td>
<td>−0.060</td>
</tr>
<tr>
<td></td>
<td>GGG</td>
<td>0.050</td>
<td>0.140</td>
<td>Tyr</td>
<td>TAT</td>
<td>−0.449</td>
<td>−0.244</td>
</tr>
<tr>
<td>His</td>
<td>CAT</td>
<td>−0.554</td>
<td>−0.247</td>
<td>Val</td>
<td>GGT</td>
<td>−0.628</td>
<td>0.152</td>
</tr>
<tr>
<td></td>
<td>CAC</td>
<td>0.370</td>
<td>0.165</td>
<td>Trp</td>
<td>TGG</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Ile</td>
<td>ATT</td>
<td>−0.334</td>
<td>−0.314</td>
<td>Tyr</td>
<td>TAT</td>
<td>−0.449</td>
<td>−0.244</td>
</tr>
<tr>
<td></td>
<td>ATC</td>
<td>0.695</td>
<td>−0.146</td>
<td>Thr</td>
<td>TGT</td>
<td>−0.628</td>
<td>0.152</td>
</tr>
<tr>
<td></td>
<td>ATA</td>
<td>−0.332</td>
<td>0.430</td>
<td>Val</td>
<td>GTC</td>
<td>0.557</td>
<td>−0.060</td>
</tr>
<tr>
<td>Leu</td>
<td>TTA</td>
<td>−0.956</td>
<td>−0.363</td>
<td>Gln</td>
<td>GAA</td>
<td>−0.309</td>
<td>−0.187</td>
</tr>
<tr>
<td></td>
<td>TTG</td>
<td>−0.377</td>
<td>−0.079</td>
<td>GAG</td>
<td>0.288</td>
<td>0.174</td>
<td></td>
</tr>
<tr>
<td></td>
<td>CTT</td>
<td>−0.562</td>
<td>0.239</td>
<td>Glu</td>
<td>GAA</td>
<td>−0.309</td>
<td>−0.187</td>
</tr>
</tbody>
</table>
Thermoplasma volcanium, Pyrobaculum aerophilum, Methanopyrus kandleri) were not significantly enriched in purine clusters. Moreover, the pseudo-control species for the mutator phenotype, the mesophilic Campylobacter jejuni, was found with an enrichment in purine clusters as in thermophilic species. Therefore, the enrichment in purine clusters in most thermophilic species was not found at the origin of the specificity of codon usage in thermophiles.

**Intergenic spaces analysis**

There was only one factor, the G+C content effect, underlying the variability of dinucleotide usage in intergenic spaces. At the between-species level, the G+C content accounted for 93.9% of total variability and the residual factor (2.5%) was not found to discriminate thermophilic species. These results were in contradiction to the hypothesis of a selective or mutational pressure working directly at the DNA level at the origin of the specificity of codon usage in thermophiles because their specificity is found only in CDS.

![Figure 6. First factorial map for synonymous codon usage in 293 bacteria.](image)

Figure 6. First factorial map for synonymous codon usage in 293 bacteria. Filled points correspond to thermophilic species and open points to mesophilic species. The first factor is the G + C content, with GC-poor species on the left and GC-rich species on the right. The second factor is linked to thermophily, with almost all thermophilic species at the top. Some outliers of interest for the discussion are outlined.
Partial genome dataset

Consistency with the complete genome dataset

The results were extremely stable when the number of codons was multiplied by 5 and the number of species by 12, suggesting that there was no systematic within-genome sampling bias problem in the partial genome dataset. As previously, codon usage variability was dominated by the G+C content effect, followed by the thermophilic factor. Comparison of codon coordinate for the synonymous codon usage analysis on the second factor, showed that the trend in thermophiles was essentially the same ($r^2 = 0.82$), with an increase in the discriminating power of codons AGG, ATA and AGA.

Synonymous codon usage in thermophiles

In addition to the 16 thermophiles already present in the complete genome dataset, there were 7 new thermophiles (taking a cutoff value of 59°C to include *Thermoplasma acidophilum*) in the partial genome dataset: *Clostridium thermocellum* ($T_{opt} \approx 62°C$), *Geobacillus stearothermophilus* ($T_{opt} \approx 65°C$), *Sulfolobus acidocaldarius* ($T_{opt} \approx 70°C$), *Thermus aquaticus* ($T_{opt} \approx 72°C$), *Thermotoga neapolitana* ($T_{opt} \approx 80°C$), *Thermus thermophilus* ($T_{opt} \approx 80°C$), and *Thermococcus litoralis* ($T_{opt} \approx 88°C$). They were all found to cluster with the thermophilic group at the top of the second axis of synonymous codon usage, with the noticeable exception of *Geobacillus stearothermophilus* (Figure 6). No mesophilic bacteria were found to cluster with thermophiles on the second
axis, with the noticeable exception of the mesophilic, $T_{opt} \approx 33^\circ C$, *Eubacterium acidaminophilum* (a *Clostridium* species in fact, *cf.* Discussion). Species from the genera *Clostridium* and *Methanosarcina* tended to be grouped with thermophiles on the second factor (Figure 6).

