Extremophiles (2007) 11:9–18
DOI 10.1007/s00792-006-0005-6

ORIGINAL PAPER

Feng-Hsu Lin · Donald R. Forsdyke

Prokaryotes that grow optimally in acid have purine-poor codons in long open reading frames

Received: 20 February 2006 / Accepted: 29 March 2006 / Published online: 7 September 2006

Abstract In nucleic acids the *N*-glycosyl bonds between purines and their ribose sugar moities are broken under acid conditions. If one strand of a duplex DNA segment were more vulnerable to mutation than the other, then the archaeon *Picrophilus torridus*, with an optimum growth pH near zero, could have adapted by decreasing the purine content of that strand. Yet, *P. torridus* has an optimum growth temperature near 60°C, and thermophiles prefer purine-rich codons. We found that, as in other thermophiles, high growth temperature correlates with the use of purine-rich codons. The extra purines are often in third, non-amino acid determining, codon positions. However, as in other acidophiles, as open reading frame lengths increase, there is increased use of purine-poor codons, particularly those without purines in second, amino acid-determining, codon positions. Thus, *P. torridus* can be seen as adapting (a) to temperature by increasing its purines in all open reading frames without greatly impacting protein amino acid compositions, and (b) to pH by decreasing purines in longer open reading frames, thereby potentially impacting protein amino acid compositions. It is proposed that longer open reading frames, being larger mutational targets, have become less vulnerable to depurination by virtue of pyrimidine for purine substitutions.

Keywords Acid environment · Base composition · Codons · Extremophiles · Optimum growth · Purine-loading

Abbreviations ORF: Open reading frame · CUTG: Codon usage tables from GenBank

Communicated by F. Robb

F.-H. Lin · D. R. Forsdyke (✉)
Department of Biochemistry, Queen's University, K7L3N6
Kingston, ON, Canada
Tel.: +1-613-5332980
Fax: +1-613-5332497
E-mail: forsdyke@post.queensu.ca
http://post.queensu.ca/~forsdyke/bioinfor.htm

Introduction

The archaeon *Picrophilus torridus* grows optimally at pH 0.7 (Fütterer et al. 2004; Ciaramella et al. 2005). However, under acid conditions purines are removed from DNA, generating apurinic acid (Tamm et al. 1952; Bernstein and Bernstein 1991). How do organisms with low pH growth optima preserve their DNA? Mechanisms to ensure near neutral cytoplasm may aid preservation in some acidophiles. Yet *P. torridus* is probably like *P. oshimae* in having an internal pH of 4.6 and being unviable in media of pH above 4.0 (Jack et al. 1998). Since in duplex DNA a pyrimidine in one strand must be matched by a purine in the other strand, vulnerability to depurination cannot be decreased by eliminating purines from DNA. However, if one strand were particularly vulnerable (Beletskii and Bhagwat 1996), or if transcription products were vulnerable, a decreased content of purines in a segment of one strand might be advantageous.

In most organisms transcripts are purine-loaded (i.e., their purine content exceeds 50%; Forsdyke and Mortimer 2000). This trend is especially prevalent in organisms with high optimum growth temperatures (Lao and Forsdyke 2000; Lambros et al. 2003; Lobry and Chessel 2003; Paz et al. 2004; Basak et al. 2004). In acidophiles vulnerability to depurination might be less if there were decreased use of purine-rich codons. This would be especially important in long open reading frames (ORFs); mutation tending to be a one-hit process, long ORFs that contain many purines constitute larger potential targets for mutation than short ORFs. But *P. torridus* grows optimally at 60°C. Can nucleic acid base composition be adjusted in *P. torridus* to adapt both to environmental pH and to temperature?

To detect possible adaptation of the base composition of a species to an environmental variable, it is usual to plot the base composition values of different species against values for the variable. For example, although a subject of some contention (Musto et al. 2005; Basak and Ghosh 2005), linear regression plots of species base composition against optimum growth temperature show a negative correlation with species GC% and a positive correlation with species AG%. We here apply this approach to examine adaptation to pH. We first demonstrate for *P. torridus* that ORF base compositions vary with ORF length. For a given base compositional parameter (AG%, GC%) there is a unique slope when that parameter is plotted against ORF length. We then economically summarize a large amount of data by expressing the corresponding slopes for various acidophiles as a function of their optimum growth pH. Our results confirm that increased purine-loading is an adaptation to temperature that affects all ORFs, whereas purine-unloading appears as an adaptation to acid growth conditions that preferentially affects long ORFs.

Materials and methods

Sequences

"Codon usage tables from GenBank" (CUTG) are derived from the available annotated protein-encoding sequences of a species and are automatically updated with each new GenBank release (Nakamura et al. 2000). It was assumed that any errors in ORF assignments would be of insufficient magnitude to affect our aggregate results. Base compositions at different codon positions were calculated from the CUTG tables (GenBank release143) using programs written in Perl (Mortimer and Forsdyke 2003).

Prokaryotes for which optimum growth pH values had been determined were identified from the literature (Schleper et al. 1995; Wiegel and Kevbrin 2003; Macario et al. 2004). For many of these, the genomic sequences had been designated complete. Two species for which there is extensive, but still incomplete sequence information (*Geobacillus stearothermus* and *Sulfolobus acidocaldarius*), were also included (Table 1). Approximate values for total genomic base compositions were calculated after summing the base compositions of the available ORFs (Zavala et al. 2005).

Optimum pH

Organisms identified as growing optimally at extremes of pH are currently scarce. The five with lowest pH values are also thermophiles, so it is necessary to distinguish adaptation to pH from adaptation to temperature. Since there are many organisms that grow optimally in neutral media, a small selection was made from this pH region, with a bias towards organisms that, like *P. torridus*, are thermophilic and are also archaea (Table 1). Values for optimum growth temperatures were found in a prokaryote growth temperature database (PGTdb; Huang et al. 2004), and in the records at the National Center for Biotechnological Information (Washington). Where a range of optimum values was given, the arithmetic center of the range was selected.

Arrowhead length plots

ORF lengths were calculated from CUTG tables. When base compositions of entire ORFs, or of individual codon positions in ORFs, are plotted against ORF lengths, the multiple data points are distributed as rightward-pointing arrowheads (Paz et al. 2004). Many features of the distributions are captured by first order linear regression analysis. Although smaller proteins dominate the statistics, and the few points near the tips of arrowheads that correspond to very long proteins usually depart a little from regression lines, higher order regressions offer little improvement. Further details of methods and terminology may be found in Rayment and Forsdyke (2005).

