5300–5309 *Nucleic Acids Research, 1998, Vol. 26, No. 23*

The hyperthermophilic bacterium *Thermotoga maritima* has two different classes of family C DNA polymerases: evolutionary implications

Yi-Ping Huang and Junetsu Ito*

Department of Microbiology and Immunology, College of Medicine, The University of Arizona, Tucson, AZ 85724, USA

Received August 28, 1998; Revised September 29, 1998; Accepted October 14, 1998 DDBJ/EMBL/GenBank accession nos+

ABSTRACT

Bacterial DNA polymerase III (family C DNA polymerase), the principal chromosomal replicative enzyme, is known to occur in at least three distinct forms which have provisionally been classified as class I (*Escherichia coli* DNA pol C-type), class II (*Bacillus subtilis* DNA pol C-type) and class III (cyanobacteria DNA pol C-type). We have identified two family C DNA polymerase sequences in the hyperthermophilic bacterium *Thermotoga maritima*. One DNA polymerase consisting of 842 amino acid residues and having a molecular weight of 97 213 belongs to class I. The other one, consisting of 1367 amino acid residues and having a molecular weight of 155 361, is a member of class II. Comparative sequence analyses suggest that the class II DNA polymerase is the principal DNA replicative enzyme of the microbe and that the class I DNA polymerase may be functionally inactive. A phylogenetic analysis using the class II enzyme indicates that *T.maritima* is closely related to the low G+C Gram-positive bacteria, in particular to *Clostridium acetobutylicum*, and mycoplasmas. These results are in conflict with 16S rRNA-based phylogenies, which placed *T.maritima* as one of the deepest branches of the bacterial tree.

INTRODUCTION

Hyperthermophiles are a fascinating group of microorganisms that have the remarkable property of growing at a temperature of 70°C or above (1,2). They have been discovered in geothermally heated environments that included deep sea hydrothermal vents. The hyperthermophiles are highly phylogenetically divergent microorganisms. Although most of the hyperthermophilic genera isolated to date belong to the domain Archaea (2,3), two groups belong to the domain Bacteria. These two groups are Thermotogales and Aquificales (2). *Thermotoga maritima* was originally isolated from a geothermally heated marine sediment at Vulcano, Italy (4). It grows at temperatures of 55–90°C, with an optimum temperature of 80°C. Thus, together with *Aquifex* species, *T.maritima* is one of the most thermophilic eubacteria presently known (2).

Thermotoga maritima is an anerobic, Gram-negative, rod-shaped bacterium which usually grows singly or in pairs. Cells of *T.maritima* are surrounded by a characteristic sheath-like structure (toga) with overballooning at the ends of the bacterium (1). *Thermotoga maritima* has a murein cell wall, unlike archaeal cell walls, and is, therefore, sensitive to lysozyme (1). Although it is susceptible to various antibiotics, such as chloramphenicol and penicillin, it is insensitive to the eubacterial RNA polymerase inhibitor rifampicin and to the action of aminoglycoside antibiotics, such as streptomycin and kanamycin (5,6). Furthermore, all species of the Thermotogales contain unique ether lipids (7). These features clearly distinguish Thermotogales from other eubacteria.

Phylogenies based on 16S rRNA sequences (8) and translational elongation factors, EF-Tu and EF-GC (9–11), indicate that Thermotogales are deep offshoots in the bacterial evolutionary tree. Indeed, it has been suggested that both *Thermotoga* and *Aquifex* are the earliest and perhaps the most slowly evolving branches in the bacterial domain (12). Since the earliest archaeal lineages are also hyperthermophilic, a theory has been advanced that the most recent common ancestor of extant life forms was hyperthermophilic (8,13). This evidence, together with geological studies suggesting that the hydrosphere was far hotter 3 billion years ago than now (14), strongly implies a hot origin of life on Earth (13). However, the recent discovery of Archaea in the cold pelagic ocean has cast some doubt on this theory (15,16). Furthermore, there are growing numbers of protein gene phylogenies which do not agree with the phylogenies based on 16S rRNA (17) and translational elongation factors (9–11). These include phylogenies based on RNA polymerase genes (18,19), ribosomal protein genes (6), glutamine synthetase genes (20–23), glutamate dehydrogenase genes (24), heat shock protein 70 genes (25), recA protein genes (26) and RNA polymerase σ factor 70-type genes (27), as well as others.

Because the phylogenetic position of the hyperthermophilic bacteria has far-reaching evolutionary implications, we have initiated phylogenetic analyses of DNA polymerase genes of *T.maritima*. Gelfand and co-workers (28) have cloned and overexpressed heat-stable DNA polymerase I (family A DNA polymerase) from *T.maritima*. However, the replicative DNA polymerase has yet to be determined. *Thermotoga maritima* contains a thermostable 4Fe ferredoxin whose properties and amino acid sequence are much more similar to those of ferredoxins from hyperthermophilic Archaea than ferredoxins from mesophilic or moderately thermophilic Bacteria (29). All hyperthermophilic Archaea examined so far have family B DNA

*To whom correspondence should be addressed. Tel: +1 520 626 7755; Fax: +1 520 626 2100; Email: ito@u.arizona.edu

