J Mol Evol (1999) 49:524–537

JOURNAL OF MOLECULAR EVOLUTION

The Archaea Monophyly Issue: A Phylogeny of Translational Elongation Factor G(2) Sequences Inferred from an Optimized Selection of Alignment Positions

Piero Cammarano,[1] Roberta Creti,[1] Anna M. Sanangelantoni,[2] Peter Palm[3]

[1] Istituto Pasteur Fondazione Cenci-Bolognetti, Dipartimento di Biotecnologie cellulari ed Ematologia, Sezione di Genetica molecolare, Universita' di Roma I "La Sapienza," Policlinico Umberto I, Viale Regina Elena 324, 00161 Roma, Italy
[2] Dipartimento di Genetica e Microbiologia A. Buzzati-Traverso, Universita' di Pavia, via Abbiategrasso 207, 27100 Pavia, Italy
[3] Max Planck Institut für Biochemie, D-8033 Martinsried, Germany

Abstract. A global alignment of EF-G(2) sequences was corrected by reference to protein structure. The selection of characters eligible for construction of phylogenetic trees was optimized by searching for regions arising from the artifactual matching of sequence segments unique to different phylogenetic domains. The spurious matchings were identified by comparing all sections of the global alignment with a comprehensive inventory of significant binary alignments obtained by BLAST probing of the DNA and protein databases with representative EF-G(2) sequences. In three discrete alignment blocks (one in domain II and two in domain IV), the alignment of the bacterial sequences with those of Archaea–Eucarya was not retrieved by database probing with EF-G(2) sequences, and no EF-G homologue of the EF-2 sequence segments was detected by using partial EF-G(2) sequences as probes in BLAST/FASTA searches. The two domain IV regions (one of which comprises the ADP-ribosylatable site of EF-2) are almost certainly due to the artifactual alignment of insertion segments that are unique to Bacteria and to Archaea–Eucarya. Phylogenetic trees have been constructed from the global alignment after deselecting positions encompassing the unretrieved, spuriously aligned regions, as well as positions arising from misalignment of the G′ and G″ subdomain insertion segments flanking the "fifth" consensus motif of the G domain (Ævarsson, 1995). The results show inconsistencies between trees inferred by alternative methods and alternative (DNA and protein) data sets with regard to Archaea being a monophyletic or paraphyletic grouping. Both maximum-likelihood and maximum-parsimony methods do not allow discrimination (by log-likelihood difference and difference in number of inferred substitutions) between the conflicting (monophyletic vs. paraphyletic Archaea) topologies. No specific EF-2 insertions (or terminal accretions) supporting a crenarchaeal–eucaryal clade are detectable in the new EF-G(2) sequence alignment.

Key words: Elongation factor G(2) — BLAST/FASTA — Sequence alignment — Insertion elements — Phylogeny — Archaea — Protein evolution

Introduction

Protein synthesis elongation factors G (EF-G) and Tu (EF-Tu) (called EF-2 and EF-1α, respectively, in Archaea and Eucarya) are paralogous GTP-binding proteins that arose from gene duplication prior to the divergence of the three major lineages, and both proteins have been used to construct unrooted (Creti et al. 1994) and rooted (Iwabe et al. 1989) global phylogenetic trees.