Results

ORF length plots for *P. torridus*

Most organisms have many small proteins and few very large ones. For an individual species, plots of the base compositions of each protein-encoding region (ORF) against the corresponding lengths (kilobases), show the multiple data points to be distributed as rightward-pointing arrowheads, with each point corresponding to an individual gene (Paz et al. 2004). Many features of the distributions can be captured by first order linear regression analysis (Rayment and Forsdyke 2005). The 1,535 genes of *P. torridus* were studied in this way.

In *P. torridus*, with increasing gene length the overall AG% values of ORFs decrease (Fig. 1a), and some 4% of the variation between points can be accounted for on this basis (adjusted $r^2 = 0.038$). All codon positions, but especially first and second positions, contribute to the decrease (Fig. 1c, e, g). In contrast, GC% values increase with increasing gene length (Fig. 1b), but less than 1% of the variation between points can be accounted for on this basis ($r^2 = 0.006$). However, when analyzed in terms of codon positions, a significant role of second codon positions emerges ($r^2 = 0.048$; Fig. 1f), which is partially countermanded by third codon positions (Fig. 1h). Thus, for both AG% and GC% the amino acid-determining codon positions (first and/or second) are involved in changes in average base compositions as ORF lengths increase.

The contributions of subsets of ORFs of increasing length to these results were examined by eliminating all ORFs above certain lengths (i.e., the arrowheads were progressively blunted). From first order linear regression analyses of these distributions, values for slopes and coefficients of determination (adjusted r^2) were extracted. Figure 2a shows that for AG%, even when

Table 1 Prokaryote genomes studied

Domain	Species	Number of available ORFs	Optimum growth		Genome base compositions[a]	
			pH	Temperature °C	GC%	AG%
Archaea	*Aeropyrum pernix*	2,694	7	92.5	57.52	52.53
	Archaeoglobus fulgidus	2,407	6.5	83	49.36	56.05
	Halobacterium sp. *NRC-1*	2,605	6.9	42	66.88	51.16
	Methanothermobacter thermautotrophicus	1,869	7.5	67.5	50.56	55.03
	Methanocaldococcus jannaschii	1,771	6	85	31.84	58.97
	Methanococcus maripaldis S2	1,801	7.25	38	34.08	56.76
	Methanopyrus kandleri AV19	1,687	6.5	98	61.2	54.42
	Methanosarcina mazei Go1	3,371	7	37	44.18	54.33
	Pyrococcus abyssi	1,791	6.9	97	45.14	57.28
	Pyrococcus horikoshii OT3	1,731	6.9	98	42.45	55.93
	Picrophilus torridus DSM9790	1,535	0.7	60	37.07	55.67
	Pyrobaculum aerophilum str. *IM2*	2,613	7	100	51.9	54.7
	Pyrococcus furiosus DSM3638	2,064	6.9	100	41.08	57.35
	Sulfolobus acidocaldarius	142	3	75	36.06	57.63
	Sulfolobus solfataricus P2	2,994	3.25	87	36.49	56.07
	Sulfolobus tokodaii str. 7	2,827	2.5	80	33.58	55.56
	Thermoplasma acidophilum DSM1228	1,522	2	59	47.37	54.62
	Thermoplasma volcanium GSS1	1,526	2	60	40.99	54.95
Bacteria	*Bacillus halodurans C-125*	4,066	9.5	47.5	44.32	53.82
	Erwinia carotovora subsp. *Atroseptica*	4,439	7.1	75	52.19	51.34
	Escherichia coli K12	5,048	6.5	37	51.81	51.39
	Geobacillus stearothermophilus	557	10	51	49.88	54.46
	Pseudomonas aeruginosa PA01	5,566	6.8	37	67.14	49.38
	Staphylococcus aureus MRSA252	2,651	7.25	33.5	33.54	54.99
	Streptococcus pneumoniae TIGR4	2,094	7.8	37	40.55	52.21

[a] Values approximated by calculating from the sum of the base compositions of all available ORFs

longer ORFs are omitted, slope values (AG%/kb) are still negative (filled symbols). Indeed, when all ORFs above 1.5 kb are omitted some 8% of the variation can be explained on this basis (r^2 = 0.08; open symbols). Although all codon positions contribute, among the codon positions, negative slope values and r^2 values tend to be consistently greatest at second positions (Fig. 2e). However, third codon positions make an increased contribution when all ORFs above 1 kb are omitted (r^2 = 0.06; Fig. 2g).

Figure 2b shows that for GC%, even when longer ORFs are omitted, slope values (GC%/kb) remain weakly positive, but become negative when ORFs above 0.75 kb are omitted. The positive slopes are largely due to second codon positions. When ORFs above 1.5 and 2.5 kb are omitted (Fig. 2f), from 7 to 8% of the variation between points can be explained on this basis (r^2 = 0.07–0.08). Figure 2h shows that negative slope values, when smaller ORFs alone are considered (i.e., when cut-off points are low), are largely due to changes at third codon positions. Here weakly negative slope values become progressively more negative as longer ORFs are omitted from the distributions. When all ORFs above 1 kb are omitted, some 2% of the variation between third codon positions can be explained on this basis (r^2 = 0.02). Apart from some differences when cut-off points were low, similar results were obtained in a study of *Thermoplasma acidophilum*, which has an optimum growth pH of 2.0 (Fig. 3).

Thus, long proteins tend to have a greater proportion of purine-poor codons than relatively shorter proteins. This holds over a wide range of length scales and, indeed, becomes statistically more certain over the low and middle range of lengths. With respect to impacting ORF length, a need to shed purines, perhaps due to a low optimum growth pH, would seem to trump a need to load purines due to a high optimum growth temperature. Does this amplify a trend towards purine loss that affects all ORFs? That is, irrespective of their lengths, are ORFs generally purine-depleted in organisms that have low growth pH optima? To examine this, and to determine the extent to which our observations on *P. torridus* are applicable to other acidophiles, we compared the overall base compositions of the ORFs of a variety of species that differ in their optimum growth conditions.