+AF063188 and AF065313

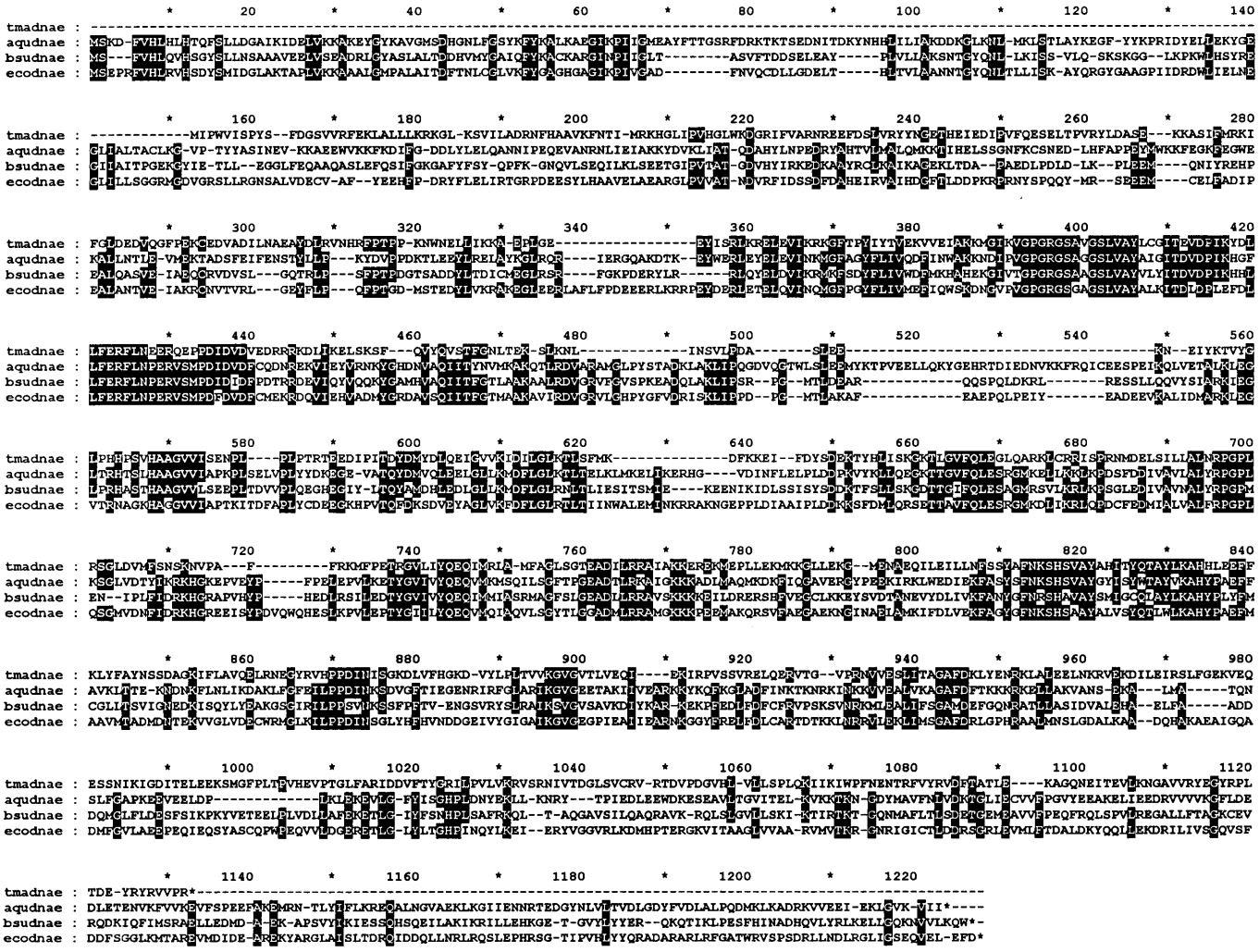


Figure 1. Amino acid sequence comparison of *T.maritima dna*E protein (class I enzyme) with those of *Aquifex aeolicus dna*E, *B.subtilis dna*E1 and *E.coli dna*E. Protein names are as follows: tmadnae, *T.maritima dna*E; aqudnae, *A.aeolicus dna*E (41); bsudnae, *B.subtilis dna*E1 (42); ecodnae, *E.coli dna*E (43). The multiple sequence alignment was performed using the PILEUP program from GCG (38). Positions are highlighted in black in which sequence conservation is either 75 or 100% identical.

polymerases (DNA polymerase II) (30–34) which are replicative enzymes. Eukaryotic replicative enzymes, DNA polymerase α, δ and ε, are also members of the family B DNA polymerases (30,35). Therefore, it was of interest to determine whether the hyperthermophilic bacterium *T.maritima* also contains a family B DNA polymerase. We first searched for possible family B-like DNA polymerase genes in *T.maritima* using PCR-mediated gene amplification technology. While family B DNA polymerase genes were successfully amplified by the PCR method from members of the γ subdivisions of Proteobacteria (Y.-P.Huang and J.Ito, in preparation), our attempt to find a family B-like DNA polymerase gene in *T.maritima* was unsuccessful. We then attempted to isolate a family C DNA polymerase gene using sets of degenerate oligonucleotide primers which were synthesized based on highly conserved regions of the family C DNA polymerases (36). Surprisingly, we were able to identify two family C DNA polymerase genes which belong to two different classes. In this report, comparative analyses of the two *T.maritima* family C DNA polymerase genes with those of other eubacteria are described and evolutionary implications are discussed.

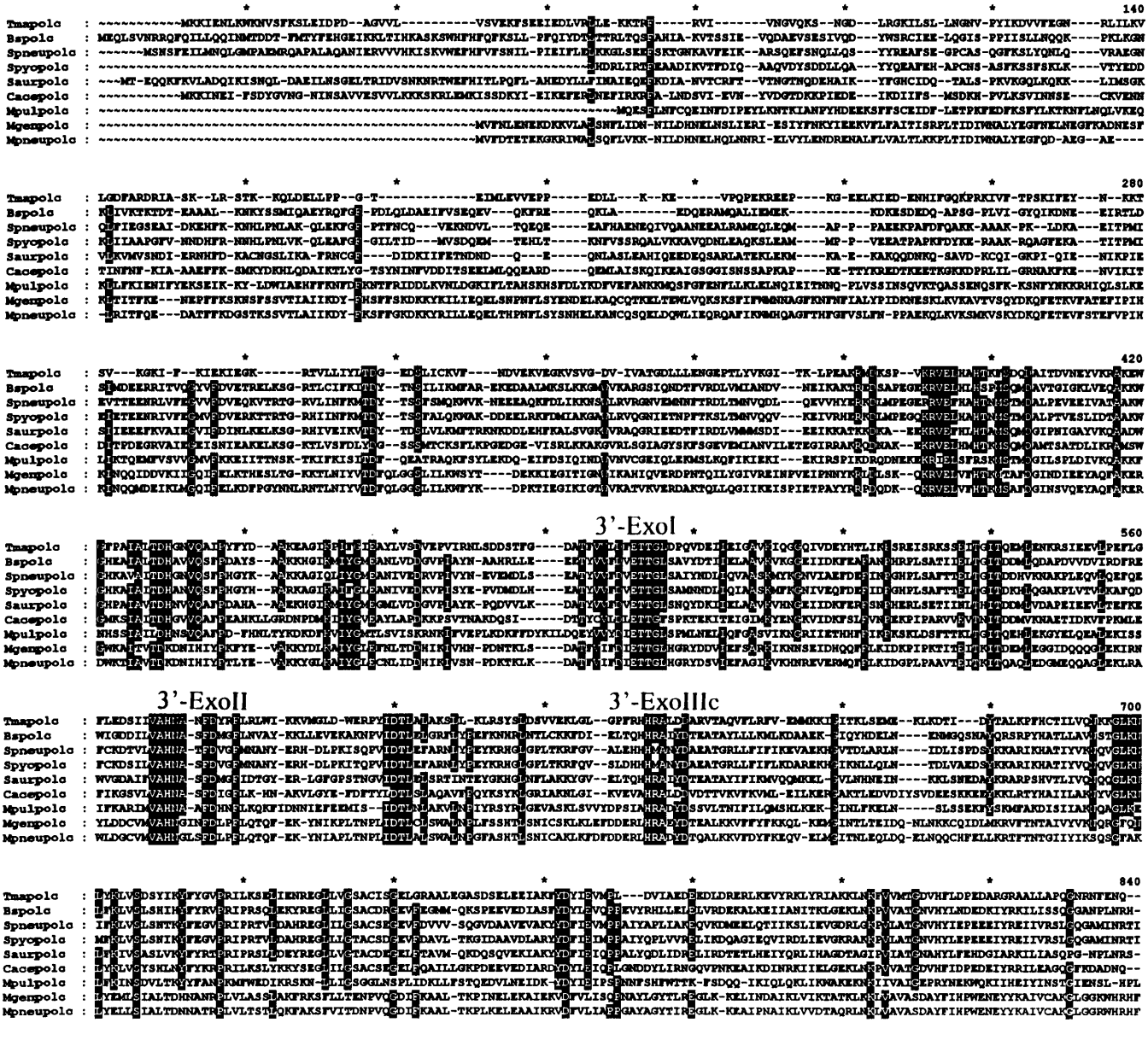
MATERIALS AND METHODS

Bacterial strain and genomic DNA

Thermotoga maritima (DSM 3109) DNA was kindly provided by Dr Robert Huber (Lehrstuhl für Mikrobiologie, Universität Regenoburg, Regensburg, Germany). *Escherichia coli* DH5α was used for cloning PCR-amplified DNA fragments and for preparation of plasmid DNA for sequencing.