**Comparison of synonymous and non-synonymous codon usage**

The species coordinates on the two factors of the synonymous and non-synonymous analyses were highly correlated (Figure 7). As previously for the complete genome dataset (Figure 5), the factors at the protein level were found to be under the control of factors at the nucleic acid level.

**Discussion**

**Absence of selection on the average amino-acid content of proteins at high temperatures**

Our results are clearly in contradiction to the hypothesis of a selective pressure on the amino-acid content at high temperatures to avoid the most thermolabile residues. The arguments are as follows. (i) This would not have any impact on synonymous codon usage as already observed (LYNN et al. 2002, Hector Musto personal communication, Figures 2 and 6). (ii) The intensity of the amino-acid response due the thermophily-related pressure at the nucleic acid level is exactly the same as in the case of the well-known effect of the influence of the G+C pressure. (iii) This residual response is quantitatively less important than the selection on amino-acid content for the sub-cellular location of proteins (membrane versus cytoplasm). (iv) The essential difference is that Glu is slightly more frequent than Gln in thermophilic proteins (KREIL, OUZOUNIS 2001). However, Asp and Asn are equally frequent, so a selective pressure to avoid deamidation at high temperatures is unlikely. (v) Mesophilic glutaminyl-tRNA-synthetase-deficient species do not cluster with thermophilic species, so that a selection to avoid amino acids whose cognate tRNA is obtained by transamidation is unlikely. (vi) Some mesophilic species cluster with thermophilic species and vice versa, so that the general trend observed in thermophilic species is not specific to thermophily.

This result is extremely surprising if one considers simultaneously the following two facts. On one hand, there is a lot of freedom left for the average amino-acid content of a protein without altering its function. From GC-poor bacteria to GC-rich bacteria, the slope of the regression line in Figures 4 and 6 means that one can change roughly a third of the residue of all proteins without being counter-selected. Another striking example is the effect of asymmetric mutation pressure on the amino-acid content of proteins in some bacteria. In *Borrelia burgdorferi* the effect is so important that one can guess, just from the amino-acid content of a protein, if the corresponding gene was on the leading strand or lagging.
strand for replication (LAFAY et al. 1999, MACKIEWICZ et al. 1999, ROCHA et al. 1999). On the other hand, selection is extremely efficient in bacteria because of their large population sizes. This is evidenced by the selection between synonymous codons for translation optimization: preferred codons match tRNAs with a high intracellular concentration, and this trend is exacerbated in highly expressed genes (AKASHI 2001 and references therein). In a bacteria-sized genome, the minute influence of a synonymous mutation, which can occur in something like $10^6$ possible positions in the genome, on cell fitness is, however, sufficient to be selected, at least in free-living bacteria. Therefore, even a minute difference between amino acids in terms of thermostability should have been selected in thermophilic proteins, and this is not the case, and that is why it is very surprising. A possible explanation for this is that at the time the amino-acid repertoire was set up, the environment was hot, and too thermolabile amino acids were discarded. This is not in contradiction to the hypothesis of a mesophilic last universal common ancestor to extant life forms (GALTIER et al. 1999), because a long time interval may have elapsed between the setting up of the amino-acid repertoire and the last universal common ancestor. If this interpretation is valid, then the current set of amino acids in proteins could be regarded as the result of a frozen accident ... at high temperatures.

Note that although it is easy to imagine positive results that would have shown a clear and understandable relationship between temperature, protein composition and stability, the observed negative result is difficult to interpret in terms of its evolutionary significance because the lack of proof is not the proof of lack.

Outliers

The separation of thermophilic species on the second axis suffered from two noticeable exceptions: the thermophilic *Geobacillus stearothermophilus* clustered with thermophilic species, and the mesophilic *Eubacterium acidaminophilum* clustered with thermophilic species.

With a threshold value of 59°C, the thermophilic status of *Geobacillus stearothermophilus* (formerly *Bacillus stearothermophilus*) is disputable. Out of more than 750 heterotrophic spore-forming strains isolated from hot compost, very few of them grow above 60°C, and only *Bacillus coagulans* and *Geobacillus stearothermophilus* were isolated at 65°C (STROM 1985a,b). There are currently 9 described species in the genus *Geobacillus*, and their permissible temperature range is wide, from 35°C to 78°C (SUNG et al. 2002). To our knowledge the best available dataset for the influence of temperature on the growth of *G. stearothermophilus* is in RATKOWSKY et al. (1983), with a well-designed experiment with 14 points covering the whole permissible range for two strains. Optimal growth temperature and parameter confidence limits ($\alpha = 0.05$) are 57.2°C [56.5-57.9] and 64.8°C [62.5-67.5] for the two strains.
Depending on the strain, we could therefore consider *G. stearothermophilus* as a mesophilic or a thermophilic bacterium.