Direct measurement of AG%

Measurements of base compositions at different codon positions were made for 25 prokaryotes (including *P. torridus*), that vary both in their optimum growth pHs and temperatures (Table 1). Whereas in Figs. 1, 2, and 3 each point corresponds to a subset of genes within a

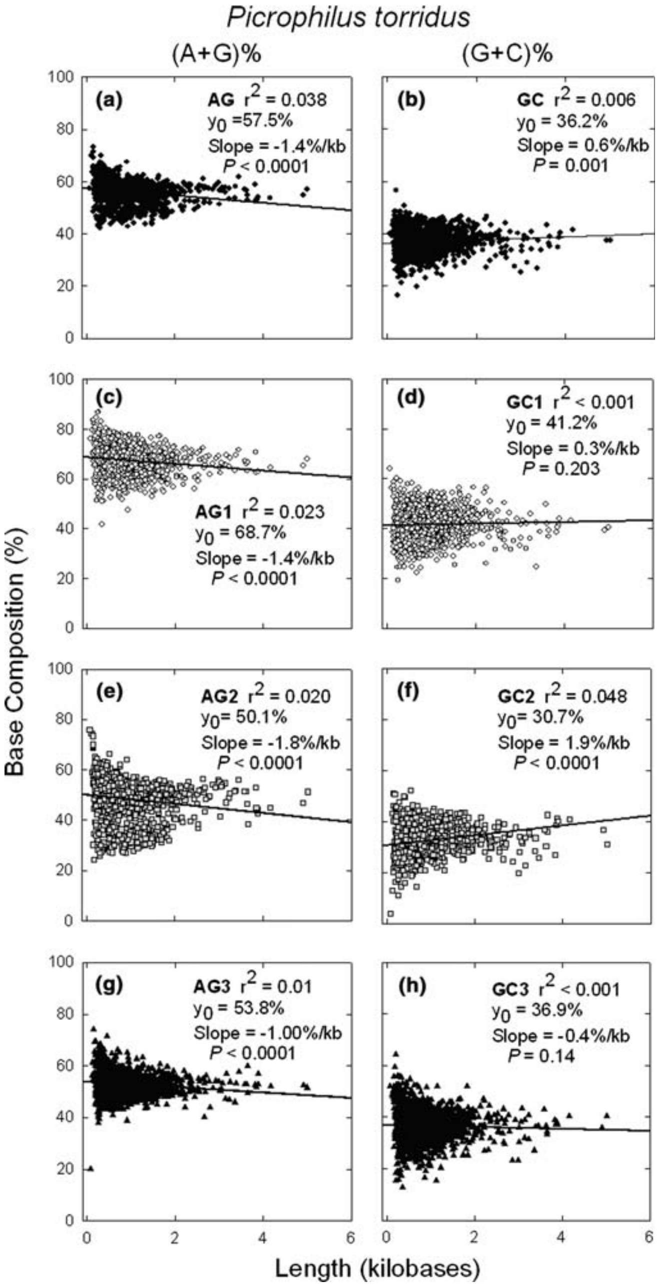


Fig. 1 Variation of base composition with ORF length for all, and for different, codon positions. Each point represents one of the 1,535 genes of *P. torridus*. Base compositions (*GC*% or *AG*%) are either for all positions. (**a** *AG*; **b** *GC*, or for different codon positions) (**c** *AG1*, **e** *AG2*, **g** *AG3*, **d** *GC1*, **f** *GC2*, **h** *GC3*). Points were fitted to first order linear regression lines (r^2 = adjusted square of the correlation coefficient, Y_0 = intercept at the ordinate, P = probability that the slope is not significantly different from zero). *AG*% values above 50% indicate "purine-loading"

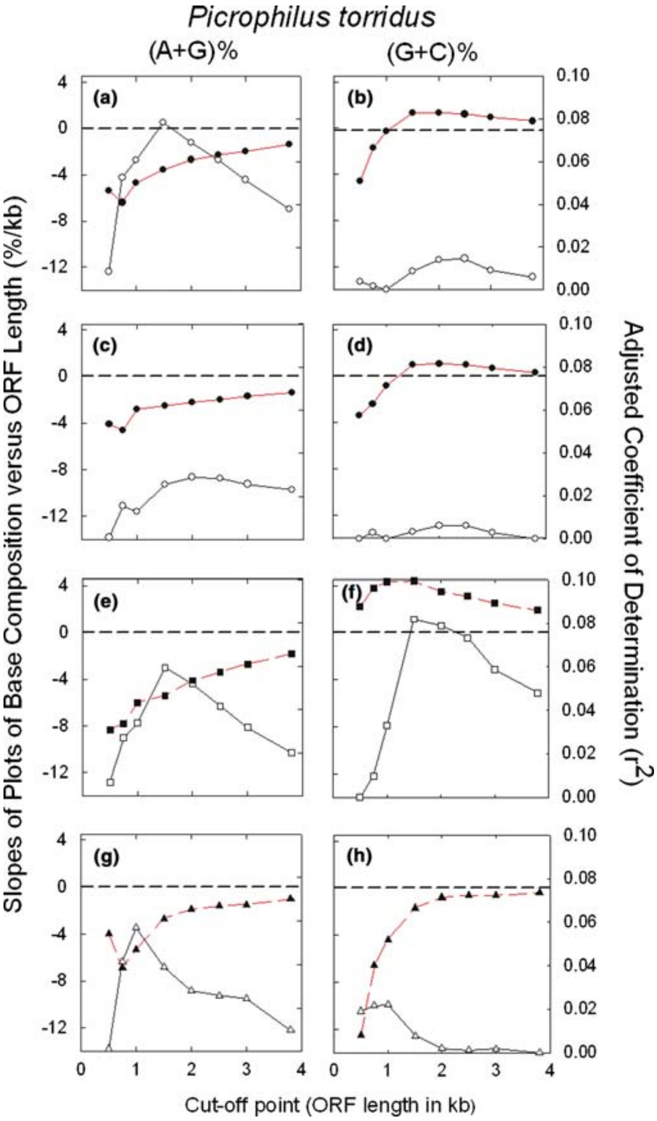


Fig. 2 Slopes (*filled symbols*) and r^2 values (*open symbols*) from plots of base composition versus ORF length for different ORF length subsets of *P. torridus* genes (e.g., a cut-off point of 2 kb means that all ORFs of length greater than 2 kb were omitted). For further details see the legend to Fig. 1

species, in Fig. 4 each point refers to all the sequenced genes of a species. There is no significant effect of optimum growth pH on either overall AG% values (Fig. 4a), or the AG% values for individual codon positions (Fig. 4c, e, g). Albeit at an extreme, the points for *P. torridus* (indicated by arrows) follow the general trend (i.e., purine-loading, mostly at first codon positions).

The same values were plotted against optimum growth temperatures. Consistent with previous studies showing that AG% increases as optimum growth temperature increases (Lambros et al. 2003), the slope is positive (r^2 = 0.15; P = 0.031; Fig. 4b). This is due largely to AG% values at first and third codon positions (Fig. 4d, h), particularly the latter (r^2 = 0.37; P = 0.001). Again, *P. torridus* follows the general trend.