PCR amplification and cloning of family C DNA polymerase genes

Degenerate oligonucleotide primers (forward and reverse) were designed based on conserved regions of multiple amino acid sequence alignments from a wide variety of family C DNA polymerases (36). Several oligonucleotide primers containing inosinic acid were also synthesized (Gibco BRL). Among them two sets of oligonucleotide primers worked well in DNA amplification for class I and class II family C DNA polymerases. The PCR primers for the class I family C DNA polymerases were



5′-ATG CCY GAC TTC GAY RTN GAY TTC (forward primer) and 5′-CYT CNG GYT TYT TYT TWC CCA TYG C (reverse primer), where Y is either C or T, R is either A or G, N is any of the four possible nucleotides and W is either A or T. The set of PCR primers for the class II family C DNA polymerases were 5′-CCI GAT ATY GAY TTR AAC TTY TC (forward primer) and 5′-GCC ATI ADD ACR TAD GCY GCI GC (reverse primer), where I is inosinic acid. PCR amplification using the above sets of primers and ~100 ng genomic DNA was carried out for 35 cycles of 60 s each at 94, 55 and 72°C, followed by a final 10 min extension step at 72°C. The amplified products for class I (*dna*E-type) and class II (DNA pol C-type) were ~960 and 940 bp, respectively. The PCR products were electrophoresed on agarose gels and the DNA fragments were extracted and purified using the Prep-Gene-DNA purification system (Bio-Rad). They were then ligated to the T-tailed vector pGEM-T (Promega) and the ligated material was used to transform *E.coli* DH5α competent cells.

In order to obtain complete nucleotide sequences of the family C DNA polymerase genes, the universal Genomewalking kit (Clontech) was used. First, we constructed the genomic libraries using six different blunt-ended restriction endonucleases (*Dra*I, *Eco*RV, *Pvu*II, *Sca*I, *Ssp*I and *Stu*I). The cleaved genomic DNA was ligated to Genomewalking adapter at 16°C for 18 h. The ligation mixture was then incubated at 70°C for 5 min to stop the ligation reaction. Based on the partial nucleotide sequence of *T.maritima* (Tma) DNA polymerase III, two pairs of oligonucleotide primers (24–28 bp length) were synthesized and used, together with the Genomewalking adapter primers, to perform PCR to extend the family C DNA polymerase sequence from both directions. These procedures were repeated several times to

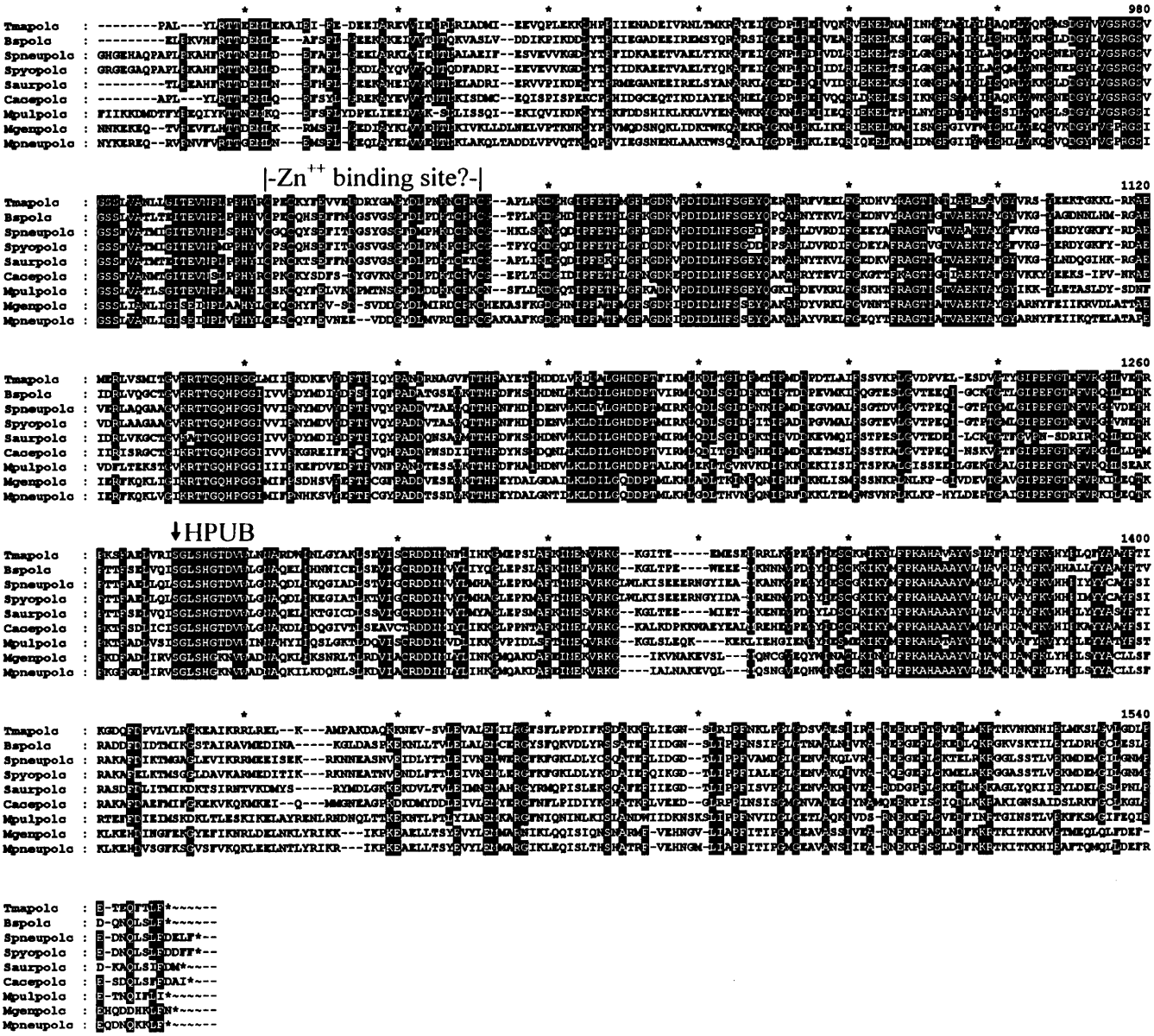


Figure 2. Amino acid sequence comparison of *T.maritima* DNA pol C (class II) with those of eight other family C DNA polymerases from various bacteria. Protein names and references are as follows: tmapolc, *T.maritima* DNA polymerase C; bsupolc, *B.subtilis* DNA polymerase C (42); spneupolc, *Streptococcus pneumoniae* DNA polymerase C (*S.pneumoniae* pol C sequence data were obtained through early release from the Institute for Genomic Research, www.tigr.org at www.ncbi.nlm.nih.gov); spyopolc, *Streptococcus pyogenes* DNA polymerase C (*S.pyogenes* pol C sequence data were obtained through early release from the University of Okalahoma, http://dna1.chem.ou.edu/strep.html); saurpolc, *Staphylococcus aureus* DNA polymerase C (44); cacepolc, *Clostridium acetobutylicum* DNA polymerase C (*C.acetobutylicum* sequence data were obtained through the Internet, http://www.ncbi.nlm.nih.gov/BLAST/unfinishedgenome.html); mpulpolc, *Mycoplasma pulmonis* DNA polymerase C (45); mgenpolc, *Mycoplasma genitalium* DNA polymerase C (46); mpneupolc, *Mycoplasma pneumoniae* DNA polymerase C (47). Conserved amino acid sequence motifs for 3′→5′ exonuclease domains are indicated as 3′-Exo I, 3′-Exo II and 3′-Exo IIIC. The 3′-Exo IIIC motif of family C DNA polymerases, HxAxxD, is different from the YxxxD motif common to family A and family B DNA polymerases (48–50). To distinguish, we have labeled the motif 3′-Exo IIIC for family C DNA polymerases and 3′-Exo III for family A and B DNA polymerases. The class II family C DNA polymerases are highly sensitive to the potent inhibitor HPura (51,52). The HPura binding site is indicated by the arrow and labeled as HPUB (HPura binding site) (53). $Cx_2Cx_{21}Cx_2C$, where x can be any amino acid, is a Zn^{2+} binding domain-like sequence. Invariant amino acid residues are highlighted in black.

complete the entire sequence of Tma family C DNA polymerase genes. All PCR products were cloned into pGEM-T Easy Vector (Promega) for DNA sequencing of the products. Nucleotide sequences were determined using the dideoxynucleotide chain termination method (37) with the Sequenase v.2.0 DNA sequencing kit (US Biochemical). Both strands were sequenced.