In contrast to the borderline status of *G. stearothermophilus*, *Eubacterium acidaminophilum* is clearly a mesophilic bacterium with an optimal growth temperature between 30°C and 35°C (ZINDEL et al. 1988). From comparisons of 16S rRNA gene sequences, *E. acidaminophilum* was found (BAENA et al. 1999) to be closely related (similarity of 96%) to the mesophilic (*T*\textsubscript{opt} ≈ 30°C) *Clostridium litorale* (FENDRICH et al. 1990). The genus *Clostridium* is a very heterogeneous group of anaerobic, gram-positive, rod-shaped, endospore-forming bacteria, comprising organisms that vary considerably in genome size (2.5–6.5 Mb), G+C content (24–55%) and optimal growth temperature, with both mesophilic (e.g. *C. botulinum* *T*\textsubscript{opt} ≈ 30°C) and thermophilic (e.g. *C. thermocellum* *T*\textsubscript{opt} ≈ 62°C) species. Although the number of *Clostridium* species is small in our dataset (8, including *E. acidaminophilum*), it is worth noting that they tend to be grouped with thermophilic species on the second axis of synonymous codon usage (Figure 6).

*Methanosarcina* species are archaea with a wide range of optimal growth temperatures, from *T*\textsubscript{opt} ≈ 20°C in *M. baltica* to *T*\textsubscript{opt} ≈ 50°C in *M. thermophila*. The three species available in the dataset (*M. acetivorans*, *M. barkeri*, *M. mazei*) are clearly mesophilic (*T*\textsubscript{opt} ≈ 37°C) but tend to be grouped with thermophilic species (Figure 6).

The outliers introduced a note of caution about the interpretation of temperature as the direct determinant of the trend of codon usage observed in thermophiles: this trend is common to all thermophiles but is not specific to thermophiles, so that another undetermined environmental factor, shared by thermophiles and outliers, could be the actual underlying cause.

**Selection to increase mRNA stability**

The high sensitivity of RNAs to hydrolysis is well-known. Under physiological conditions, hydrolysis of RNA involves an S\textsubscript{N}2-like mechanism with a nucleophilic attack by the 2’-oxygen on the adjacent phosphorus center, to form a pentacoordinate intermediate, followed by the departure of the 5’-oxyanion leaving group, yielding two RNA fragments with 2’,3’-cyclic phosphate and 5’-hydroxyl termini, respectively. Spontaneous hydrolysis of unstructured RNAs by this intramolecular transesterification is expected to be ~100,000-fold faster in thermophilic organisms than in mesophilic organisms (LI, BREAKER 1999). Stabilization of rRNAs and tRNAs in thermophilic species is achieved by double-stranded duplex, which precludes an in-line attack of the 2’-oxygen (SOUKUP, BREAKER 1999), an increased G+C content of the stems (GALTIER, LOBRY 1997), and many post-transcriptionally modified nucleosides, such as the 2’-O-methylation. However, there are many others, for instance in *Methanococcales* out of 26 modified nucleosides, 3 were found only in the tRNAs from hyperthermophilic species (MCCLOSKEY et al. 2001). We are not
aware of any documented mechanisms for the stabilization of mRNAs in thermophiles, although this is achieved in some way in vivo with some mRNAs having half-lives in hours in the hyperthermophile *Sulfolobus solfataricus* (BINI et al. 2002).

Under physiological conditions, 5′-UpA-3′ and 5′-CpA-3′ phosphodiester bonds in RNAs have been observed as less stable than those flanked by other nucleosides (BIBILLO et al. 1999), although this instability is not intrinsic, but rather results from an enhanced reactivity only within some oligonucleotide sequences (KAUKINEN et al. 2002). If there was a selective pressure to increase mRNA stability, we would expect thermophilic bacteria to strongly avoid these bonds in coding sequences when there is a synonymous alternative. This is not the case. Ile is encoded by the 3 codons, but in thermophilic species the codon ATA is favored over its synonymous competitors ATT and ATC, although it is the only one to yield a 5′-UpA-3′ bond at the mRNA level. Leu is encoded by 6 codons, and the codon CTA is not avoided, as compared to its synonymous competitors avoiding a 5′-UpA-3′. Ala is encoded by the 4 GCN codons, and the codon GCA is not especially avoided in thermophilic bacteria, although it is the only one to yield a 5′-CpA-3′ bond. Therefore, we conclude that there is no clear evidence for a selective pressure to increase mRNA resistance to hydrolysis by avoiding the most scissile bonds. Either there is no need for this, or it is achieved in another way.

**Acknowledgments.** We thank the ADE4 core team (especially Jean THIOULOUSE and Anne-Béatrice DUFOUR) for providing access to the R-ADE4 library before official release. We also thank Maria Luisa CHIUSANO for sharing her bibliographical data, Guy PERRIÈRE and Donal HICKEY for sharing preprints (before formal acceptance) of their work, and Hector MUSTO for personal communication of his team observations and sharing preprints before formal acceptance. Some important improvements of the discussion are due to Robin KNIGHT. Important key papers were pointed out to us by Patrick FORTERRE. Helpful comments were given by two anonymous referees. We thank Sylwia UFNALSKA for editing the English of the manuscript.

**REFERENCES**