Thus, apart from participating in a barely significant decline in AG% at second codon positions (r^2 = 0.098; P = 0.070; Fig. 4e), from direct measurement of base composition it must be concluded that *P. torridus* has resolved the putative conflict between pH and temperature in favor of temperature. As with other thermophiles, purines are loaded, not shed.

Direct measurement of GC%

Values for GC% were plotted similarly. These decrease as optimum growth pH decreases (Fig. 4a). The decrease is attributable to bases at all codon positions, but is significant only at first codon positions (Fig. 4c; $r^2 = 0.197$; $P = 0.015$). The points for *P. torridus* follow the general trend among organisms with low optimum growth pH. Thus, *P. torridus* has a lower overall GC% (i.e., it is relatively AT-rich) relative to organisms that have higher optimum growth pHs. The trend is weakly countermanded in *P. torridus* by a relative enrichment of GC% in ORFs encoding longer proteins (Fig. 1b, d, f; Fig. 2b, d, f). The same group of organisms shows no significant trend in GC% with optimum growth temperature (Fig. 4b, d, f, h). Thus, differences that correlate with optimum growth pH are unlikely to be explained by differences in optimum growth temperature.

Variation of AG slopes with growth conditions

For *P. torridus* there is an apparent discord between the purine-unloading of long ORFs relative to short ORFs (Figs. 1a, c, e, g; 2a, c, e, g), and an overall pH-independent purine-loading of all ORFs (Fig. 4a, c). To investigate this, slope values from arrowhead plots of base composition versus ORF length were obtained for the 24 other organisms whose base compositions are shown in Table 1 and Fig. 4. In every case all available ORFs were included, since a complete set of ORFs can give significant results (as in pointed arrow plots; Fig. 1), albeit not with the best r^2 values (as with blunted arrow plots; Figs. 2, 3). We chose not to make arbitrary cuts (to generate blunted arrow plots) among genomes that differ greatly in ORF number (Table 1). The slope values, including those for *P. torridus*, were plotted against the optimum growth parameters (pH and temperature) for each organism.

The unloading of purines as ORF length increases, as observed with *P. torridus*, occurs in other organisms with a low optimum growth pH (negative slopes of plots at low pH; Fig. 5a). This is largely due to differences at first and second codon positions, particularly the latter ($r^2 = 0.32$; $P = 0.002$; Fig. 5e). Thus, the weak tendency to unload purines in ORFs at second codon positions (Fig. 4e) is probably due to a preferential purine loss from the codons of longer ORFs.

These data indicate that in acidophilic microorganisms there is an increasing substitution (or insertion) of amino acids corresponding to purine-poor codons as protein length increases. No correlations were noted when the same slope values were plotted against optimum growth temperatures (Fig. 5b, d, f, h). Thus, the correlations relate to differences in optimum growth pH. Whereas increased optimum environmental temperature elevates purines generally (Fig. 4b) but does not differentially affect purines in long ORFs (Fig. 5b), decreased optimum environmental pH does not decrease purines generally (Fig. 4a) but does differentially affect purines in long ORFs (Fig. 5a).

Variation of GC-slopes with growth conditions

Figure 6 shows similar slope plots for GC%. The weak increase in GC% with ORF length observed in *P. torridus* (Fig. 1b, f) appears as part of a general trend, which actually becomes more evident as optimum growth pH increases. While remaining positive, slope values progressively decline as optimum growth pH decreases (Fig. 6a). This, however, is due mainly to differences at third codon positions ($r^2 = 0.16$; $P = 0.026$; Fig. 6g), so does not involve amino acid differences. Proteins would have needed to exchange (or insert) no particular amino acid to account for this. But for *P. torridus* the second codon position is primarily involved (Figs. 1f, 2f), so for acidophiles a differential usage of amino acids is implied (Fig. 6e).

In contrast to these results concerning optimum growth pH (Fig. 6a), with increasing optimum growth temperature slope values for GC% versus ORF length decrease ($r^2 = 0.266$; $P = 0.005$; Fig. 6b) but, in general, remain positive. This decrease primarily involves second codon positions (Fig. 6f), indicating a difference in amino acid content. Thus, relative to organisms growing optimally at 37°C, differences in GC% between ORFs encoding long and short proteins, although present, are less marked in thermophiles. ORFs for long proteins are enriched in GC relative to ORFs for short proteins, but this is a characteristic of most of the 25 species irrespective of optimum growth temperatures.

Discussion

There has been much interest in identifying differences between the proteins of mesophiles and the proteins of thermophiles that can account for the greater stability of the latter (Jaenicke and Bohm 1998; Kumar and Nussinov 2001; Fukuchi and Nishikawa 2001). Proteins of thermophiles are under pressure to reduce interdomain space, and their structures are more compact with fewer external loops corresponding to interdomain regions than in the orthologous structures of mesophiles (Thomson and Eisenberg 1999; Schafer et al. 2004). There are also differences in amino acid composition (Suhre and Claverie 2003). However, some amino acid differences appear not as a primary adaptation to thermophilia at the protein level, but secondary to adaptations at the nucleic acid level that result in nonsynonymous (therefore, amino acid changing) codon