Sequence analysis

Sequence data processing and editing were performed using a GCG package (38). Sequence alignments were carried out using the PILEUP program of the GCG or the CLUSTAL W program (39), followed by manual adjustment where necessary. For the

phylogenetic analysis of the class II family C DNA polymerases, only the most conserved sections, where the alignment was unambiguous, were used. Phylogenetic trees were generated using algorithms implemented in the PAUP* test version software package (40). Distance and parsimony analyses were based on 250 bootstrap replicates.

RESULTS

Nucleotide and predicted amino acid sequence of the Tma family C DNA polymerases

Using the PCR-mediated gene amplification method, we searched for the family C DNA polymerase gene of *T.maritima*. Several sets of PCR primers were designed based on multiple amino acid sequence alignments from a wide variety of family C DNA polymerases (36). The PCR-amplified DNA fragments (~940–980 bp) were cloned into a pGEM-T vector and sequenced as described in Materials and Methods.

To our surprise, we found two different sequences from PCR-amplified DNA fragments. The deduced amino acid sequences from both DNA fragments suggest that one is related to the class I and the other is related to the class II family C DNA polymerases. The nucleotide sequences and deduced amino acid sequences of these two DNA polymerase III genes of *T.maritima*, together with the flanking nucleotide sequences, were deposited in GenBank with the accession nos AF063188 for the Tma *dna*E and AF065313 for the Tma DNA pol C gene.

In Figure 1, the Tma *dna*E protein sequence is compared with those of *Aquifex dna*E (41), *Bacillus subtilis dna*E1 (42) and *E.coli dna*E (43). The Tma *dna*E protein consists of 842 amino acid residues with a calculated molecular weight of 96.516 kDa, whereas *Aquifex dna*E, *B.subtilis dna*E1 and *E.coli dna*E proteins are 133.200, 125.348 and 129.903 kDa, respectively. As can be seen from Figure 1, Tma *dna*E protein is short at both the N- and C-terminal ends. In addition, it contains several internal deletions. The Tma *dna*E protein is ~20% shorter than others. Therefore, whether or not this protein is functional remains to be determined.

In Figure 2, the amino acid sequence of Tma DNA pol C is compared with those of eight other DNA polymerase Cs from various bacteria. In contrast to the Tma *dna*E protein, Tma DNA pol C appears to be a bona fide DNA polymerase C. It contains three highly conserved amino acid sequence motifs in an editing 3′→5′ exonuclease domain, which are indicated as 3′-Exo I, 3′-Exo II and 3′-Exo IIIc (48–50). Although specific amino acid residues for the DNA polymerase catalytic domain have yet to be identified, a highly specific DNA pol C inhibitor [6-(*p*-hydroxyphenylazo)-uracil, HPura] binding site (53,54) and surrounding sequences are highly conserved, as indicated by the arrow in Figure 2. In addition, all the class II DNA polymerases C contain highly conserved 'Zn^{2+} binding domain-like' sequences, $Cx_2Cx_{21}Cx_2C$, where x can be any amino acid.

Phylogenetic analysis of the Tma class II family C DNA polymerase

In reconstructing our phylogenetic tree, we removed all gaps from the multiple sequence alignment (Fig. 2). A phylogenetic analysis using the edited sequence alignment is shown in Figure 3. Two phylogenetic inference methods, distance and parsimony, produced very similar topologies (40). Figure 3 shows only the tree produced as a result of the neighbor joining method (55). *Escherichia coli dna*E protein was used as an outgroup. The bootstrap values are annotated at nodes. Clearly, the Tma DNA pol C is closely related to *Clostridium acetobutylicum* DNA pol C (Cac DNA pol C). An amino acid sequence similarity test also indicated that the Tma DNA pol C is most closely related to the *C.acetobutylicum* DNA pol C (with 69% similarity and 47% identity) (36). Although three *Mycoplasma* DNA polymerase Cs appear to be ancestral to the other DNA pol C genes, this antiquity may be an artifact due to the unique character of *Mycoplasma* genes, which are well documented to have undergone rapid evolution (56–58). While some investigators place the mycoplasmas within the low G+C Gram-positive group (57,58), others argue that the mycoplasmas form a phylogenetically distinct group that is distantly related to the low G+C Gram-positive bacteria (59). It is interesting to note that the mycoplasmas do not have family A DNA polymerases (46,47).

Comparison of nucleotide changes

Since all bacteria so far tested contain class I family C DNA polymerases (36) and only members of three bacterial phyla (low G+C Gram-positive bacteria such as *Clostridium*, *Bacillus* and *Lactobacillus*, mycoplasmas and Thermotogales) use class II enzyme for their chromosomal replication, one wonders if the class II family C DNA polymerases were acquired laterally from another domain of life, perhaps from Archaea. To investigate this possibility, the G+C contents of the Tma family C DNA polymerases were compared with that of whole genome DNA. Similar comparisons were also made with other family C DNA polymerase genes and respective chromosomal DNA (Table 1). Horizontally transferred genes often contain a different G+C content than the host genomic DNA. Although the G+C content of the Tma class II enzyme is sightly higher than that of the genomic DNA, the G+C content of the *E.coli dna*E gene is even higher than that of its chromosomal DNA. Thus, it is not apparent whether or not the class II family C DNA polymerase gene arose by horizontal evolution.

We next compared nucleotide change between the respective classes of the family C DNA polymerases (Table 2). Allowing for gaps, 570 codons which are synonymous and non-synonymous changes can be compared directly between the Tma DNA pol C and *B.subtilis* DNA pol C (Table 2). Similarly, 305 codons which are synonymous and non-synonymous changes can be directly compared between the Tma *dna*E and *Aquifex dna*E genes. The results indicate that between the Tma DNA pol C and *B.subtilis* DNA pol C genes many nucleotide changes have occurred, but most of them (70.35%) are synonymous changes which are without consequence for the amino acid sequence, because most of them occurred at the third position of the codon. In contrast, many nucleotide changes have occurred in Tma *dna*E and *Aquifex dna*E genes as well, but nearly 46% of them result in amino acid substitution. This clearly suggests that the Tma DNA pol C and Tma *dna*E genes are not changing at the same rate. These results appear to support the idea that the Tma *dna*E protein may not be functional and, thus, is under less selective pressure.

Table 1. Comparison of G + C content of the family C DNA polymerase genes

	T.maritima (%)		*A.aerolicus* (%)	*B.subtilis* (%)	*E.coli* (%)
	*dna*E	pol C	*dna*E	pol C	*dna*E
C DNA pol family	Class I	Class II	Class I	Class II	Class I
DNA pol coding sequences	45.8	49.0	46.6	42.7	55.0
Codon fisrt position	45.6	53.4	53.5	47.8	68.2
Codon second position	32.6	32.8	28.7	33.4	37.8
Codon third position	59.9	51.1	57.4	40.7	63.7
Whole genome	46.0		43.0	43.0	51.0

Table 2. Comparison of nucleotide changes

	Tma-Bsu pol C (%)	Tma-Eco *dna*E (%)	Tma-Aqu *dna*E (%)	Eco-Aqu *dna*E (%)
Total synonymous changes	70.35	60.40	54.10	68.26
First codon	0.53	1.65	0.33	0.24
Second codon	0	0	0	0
Third codon	62.28	50.17	49.83	60.92
Other[a]	7.54	8.58	3.94	7.40
Total non-synonymous changes	29.65	39.40	45.90	31.74
First codon	12.28	21.12	20.98	14.08
Second codon	9.65	10.23	15.74	9.55
Third codon	7.72	8.25	9.18	8.11
Total synonymous and non-synonymous changes	570	303	305	420

[a]More than one codon change in Arg, Leu or Ser.