Correspondence to: P. Cammarano; *e-mail:* cammarano@bce.med.uniromal.it

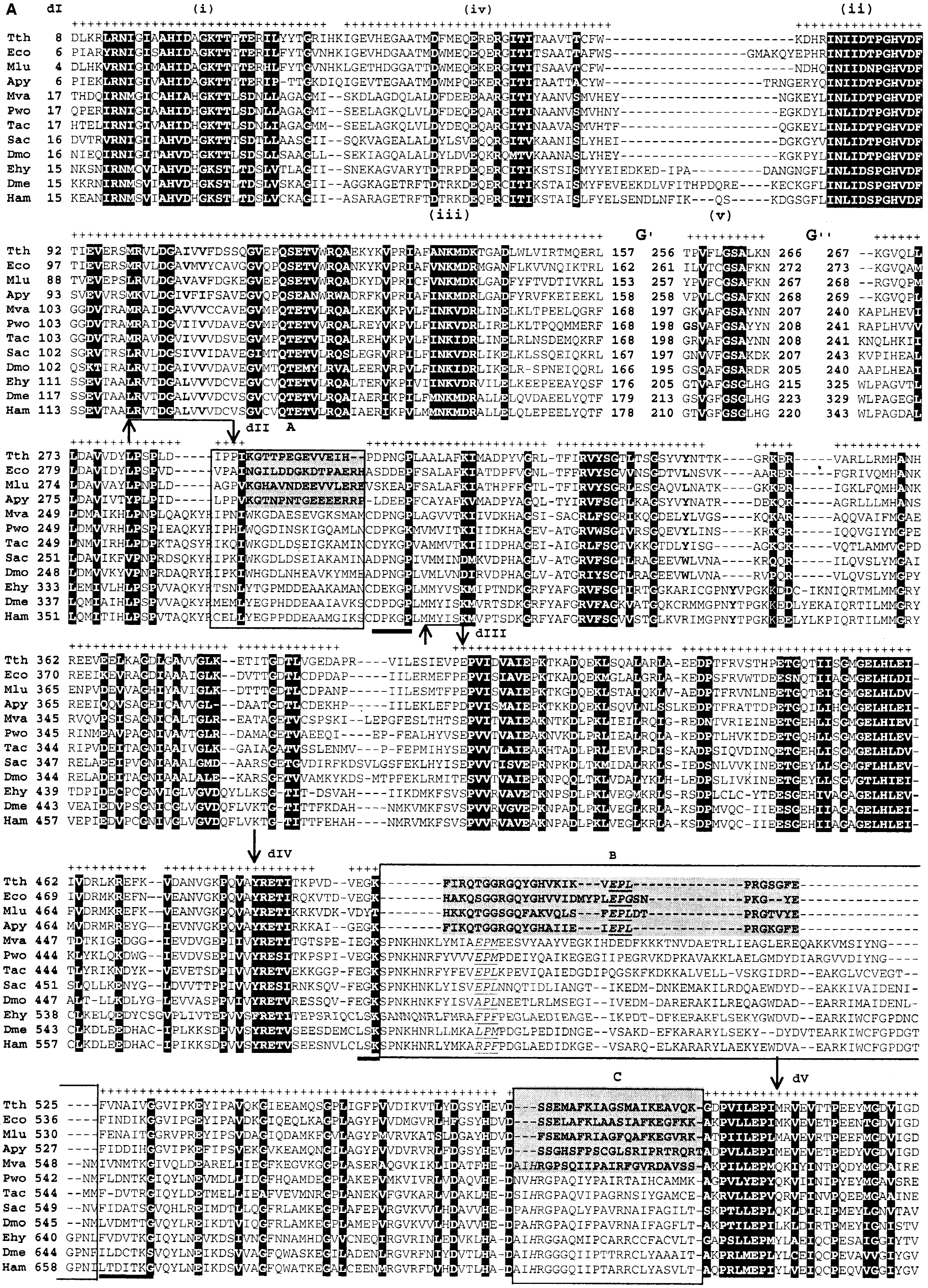
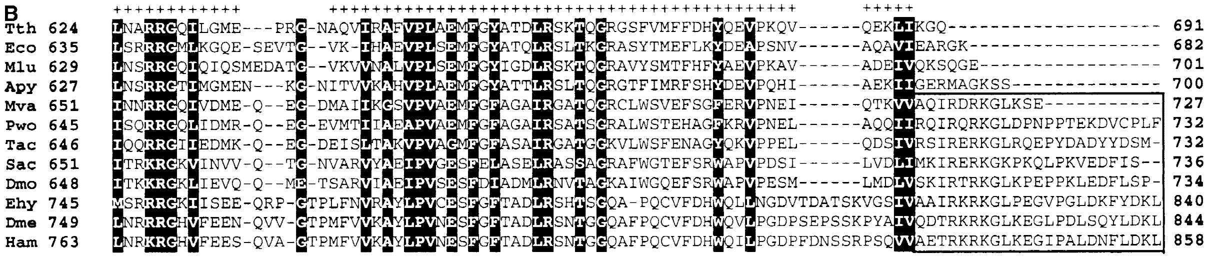