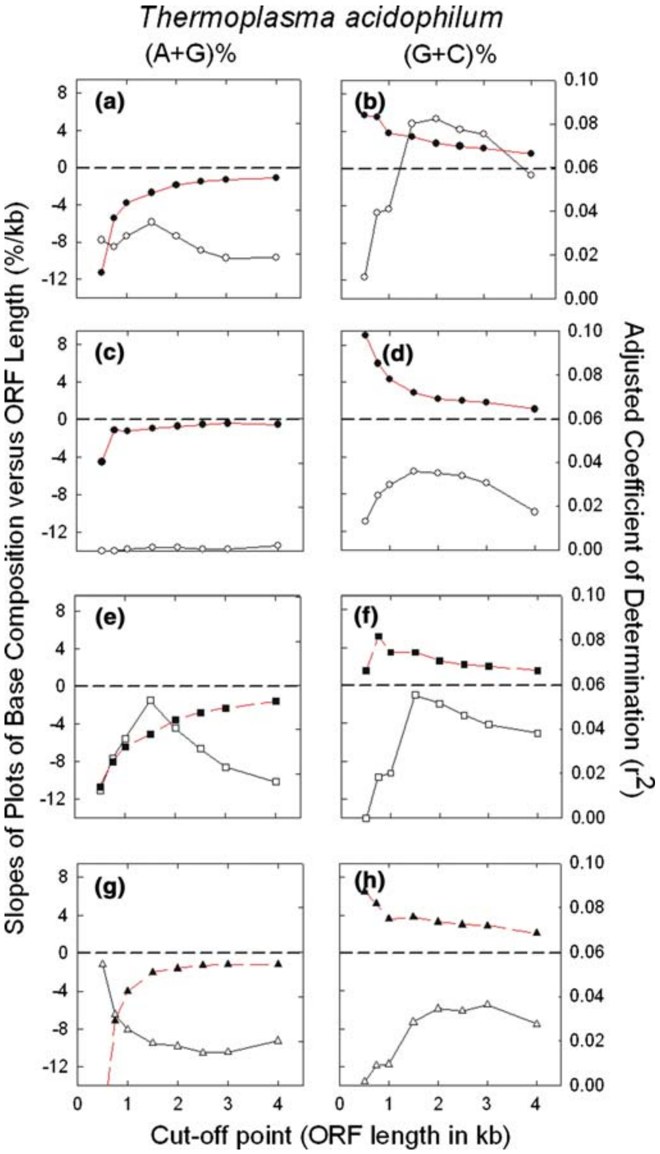


Fig. 3 Slopes (*filled symbols*) and r^2 values (*open symbols*) from plots of base composition versus ORF length for different ORF length subsets of *Thermoplasma acidophilum* genes. For further details see the legends to Figs. 1 and 2

changes (Lao and Forsdyke 2000; Lambros et al. 2003; Lobry and Chessel 2003; Paz et al. 2004; Basak et al. 2004). While changes in amino acid composition that might aid stability at low pH were anticipated in acidophiles, the only change in *P. torridus* is a slight increase in isoleucine (Fütterer et al. 2004). Whether there is a general reduction in interdomain space in acidophiles is not currently known (Schafer et al. 2004), but there is a reduction in intergenic space and coding sequences account for 91.7% of the *P. torridus* genome (Fütterer et al. 2004).

The primary nucleic acid level adaptation in thermophiles is enhanced loading of ORFs with purines. This is a phenomenon of protein-encoding regions, not of intergenic regions or introns, suggesting that it might reflect an adaptation operative at the mRNA level. The adaptation might reflect an evolutionary pressure primarily affecting transcripts, rather than differentially affecting one of the strands of ORFs encoding those transcripts. It has been proposed that the loading of the loop regions of mRNA secondary structures with one type of Watson–Crick-pairing base (e.g., purines rather than pyrimidines), would militate against the RNA–RNA interactions that could lead to formation of lengths of double-strand RNA extending beyond two helical turns. Such structures could delay mRNA-dependent processes (i.e., protein synthesis), and trigger false alarms indicative of cell stress (Cristillo et al. 2001; Forsdyke 2001, 2006; Forsdyke et al. 2002).

According to this argument, wherever possible, over evolutionary time pyrimidine-to-purine mutations have been accepted. *Multiple* such mutations would be needed to effectively "purine-load" an ORF, and a single mutation would have little impact. Since multiple mutations would be necessary, then they would be expected to be predominantly synonymous (Fig. 4h; Lambros et al. 2003). Organisms with multiple non-synonymous mutations would more likely be counterselected due to impairment of protein functions, but the pressure to purine-load does invoke some non-synonymous mutations (Lao and Forsdyke 2000; Mortimer and Forsdyke 2003).

We envisage a selective pressure on nucleic acids in acidophiles relating to their increased susceptibility to depurination under acid conditions. Error-prone DNA repair processes at a site of depurination would then lead to base changes (Bernstein and Bernstein 1991). A *single* mutation in a first or second (i.e., mainly non-synonymous) codon position in an ORF could, depending on its nature and position, impact the function of an entire protein. Long ORFs that are likely to contain more purines than short ORFs would be particularly vulnerable. So, wherever feasible, adaptation to acid environments should involve non-synonymous purine-to-pyrimidine substitutions in long ORFs (Fig. 5). Since acidophiles other than *Picrophilus* are, despite their acid environment, held to maintain their internal pH close to neutral (Ciaramella et al. 2004), our observation that these organisms follow the same depurination trend in long ORFs as *P. torridus* (Fig. 5), suggests that their neutrality controls may be imperfect. In other words, while they may maintain near-neutral internal pH under laboratory conditions, environmental fluctuations (e.g., nutrient deprivation, hypothermia) may decrease this control, so allowing occasional transient excursions of internal pH to that of the environment. While decreasing susceptibility to future mutation, purine-to-pyrimidine substitutions could also introduce amino acids that are more consistent with protein function under acid conditions. In long ORFs the advantage with respect to thermophilia conferred by purine-loading, would be trumped by the advantage with respect to acidophilia conferred by purine-unloading. This predicts, for those seeking to detect amino acid differences between