Comparison of codon usages

Codon usages in the family C DNA polymerase genes are compared in Table 3. For comparison, codon frequencies in the class I *dna*E genes of *Aquifex* and *E.coli* and in the class II DNA pol C gene of *B.subtilis* are also included in Table 3. It is evident that some strong codon bias exists which is significantly different from *E.coli* and *B.subtilis*. The most dramatic example is the difference in usage of the two arginine codons, AGA and AGG. In both *T.maritima* class I and class II enzymes, AGA and AGG are the preferred codons, whereas CGT and CGC are the favored codons in *E.coli dna*E. In *Aquifex dna*E also, AGA and AGG codons are overwhelmingly favored. General trends in codon usage that have been noted by Kim *et al*. (60) are also evident in the Tma *dna*E and Tma DNA pol C genes. These include TCT (cysteine), CAG (glutamine), CAC (histidine), TTC (phenylalanine) and TAC (tyrosine).

DISCUSSION

We have shown that the hyperthermophilic bacterium *T.maritima* has two different classes of family C DNA polymerases. One is a homolog of the *E.coli dna*E protein, an α subunit of the DNA polymerase III holoenzyme (class I). The second one is a homolog of the *B.subtilis* DNA polymerase C (class II). Comparative sequence analyses among the class I family C DNA polymerases suggest that the Tma *dna*E protein may not be functional. The nucleotide substitution pattern also suggests that this might be the case. In contrast, the Tma DNA polymerase C appears to be the bona fide replicative DNA polymerase. Figure 4 shows a schematic comparison of the class I and class II family C DNA polymerases of *T.maritima* with those of other bacteria. Tma *dna*E is ~20% smaller than those of *E.coli* and *Aquifex*. In addition, Tma *dna*E has some internal sequence deletions (Fig. 1). It has been well established genetically as well as biochemically that the *B.subtilis* DNA pol C is absolutely essential for chromosome replication (51–54). Thus, the *B.subtilis* DNA pol C is a prototype of the class II family C DNA polymerases. As shown in Figure 4, Tma DNA pol C contains a conserved editing 3′→5′ exonuclease domain as well as a highly conserved DNA polymerase domain, including an HPura-binding site (HPUB). It has been generally accepted that class I family C DNA polymerase is the principal replicative DNA polymerase subunit in Gram-negative bacteria and that the class II family C DNA polymerase is the major replicative DNA polymerase of the Gram-positive bacteria (35). Our recent survey indicates that this is not the case (36). It now appears clear that only eubacterial members which belong to three bacterial phyla use class II family C DNA polymerase for chromosome replication. These include low G+C Gram-positive bacteria, mycoplasmas and *T.maritima* (this study). Surprisingly, most other bacteria, including high G+C Gram-positive bacteria such *as Actinomyces*, *Streptomyces*, *Mycobacterium* and *Corynebacterium* utilize class I family C DNA polymerases as their replicative enzymes. Cyanobacteria use class III family C DNA polymerases (36).

Table 3. Codon usage in class I and class II family C DNA polymerases

		Class II		Class I		
		Tma	Bsu	Tma	Aqu	Eco
Arg	AGA	36	17	32	16	0
	AGG	25	5	15	22	4
	CGT	6	18	3	0	36
	CGC	1	9	2	1	29
	CGA	3	2	4	0	1
	CGG	4	8	0	0	8
Ala	GCT	10	30	3	18	11
	GCC	31	17	18	15	25
	GCA	10	24	10	20	11
	GCG	26	23	10	21	55
Pro	CCT	10	17	12	7	7
	CCC	26	0	13	25	5
	CCA	13	7	10	4	8
	CCG	19	33	7	7	34
Asn	AAT	10	29	4	3	5
	AAC	38	22	29	42	27
Gln	CAA	3	30	5	7	12
	CAG	22	26	13	22	28
Lys	AAA	59	76	46	56	38
	AAG	61	30	21	81	20
Met	ATG	32	36	15	25	37
Leu	TTA	1	24	2	11	11
	TTG	11	20	9	3	13
	CTT	22	43	21	30	20
	CTC	59	11	30	56	12
	CTA	5	3	7	9	2
	CTG	36	32	21	17	64
Gly	GGT	29	19	16	15	31
	GGC	8	25	11	7	31
	GGA	40	31	19	43	10
	GGG	11	14	3	8	22
Thr	ACT	5	11	2	7	2
	ACC	16	12	9	20	29
	ACA	8	34	14	8	3
	ACG	27	22	10	18	11
Asp	GAT	47	68	24	18	43
	GAC	41	31	20	52	43
Glu	GAA	78	79	62	77	61
	GAG	55	44	16	42	31
Phe	TTT	15	40	13	17	24
	TTC	48	21	27	36	26
Trp	TGG	7	9	4	9	8
Ser	AGT	10	9	4	6	1
	AGC	12	10	9	7	15
	TCT	7	22	13	4	7
	TCC	18	13	7	11	8
	TCA	9	21	7	5	4
	TCG	9	3	5	4	11
Val	GTT	34	27	27	30	11
	GTC	23	20	15	15	20
	GTA	4	25	6	15	18
	GTG	45	17	25	20	26
Ile	ATT	8	61	10	11	12
	ATC	51	34	30	8	52
	ATA	43	11	18	48	2
Cys	TGT	6	9	1	3	5
	TGC	3	6	3	2	7
His	CAT	7	28	6	0	8
	CAC	17	15	11	20	16
Tyr	TAT	9	35	11	7	20
	TAC	37	19	22	50	19

Barnes *et al.* (45) were the first who observed that *Mycoplasma pulmonis* has a *B.subtilis* DNA pol III-type (class II) DNA polymerase and an additional DNA polymerase which is insensitive to HPura inhibitor. Subsequent complete genome sequence analyses of two mycoplasmas, *Mycoplasma genitalium* (46) and *Mycoplasma pneumoniae* (47), revealed that indeed these mycoplasma chromosomes contain a class II family C DNA polymerase and a class I family C DNA polymerase homolog. More recently, *B.subtilis* genome analyses showed that it contains in addition to the one class II family C DNA polymerase, two isoforms of the class I family C DNA polymerases (42). Interestingly, all these bacteria do not have family B (pol II) DNA polymerases. Instead, they contain multigenes for the family C DNA polymerases.

What can be said about the origin of the multigene of family C DNA polymerases? There are two simple explanations. One is that the class II family C DNA polymerases may be xenologous enzymes. Since only three bacterial phyla (low G+C Gram-positive bacteria, mycoplasmas and Thermotogales) contain class II family C DNA polymerases, it is possible that they arose by lateral gene transfer from another domain, perhaps from Archaea. However, this seems unlikely, because so far no family C DNA polymerase has been found in Archaea, with the possible exception of the hyperthermophilic, sulfate-reducing archaeon *Archaeoglobus fulgidus*, which has a homolog of the *dna*Q gene (33). The *dna*Q gene encodes the ε subunit of the *E.coli* DNA polymerase III holoenzyme (63). The second explanation is that the class I and class II gene variants of the family C DNA polymerases are the result of an early gene duplication. Koonin and Bork (64) favor the second view and suggested that an ancient duplication of the DNA pol III gene took place predating the divergence of Gram-positive and Gram-negative bacteria and that one of the copies in the Gram-negative lineage was lost during the course of evolution. If this is the case, then the class I and class

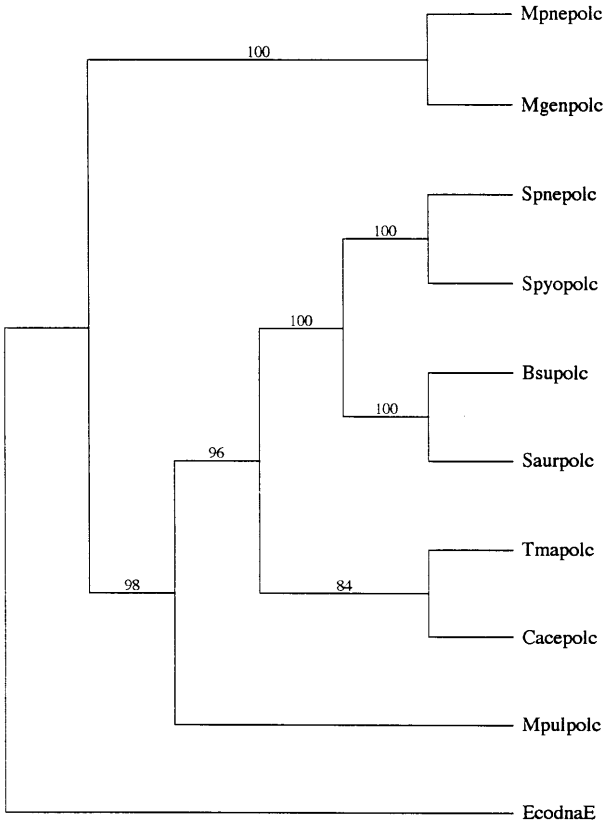


Figure 3. Phylogenetic tree based on class II family C DNA polymerase sequences. Neighbor joining consensus tree obtained with 250 bootstrap replicates. Protein names and sequence accession numbers are as follows: mpneupolc, *M.pneumoniae* DNA pol C (P75080); mgenpolc, *M.genitalium* DNA pol C (P47277); spneupolc, *S.pneumoniae* DNA pol C (*S.pneumoniae* pol C sequence data were obtained through early release from the Institute for Genomic Research, www.tigr.org at www.ncbi.nlm.nih.gov); spyopolc, *S.pyogenes* DNA pol C (*S.pyogenes* pol C sequence data were obtained through early release from the University of Okalahoma, http://dna1.chem.ou.edu/strep.html); bsupolc, *B.subtilis* DNA pol C (P13267); saurpolc, *S.aureus* DNA pol C (D45368); tmapolc, *T.maritima* DNA pol C (this study); cacepolc, *C.acetobutylicum* DNA pol C (*C.acetobutylicum* sequence data were obtained through the Internet, http://www.ncbi.nlm.nih.gov/BLAST/unfinishedgenome.html); Mpulpolc, *M.pulmonis* DNA pol C (P47729); ecodnaE, *E.coli dna*E (P10443). *E.coli dna*E protein was used as an outgroup.

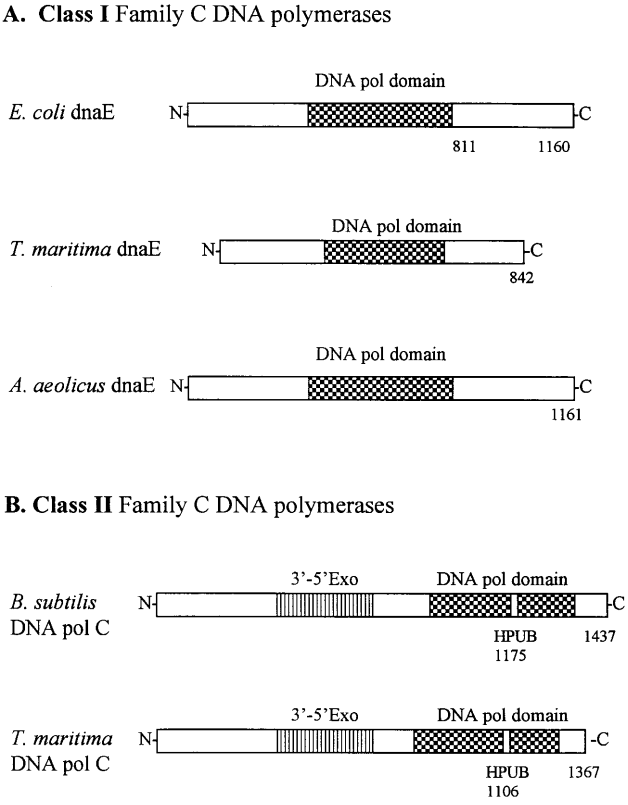


Figure 4. Schematic representation of domain structure of class I and class II family C DNA polymerases. DNA pol domains represent DNA polymerase catalytic regions. The N- and C-terminal regions are the sites of association with other subunits of the DNA polymerase III holoenzyme (61,62). 3'-5'Exo represents the editing 3′→5′ exonuclease domain. HPUB represents the HPura binding site (53). Numerals indicate total amino acid residues of each DNA polymerase

II family C DNA polymerases are paralogous proteins. An intriguing question would be the nature of the primordial family C DNA polymerase gene before its duplication into class I and class II genes. In this regard, we have previously suggested that the primordial family C DNA polymerase gene could have been like the present class II gene (49). Although a close evolutionary relationship between mycoplasmas and low G+C Gram-positive bacteria was not surprising, the even closer relationship between *T.maritima* and *Clostridium* was unexpected, because of their phylogenetic position on the 16S rRNA-based evolutionary tree. However, a linkage of *T.maritima* and the low G+C Gram-positive group of bacteria has also been observed for a number of other protein gene sequences, including those of glutamine synthetase (20–21), type I DNA topoisomerase (65), type II DNA topoisomerase (66), anthranilate synthetase component I (60), the σ factor σ^{70} family (26) and recA protein (25).

Evolutionary implications

The origin and evolution of hyperthermophilic microorganisms has been an intriguing question to the scientific community for many years. Two major hypotheses have been advanced (67). One is that the first cellular organism arose in a high temperature environment (hot origin of life theory). The second one is that life originated at moderate temperatures (mesophilic) and that the hyperthermophilic microorganisms arose by adaptation to a hot environment (adaptation theory). According to the current 'universal tree of life', primarily based on rRNA sequences, the deepest branches are occupied by hyperthermophiles (3,68). A number of geochemists who appreciate the abundance of hydrothermal systems on the early Earth and the potent chemical reactions possible in such systems, tend to support the hot origin of life theory (69). The 'universal tree of life' further suggests that Eubacteria rather than Archaea arose first on the Earth (3). Thus, the hyperthermophilic bacteria *Aquifex* and *Thermotoga* are generally regarded as the earliest and most slowly evolving microorganisms known today. Interestingly, these two microorganisms are very different. While *Aquifex* is a microaerophile, growing chemolitho-autotrophically by reducing oxygen and using hydrogen and sulfur as electron donors (70), *Thermotoga* is an anerobic,

chemoorganoheterotrophic microorganism which grows either on various sugars or on complex organic substrates as carbon and energy sources (4). We now know that *Aquifex* uses class I DNA polymerase for its chromosomal replication, but *Thermotoga* uses class II DNA polymerase. It is of crucial importance to determine accurately the phylogenetic positions of these hyperthermophilic microorganisms to assess the alternative theories, hot origin of life or adaptation. However, as Forteree (71) pointed out, the phylogenetic positions of these hyperthermophiles are currently controversial. Klenk *et al*. (19) were the first to show, using DNA-dependent RNA polymerase genes, that the hyperthermophilic bacterium *Aquifex pyrophilus* is linked closely to *E.coli* and *Pseudomonas putida*, both of which belong to the γ subdivision of Proteobacteria.. They suggested that the deep branching lineages of the hyperthermophiles based on 16S rRNA sequences could be erroneous because the G+C contents of rRNA differ drastically among microorganisms. Very recently, the complete genome of the hyperthermophilic bacterium *Aquifex aeolicus* has been determined (41). Because of the availability of >1500 open reading frames from this *Aquifex* lineage, it was expected to offer a definitive resolution of the phylogenetic position. However, analyses of various protein coding genes did not yield a statistically significant placement of the *Aquifex* lineage (41). In fact, the genome sequences have been described as proving to be more confusing than enlightening (72).