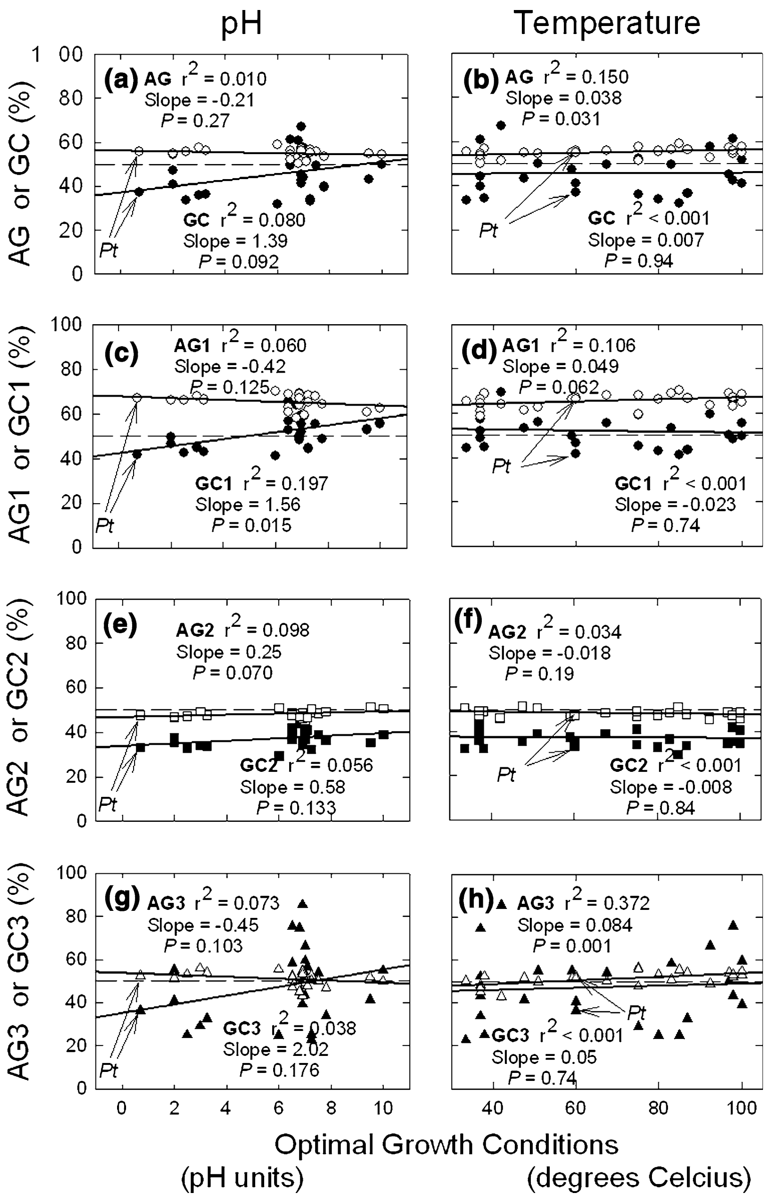


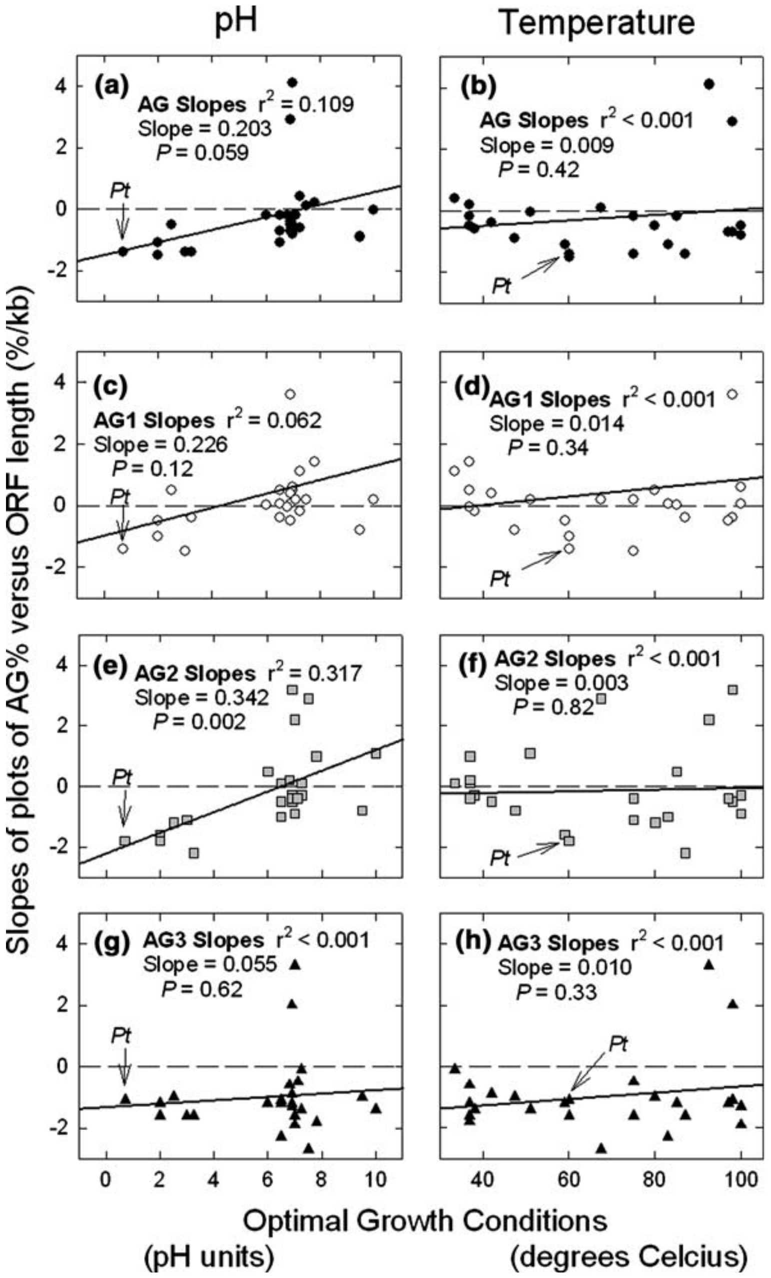
Fig. 4 Variation of base composition for all available sequenced genes of each of 25 prokaryotic species, either with optimum growth pH (**a**, **c**, **e**, **g**) or optimum growth temperature (**b**, **d**, **f**, **h**). Each point corresponds to a species. Points for *P. torridus* are indicated by arrows (*Pt*). *Open symbols* refer to AG% values, and *closed symbols* refer to GC% values at either, all codon positions (**a**, **b**), or individual codon positions (**c**, **d**, **e**, **f**, **g**, **h**), as indicated in the figures. Points were fitted to first order linear regression lines (r^2 = adjusted square of the correlation coefficient; *P* = probability that a slope is not significantly different from zero). *Horizontal dashed lines* refer to base compositions of 50%. *Open symbols* above these lines indicate "purine-loading"

organisms growing optimally in different environments (Suhre and Claverie 2003), that it would be more informative to compare amino acid differences between long orthologs.

The above explanation for the negative effect of ORF length on purine-loading in acidophiles (Figs. 1, 2, 3, and 5), seems more likely than another argument for length-dependent effects on ORF base composition, namely that long proteins are more likely to consist of multiple independent domains, so that the corresponding ORFs are likely to have a greater content of more mutable interdomain sequence, which can accept new "placeholder" amino acids (Rayment and Forsdyke 2005). As discussed above, we expect extremophile proteins to have compact structures with less opportunity for interdomain mutations. Furthermore, adding a pyrimidine-rich codon would do nothing to address the problem, a need to decrease the number of purine-rich codons.

Another explanation for ORF length effects on base composition is that mutations to stop-codons (classically UAA, UAG, and UGA) are more likely to be lethal in long ORFs (Oliver and Marin 1996). These codons are all AG-rich and GC-poor. Accordingly, members of the set of codons able to undergo one-step mutation to a stop-codon would be likely to accept mutations that would generate AG-poor and GC-rich synonymous codons. For example, serine codons UCA and UCG would accept pyrimidines at third codon positions generating the codons UCU and UCC, which would not give rise to stop-codons by one-step mutation. The glycine codon GGA would be more likely to accept a pyrimidine at its third position, so generating GGU and GGC. Thus, long ORFs should be relatively AG depleted (Fig. 1a) and GC enriched (Fig. 1b). However, it is difficult to envisage why this should be specific to acidophiles.