Our finding that the *T.maritima* family C DNA polymerase (class II) is closely related to those of low G+C Gram-positive bacteria and mycoplasmas also contrasts strikingly with the 16S rRNA-based phylogeny (3). As described above, our results are in good agreement with those of a number of protein gene sequences (20,21,25–27,45,60,73). In an extensive analysis comparing phylogenetic results of the σ factor proteins with small subunit rRNA data, Gruber and Bryant (27) showed that despite the overall branching patterns and resolution of the two markers being very similar, the *T.maritima* σ factor gene was most closely related to the mycoplasma *M.genitalium*. Similarly, the *T.maritima* recA protein gene was placed in between low and high G+C Gram-positive bacteria (26). Interestingly, Cavalier-Smith had suggested previously that *Thermotaga* is close to Gram-positive bacteria because it does not have an outer membrane (74). Together, this phylogenetic evidence is compelling and appears to refute the idea that all these genes which contradict the rRNA-based phylogeny arose by horizontal evolution. Also, this evidence is not compatible with the notion of a hot origin of life. However, it is of paramount importance to keep in mind that gene trees do not necessarily represent organismal phylogenies. If *T.maritima* represents one of the earliest branches of the bacterial tree, then one must consider the antiquity of low G+C Gram-positive bacteria as well, in particular the clostridia.

ACKNOWLEDGEMENTS

We are very grateful to Dr Robert Hubber (Universität Regensburg, Germany) for his generous gift of *T.maritima* DNA. We are also grateful to Dr David L. Swofford for allowing us to perform analyses with the PAUP* 4d64 program and publish the results. We thank the Institute for Genomic Research for availability of *Streptococcus pneumoniae* pol C sequence data prior to publication, the Streptococcal Genome Sequencing team at the University of Okalahoma Health Science Center and the *C.acetobutylicum* Genome Sequencing team at Genome Therapeutic Corporation. We would like to thank Harris Bernstein, James W. Moulder and Vivian Gage for their critical reading of the manuscript. This research was supported by a grant from the NIH (GM28013).

REFERENCES

1 Huber,R. and Stetter,K.O. (1992) In Kristjansson,J.K. (ed.), *Thermophilic Bacteria*. CRC Press, Boca Raton, FL, pp. 185–194.
2 Stetter,K.O. (1998) *FEMS Microbiol. Rev.*, **18**, 149–158.
3 Woese,C.R., Kandler,O. and Wheelis,M.L. (1990) *Proc. Natl Acad. Sci. USA*, **87**, 4576–4579.
4 Huber,R., Langworthy,T.A., König,H., Thomm,M., Woese,C.R., Sleytr,U.B. and Stetter,K.O. (1986) *Arch. Microbiol.*, **144**, 324–333.
5 Sanangelantoni,A.M., Bocchetta,M., Cammarano,P. and Tiboni,O. (1994) *J. Bacteriol.*, **176**, 7703–7710.
6 Londei,P., Altamura,S., Huber,R., Stetter,K.O. and Cammarano,P. (1988) *J. Bacteriol.*, **170**, 4353–4360.
7 De Rose,M., Gambacorta,A., Huber,R., Lanzotti,D., Nicholaus,B., Stetter,K.O. and Trincone,A. (1988) *J. Chem. Soc. Chem. Commun.*, **13**, 1300.
8 Achenbach-Richter,L., Gupta,R.S., Stetter,K.O. and Woese,C.R. (1987) *Syst. Appl. Microbiol.*, **9**, 34–39.
9 Bachleitner,M., Ludwig,W., Stetter,K.O. and Schleifer,K.H. (1989) *FEMS Microbiol. Lett.*, **48**, 115–120.
10 Woese,C.R. (1994) *Proc. Natl Acad. Sci. USA*, **91**, 1601–1603.
11 Baldauf,S.L., Palmer,J.D. and Doolittle,W.F. (1996) *Proc. Natl Acad. Sci. USA*, **93**, 7749–7754.
12 Pace,N.R. (1997) *Science*, **276**, 734–740.
13 Pace,N.R. (1991) *Cell*, **65**, 531–533.
14 Ernst,W.G. (1983) In Schopf,J.W. (ed.), *Earth's Earliest Biosphere*. Princeton University Press, Princeton, NJ, pp. 41–52.
15 DeLong,E.F. (1992) *Proc. Natl Acad. Sci. USA*, **89**, 5685–5689.
16 Schleper,C., Swanson,R.V., Mathur,E.J. and DeLong,E.F. (1997) *J. Bacteriol.*, **179**, 7803–7811.
17 Olsen,G.J., Woese,C.R. and Overbeek,R. (1994) *J. Bacteriol.*, **176**, 1–6.
18 Palm,P., Schleper,C., Arnold-Ammer,I., Holz,I., Meier,T., Lottspeich,F. and Zillig,W. (1993) *Nucleic Acids Res.*, **21**, 4904–4908.
19 Klenk,H.-P., Palm,P. and Zillig,W. (1994) *Syst. Appl. Microbiol.*, **16**, 638–647.
20 Tiboni,O., Cammarano,P. and Sanangelantoni,A.M. (1993) *J. Bacteriol.*, **175**, 2961–2969.
21 Brown,J.R., Masuchi,Y., Robb,F.T. and Doolittle,W.F. (1994) *J. Mol. Evol.*, **38**, 566–576.
22 Pesole,G., Gissi,C., Lanave,C. and Saccone,C. (1995) *Mol. Biol. Evol.*, **12**, 189–197.
23 Saccone,C., Gissi,C., Lanave,C. and Pesole,G. (1995) *J. Mol. Evol.*, **40**, 273–279.
24 Benachenhou-Lahfa,N., Forterre,P. and Labedan,B. (1993) *J. Mol. Evol.*, **36**, 335–346.
25 Gupta,R.S., Bustard,K., Falah,M. and Singh,D. (1997) *J. Bacteriol.*, **179**, 345–357.
26 Eisen,J.A. (1995) *J. Mol. Evol.*, **41**, 1105–1123.
27 Gruber,T.M. and Bryant,D.A. (1997) *J. Bacteriol.*, **179**, 1734–1747.
28 Bost,D.A., Stoffel,S., Landre,P., Lawyer,F.C., Akers,J., Abramson,R.D. and Gelfand,D.H. (1994) *FASEB J.*, **8**, A1395.
29 Blamey,J.M., Mukund,S. and Adams,M.W. (1994) *FEMS Microbiol. Lett.*, **121**, 165–169.
30 Braithwaite,D.K. and Ito,J. (1993) *Nucleic Acids Res.*, **21**, 787–802.
31 Bult,C.J., White,O., Olsen,G.J., Zhou,L., Fleischmann,R.D., Sutton,G.G., Blake,J.A., FitzGerald,L.M., Clayton,R.A., Gocayne,J.D., Kerlavage,A.R., Dougherty,B.A., Tomb,J.F., Adams,M.D., Reich,C.I., Overbeek,R., Kirkness,E.F., Weinstock,K.G., Merrick,J.M., Glodek,A., Scott,J.L., Geoghagen,N.S.M. and Venter,J.C. (1996) *Science*, **273**, 1058–1073.
32 Smith,D.R., Doucette-Stamm,L.A., Deloughery,C., Lee,H., Dubois,J., Aldredge,T., Bashirzadeh,R., Blakely,D., Cook,R., Gilbert,K., Harrison,D., Hoang,L., Keagle,P., Lumm,W., Pothier,B., Qiu,D., Spadafora,R., Vicaire,R., Wang,Y., Wierzbowski,J., Gibson,R., Jiwani,N., Caruso,A., Bush,D. and Reeve,J.N. (1997) *J. Bacteriol.*, **179**, 7135–7155.
33 Klenk,H.P., Clayton,R.A., Tomb,J.F., White,O., Nelson,K.E., Ketchum,K.A., Dodson,R.J., Gwinn,M., Hickey,E.K., Peterson,J.D., Richardson,D.L., Kerlavage,A.R., Graham,D.E., Kyrpides,N.C., Fleischmann,R.D., Quackenbush,J., Lee,N.H., Sutton,G.G., Gill,S., Kirkness,E.F., Dougherty,B.A., McKenney,K., Adams,M.D., Loftus,B. and Venter,J.C. (1997) *Nature*, **390**, 364–370.

34 Perler,F.B., Kumar,S. and Kong,H. (1996) *Adv. Protein Chem.*, **48**, 377–435.
35 Kornberg,A. and Baker,T. (1992) *DNA Replication*, 2nd Edn. Freeman, San Francisco, CA.
36 Huang,Y.-P. and Ito,J. (1998) *J. Mol. Evol*., submitted for publication.
37 Sanger,F., Nicklen,S. and Coulson,A.R. (1977) *Proc. Natl Acad. Sci. USA*, **74**, 5463–5467.
38 Devereux,J., Haeberli,P. and Smithies,O. (1984) *Nucleic Acids Res*., **12**, 387–395.
39 Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) *Nucleic Acids Res*., **22**, 4673–4680.
40 Swofford,D.L. (1998) PAUP* 4.0 beta version. Sinauer Associates, Sunderland, MA.
41 Deckert,G., Warren,P.V., Gaasterland,T., Young,W.G., Lenox,A.L., Graham,D.E., Overbeek,R., Snead,M.A., Keller,M., Aujay,M., Huber,R., Feldman,R.A., Short,J.M., Olsen,G.J. and Swanson,R.V. (1998) *Nature*, **392**, 353–358.
42 Kunst,F., Ogasawara,N., Moszer,I., Albertini,A.M., Alloni,G., Azevedo,V., Bertero,M.G., Bessieres,P., Bolotin,A., Borchert,S., Borriss,R., Boursier,L., Brans,A., Braun,M., Brignell,S.C., Bron,S., Brouillet,S., Bruschi,C.V., Caldwell,B., Capuano,V., Carter,N.M., Choi,S.K., Codani,J.J., Connerton,I.F. and Danchin,A. (1997) *Nature*, **390**, 249–256.
43 Tomasiewicz,H.G. and McHenry,C.S. (1987) *J. Bacteriol*., **169**, 5735–5744.
44 Pacitti,D.F., Barnes,M.H., Li,D.H. and Brown,N.C. (1995) *Gene*, **165**, 51–56.
45 Barnes,M.H., Tarantino,P.M.J., Spacciapoli,P., Brown,N.C., Yu,H. and Dybvig,K. (1994) *Mol. Microbiol*., **13**, 843–854.
46 Fraser,C.M., Gocayne,J.D., White,O., Adams,M.D., Clayton,R.A., Fleischmann,R.D., Bult,C.J., Kerlavage,A.R., Sutton,G. and Kelley,J.M. (1995) *Science*, **270**, 397–403.
47 Himmelreich,R., Hilbert,H., Plagens,H., Pirkl,E., Li,B.C. and Herrmann,R. (1996) *Nucleic Acids Res*., **24**, 4420–4449.
48 Barnes,M.H., Spacciapoli,P., Li,D.H. and Brown,N.C. (1995) *Gene*, **165**, 45–50.
49 Huang,Y., Braithwaite,D.K. and Ito,J. (1997) *FEBS Lett*., **400**, 94–98.
50 Ito,J. and Braithwaite,D.K. (1998) *Mol. Microbiol*., **27**, 235–236.
51 Neville,M.M. and Brown,N.C. (1972) *Nature New Biol*., **240**, 80–82.
52 Brown,N.C. (1970) *Proc. Natl Acad. Sci. USA*, **67**, 1454–1461.
53 Barnes,M.H., Hammond,R.A., Foster,K.A., Mitchener,J.A. and Brown,N.C. (1989) *Gene*, **85**, 177–186.
54 Hammond,R.A., Barnes,M.H., Mack,S.L., Mitchener,J.A. and Brown,N.C. (1991) *Gene*, **98**, 29–36.
55 Saitou,N. and Nei,M. (1987) *Mol. Biol. Evol*., **4**, 406–425.
56 Woese,C.R., Stackebrandt,E. and Ludwig,W. (1984) *J. Mol. Evol*., **21**, 305–316.
57 Weisburg,W.G., Dobson,M.E., Samuel,J.E., Dasch,G.A., Mallavia,L.P., Baca,O., Mandelco,L., Sechrest,J.E., Weiss,E. and Woese,C.R. (1989) *J. Bacteriol*., **171**, 4202–4206.
58 Woese,C.R. (1987) *Microbiol. Rev*., **51**, 221–271.
59 Razin,S. (1978) *Microbiol. Rev*., **42**, 414–470.
60 Kim,C.W., Markiewicz,P., Lee,J.J., Schierle,C.F. and Miller,J.H. (1993) *J. Mol. Biol*., **231**, 960–981.
61 Kim,D.R., Pritchard,A.E. and McHenry,C.S. (1997) *J. Bacteriol*., **179**, 6721–6728.
62 Fijalkowska,I.J. and Schaaper,R.M. (1993) *Genetics*, **134**, 1039–1044.
63 Scheuermann,R.H. and Echols,H. (1984) *Proc. Natl Acad. Sci. USA*, **81**, 7747–7751.
64 Koonin,E.V. and Bork,P. (1996) *Trends Biochem. Sci*., **21**, 128–129.
65 Bouthier,d.l.T., Kaltoum,H., Portemer,C., Confalonieri,F., Huber,R. and Duguet,M. (1995) *Biochim. Biophys. Acta*, **1264**, 279–283.
66 Guipaud,O., Labedan,B. and Forterre,P. (1996) *Gene*, **174**, 121–128.
67 Brock,T.D. (1986) In Brock,T.D. (ed.), *Thermophiles: General, Molecular and Applied Microbiology*. John Wiley & Sons, New York, NY, pp. 1–16.
68 Stetter,K.O. (1996) In *Evolution of Hydrothermal Ecosystems on Earth (and Mars?)*, Ciba Foundation Symposium 202. John Wiley & Sons, Chichester, UK, pp. 1–18.
69 Shock,E.L. (1996) In *Evolution of Hydrothermal Ecosystems on Earth (and Mars?)*, Ciba Foundation Symposium 202. John Wiley & Sons, Chichester, UK, pp. 40–60.
70 Huber,R., Wilharm,T., Huber,D., Trincone,A., Burggraf,S., König,H., Rochel,R., Rochinga,I., Fricke,H. and Stetter,K.O. (1992) *Syst. Appl. Microbiol*., **15**, 349–351.
71 Forterre,P. (1996) *Cell*, **85**, 789–792.
72 Pennisi,E. (1998) *Science*, **280**, 672–674.
73 Brown,J.R. and Doolittle,W.F. (1997) *Microbiol. Mol. Biol. Rev*., **61**, 456–502.
74 Cavalier-Smith,T. (1992) *Secondary Metabolites; Their Function and Evolution*, Ciba Foundation Symposium 171. John Wiley & Sons, Chichester, UK, pp. 64–87.