Fig. 5 Variation of slopes of plots of base composition (AG%) versus ORF lengths for various species (as shown for *P. torridus* in Fig. 1), with either optimum growth pH (**a**, **c**, **e**, **g**) or optimum growth temperature (**b**, **d**, **f**, **h**). Each point corresponds to a species. Those for *P. torridus* are indicated by *arrows* (*Pt*)



Regarding amino acid content, we should note that, consistent with the AT richness of acidophiles (Fig. 4a), the reported slight elevation in isoleucine in *P. torridus* would correspond to an increase in one or more GC-poor codons (AUU, AUC, AUA). Whether the isoleucine elevation serves a function at the protein level, or is actually secondary to adaptations at the nucleic acid level, remains to be determined. Regarding nucleic acid level adaptations, there is a reciprocal relationship between AG% and GC% (Lao and Forsdyke 2000; Saccone et al. 2001; Mortimer and Forsdyke 2003). So a primary effect on one, may secondarily affect the other. The underlying basis of this relationship can be determined from first principles. There are two ways to modulate GC% when the total number of bases is constant; either by changing the number of G's, or by changing the number of C's. As GC% increases, trading A for G would not affect the AG%. Likewise, trading T for C would not affect the AG%. However, if G replaced T, AG% would increase as GC% increases (non-reciprocal relationship). Conversely, if A were replaced with C, AG% would decrease as GC% increases (reciprocal relationship; Lambros et al. 2003). So A-for-C trading may partly explain some of the effects observed here. For example, such trading would decrease the vulnerability of a strand to cytosine deamination, which is favored at extremes of pH, and preferentially affects single strands of DNA (Beletskii and Bhagwat 1996).

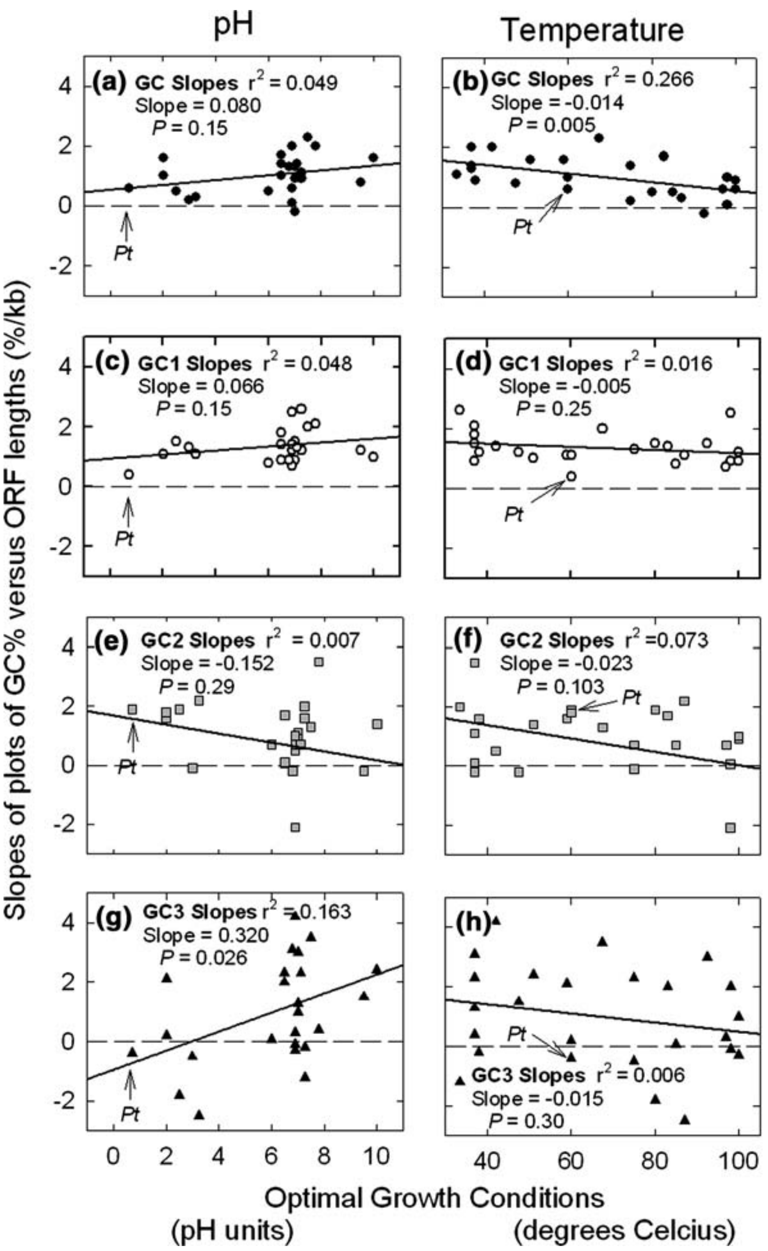


Fig. 6 Variation of slopes of plots of base composition (GC%) versus ORF lengths for various species (as shown for *P. torridus* in Fig. 1), with either optimum growth pH (**a**, **c**, **e**, **g**) or optimum growth temperature (**b**, **d**, **f**, **h**). Each point corresponds to a species. Those for *P. torridus* are indicated by *arrows* (*Pt*)

We have not here considered direct effects of low pH on mRNAs. Since mRNAs are multiple and ephemeral, a random change in a base in one mRNA is unlikely to have a long-term consequence for the species. Intriguingly, Amosova et al. (2006) have noted that when duplex DNA adopts extruded stem-loop configurations, certain short sequences of guanine residues can catalyze self-depurination under physiological conditions. Thus, there are depurination-mediated "hot-spots" for mutation in DNA. Such self-depurination increases when the pH is lowered, but decreases at high temperature due to disruption of the required DNA stem-loop structure. Possible implications of the present study for the base composition of nucleic acids in the relatively alkaline environment of mitochondria are discussed elsewhere (Forsdyke 2006).

Conclusion

There are circumstances under which it is of greater advantage to have a particular base in a particular position in a nucleic acid, than to have a particular amino acid in a particular position in a protein. The pressure to encode a particular amino acid is great, but it can be over-ridden by other pressures on the genome phenotype (Sueoka 1961; Schaap 1971; Lao and Forsdyke 2000; Chang and Benner 2004; Forsdyke 2006). Environmental pH appears to be one such pressure.

Acknowledgments We thank J. R. Mortimer and J. H. Rayment for programs that extract and display base compositions. Queen's University host's Forsdyke's webpages which contain full-text versions of some of the cited references (http://www.post.queensu.ca/~forsdyke/bioinfor.htm).

References

Amosova O, Coulter R, Fresco JR (2006) Self-catalyzed site-specific depurination of guanine residues within gene sequences. Proc Natl Acad Sci USA 103:4392–4397

Basak S, Ghosh TC (2005) On the origin of genomic adaptation at high temperature for prokaryotic organisms. Biochem Biophys Res Comm 330:629–632

Basak S, Banerjee T, Gupta SK, Ghosh TC (2004) Investigation on the causes of codon and amino acid usages variation between thermophilic *Aquifex aeolicus* and mesophilic *Bacillus subtilis*. J Biomol Struct Dyn 22:205–214

Beletskii A, Bhagwat AS (1996) Transcription-induced mutations: increase in C to T mutations in the nontranscribed strand during transcription in *Escherichia coli*. Proc Natl Acad Sci USA 93:13919–13924

Bernstein C, Bernstein H (1991) Aging, Sex and DNA Repair. Academic, San Diego

Chang MSS, Benner SA (2004) Empirical analysis of protein insertions and deletions determining parameters for the correct placement of gaps in protein sequence alignments. J Mol Biol 341:617–631

Ciaramella M, Napoli A, Rossi M (2005) Another extreme genome: how to live at pH 0. Trends Microbiol 13:49–51

Cristillo AD, Mortimer JR, Barrette IH, Lillicrap TP, Forsdyke DR (2001) Double-stranded RNA as a not-self alarm signal: to evade, most viruses purine-load their RNAs, but some (HTLV–1, EBV) pyrimidine-load. J Theor Biol 208:475–491

Forsdyke DR (2001) The Origin of Species, Revisited. McGill-Queen's University Press, Montreal

Forsdyke DR (2006) Evolutionary Bioinformatics. Springer, New York

Forsdyke DR, Mortimer JR (2000) Chargaff's Legacy. Gene 261:127–137

Forsdyke DR, Madill CA, Smith SD (2002) Immunity as a function of the unicellular state: implications of emerging genomic data. Trends Immunol 23:575–579

Fukuchi S, Nishikawa K (2001) Protein surface amino acid compositions distinctively differ between thermophilic and mesophilic bacteria. J Mol Biol 309:835–843

Fütterer O, Angelov A, Liesegang H, Gottschalk G, Schleper C, Schepers B, Dock C, Antranikian G, Liebl W (2004) Genome sequence of *Picrophilus torridus* and its implications for life around pH 0. Proc Natl Acad Sci USA 101:9091–9096

Huang S-L, Wu L-C, Liang H-K, Pan K-T, Horng J-T, Ko M-T (2004) PGTdb: a database providing growth temperatures of prokaryotes. Bioinformatics 20:276–278

Jack LCM de V, Driessen AJM, Zillig W, Konings WN (1998) Bioenergetics and cytoplasmic membrane stability of the extremely acidophilic, thermophilic archaeon *Picrophilus oshimae*. Extremophiles 2:67–74

Jaenicke R, Bohm G (1998) The stability of proteins in extreme environments. Curr Opin Struct Biol 8:738–748

Kumar S, Nussinov R (2001) How do thermophilic proteins deal with heat? Cell Mol Life Sci 58:1216–1233

Lambros RJ, Mortimer JR, Forsdyke DR (2003) Optimum growth temperature and the base composition of open reading frames in prokaryotes. Extremophiles 7:443–450

Lao PJ, Forsdyke DR (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 JR, Chessel D (2003) Internal correspondence analysis of codon and amino acid usage in thermophilic bacteria. J Appl Genet 44:235–261

Macario A, Malz M, Macario EC de (2004) Evolution of chaperoning systems within the phylogenetic domain archaea. Front Biosci 9:1318–1332

Mortimer JR, Forsdyke DR (2003) Comparison of responses by bacteriophages and bacteria to pressures on the base composition of open reading frames. Appl Bioinf 1:47–62

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

Nakamura Y, Gojobori T, Ikemura T (2000) Codon usage tabulated from the international DNA sequence databases: status for the year. Nucleic Acids Res 28:292

Oliver JL, 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, Korol A (2004) Adaptive role of increased frequency of polypurine tracts in mRNA sequences of thermophilic prokaryotes. Proc Natl Acad Sci USA 101:2951–2956

Rayment J, Forsdyke DR (2005) Amino acids as placeholders: base composition pressures on protein length in malaria parasites and prokaryotes. Appl Bioinf 4:117–130

Saccone C, Gissi C, Lanave C, Larizza A, Pesole G, Reyes A (2000) Evolution of the mitochondrial genetic system: an overview. Gene 261:153–159

Schaap T (1971) Dual information in DNA and the evolution of the genetic code. J Theor Biol 32:293–298

Schäfer K, Magnusson U, Scheffel F, Schiefner A, Sandgren MOJ, Diederichs K, Welte W, Hülsmann A, Schneider E, Mowbray SL (2004) X-ray structures of the maltose–maltodextrin-binding protein of the thermoacidophilic bacterium *Alicyclobacillus acidocaldarius* provide insight into acid stability of proteins. J Mol Biol 335:261–274

Schleper C, Puehler G, Holz I, Gambacorta A, Janekovic D, Santarius U, Klenk H-P, Zillig W (1995) *Picrophilus* gen. nov., fam. nov.: a novel aerobic, heterotrophic thermoacidophilic genus and family comprising archaea capable of growth around pH 0. J Bacteriol 177:7050–7059

Sueoka N (1961) Compositional correlation between deoxyribonucleic acid and protein. Cold Spring Harb Symp Quant Biol 26:35–43

Suhre K, Claverie J-M (2003) Genomic correlates of hyperthermostability, an update. J Biol Chem 278:17198–17202

Tamm C, Hodes ME, Chargaff E (1952) The formation of apurinic acid from the desoxyribonucleic acid of calf thymus. J Biol Chem 195:49–63

Thomson MJ, Eisenberg D (1999) Transproteomic evidence of a loop-deletion mechanism for enhancing protein stability. J Mol Biol 290:595–604

Wiegel J, Kevbrin V (2003) Alkalithermophiles. Biochem Soc Trans 32:193–198

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