Fig. 1. See legend on p. 526.



```
B           ++++++++++++          ++++++++++++++++++++++++++++++++++++++++++          +++++
Tth 624  LNARRGQILGME---PRG-NAQVIRAFVPLAEMFGYATDLRSKTQGRGSFVMFFDHYQEVPKQV------QEKLIKGQ-------------------  691
Eco 635  LSRRRGMLKGQE-SEVTG--VK-IHAEVPLSEMFGYATQLRSLTKGRASYTMEFLKYDEAPSNV------AQAVIEARGK-----------------  682
Mlu 629  LNSRRGQIQIQSMEDATG--VKVVNALVPLSEMFGYIGDLRSKTQGRAVYSMTFHFYAEVPKAV------ADEIVQKSQGE----------------  701
Apy 627  LNSRRGTIMGMEN---KG-NITVVKAHVPLAEMFGYATTLRSLTQGRGTFIMRFSHYDEVPQHI------AEKIIGERMAGKSS-------------  700
Mva 651  INNRRGQIVDME-Q--EG-DMAIIKGSVPVAEMFGFAGAIRGATQGRCLWSVEFSGFERVPNEI------QTKVVAQIRDRKGLKSE----------  727
Pwo 645  ISQRRGQLIDMR-Q--EG-EVMTIIAEAPVAEMFGFAGAIRSATSGRALWSTEHAGFKRVPNEL------AQQIIRQIRQRKGLDPNPPTEKDVCPLF  732
Tac 646  IQQRRGIIEDMK-Q--EG-DEISLTAKVPVAGMFGFASAIRGATGGKVLWSFENAGYQKVPPEL------QDSIVRSIRERKGLRQEPYDADYYDSM-  732
Sac 651  ITRKRGKVINVV-Q--TG-NVARVYAEIPVGESFELASELRASSAGRAFWGTEFSRWAPVPDSI------LVDLIMKIRERKGKPKQLPKVEDFIS--  736
Dmo 648  ITKKRGKLIEVQ-Q--ME-TSARVIAEIPVSESFDIADMLRNVTAGKAIWGQEFSRWAPVPESM------LMDLVSKIRTRKGLKPEPPKLEDFLSP-  734
Ehy 745  MSRRRGKIISEE-QRP-GTPLFNVRAYLPVCESFGFTADLRSHTSGQA-PQCVFDHWQLLNGDVTDATSKVGSIVAAIRKRKGLPEGVPGLDKFYDKL  840
Dme 749  LNRRRGHVFEEN-QVV-GTPMFVVKAYLPVNESFGFTADLRSNTGGQAFPQCVFDHWQVLPGDPSEPSSKPYAIVQDTRKRKGLKEGLPDLSQYLDKL  844
Ham 763  LNRKRGHVFEES-QVA-GTPMFVVKAYLPVNESFGFTADLRSNTGGQAFPQCVFDHWQILPGDPFDNSSRPSQVVAETRKRKGLKEGIPALDNFLDKL  858
```

Fig. 1. Aligned predicted EF-G(2) sequences from the three phylogenetic domains. The five EF-G structural domains (Ævarsson et al. 1994) are numbered consecutively by *uppercase roman numerals* (dl to dV) along the *T. thermophilus* sequence. *Arrows* indicate starts (↓) and ends (↑) of structural domains; $\overline{\uparrow \quad \downarrow}$ delimits sequence elements that are not assigned to either one or the other of two neighboring domains (*T. thermophilus* residues 323–335). Only 12 sequences are shown for reasons of space. *Boldface characters* indicate sites occupied by identical or similar amino acids (ILVM, DEKRH, ST, GA, FYW, NQ) in no less than 80% of the aligned sequences. The positions of insertion sequences constituting the G′ and G″ subdomains (Ævarsson 1995) are indicated. *Lowercase roman numerals* (i–v) indicate the regions comprising the four consensus motifs of the G domain that are common to all of the translational GTPases (consensus motifs I–III and the RGITI sequence) and the EF-G(2) variant (VXXGS[G,A]) of the fifth consensus motif; the fourth element ([L,K] of the general fifth consensus motif (GSA[L,K]) proposed by Ævarsson (1995) is not confirmed by the alignment. The alignment of sequences comprising structural domain 1 differs from that of Ævarsson (1995) in the introduction, in the present alignment, of a single gap in the EF-G sequences (between *T. thermophilus* residues 48 and 49), which generates a universally conserved glycine at position 48 of the *T. thermophilus* sequence. The EFG(2) version of the consensus motif for domain II of the translational GTPases (GX[L,I,V,F][Y,F,*del*]XXXR[L,V,I] [F,W,Y]SGX [L,I,V]) spans *Thermus thermophilus* residues 323–335). *Underlined* sites in domains II and IV are (i) a consensus sequence [N,K,D,E-][G,A,E]P (*T. thermophilus* residues 304–308), (ii) variants of a dominant EGK theme (*Thermus* residues 494–496), (iii) the motif [F,I,L,M-,V]X[ND]X[I,T]XG that delimits C-terminally a putative archaeal–eucaryal insert in box B (*T. thermophilus* residues 525–531). *Plus signs* indicate characters selected for phylogenetic analysis. *Numbers* indicate amino acid sequence positions. *Underlined characters* in the *boxed* B region indicate the archaeal–eucaryal and the bacterial proline-containing element that have been matched in Fig. 2B. The histidine which is ADP-ribosylatable by the diphtheria toxin reaction (Kessel and Klink 1980; Lechner et al. 1988) is given in *italics* in the *boxed* region C. Abbreviations and sources of sequences: Tth (*Thermus thermophilus,* P13551); Eco (*Escherichia coli,* P02996); Mlu (*Micrococcus luteus,* P09952); Apy (*Aquifex pyrophilus,* X74277); Mva (*Methanococcus vannielii,* P09604); Pwo (*Pyrococcus woesei,* P29050); Tac (*Thermoplasma acidophilum,* P26752); Sac (*Sulfolobus acidocaldarius,* P23112); 4B7 (uncultivated planktonic marine Archaeon, UA41261); Ehy (*Entamoeba histolytica,* QO6193); Dme (*Drosophila melanogaster,* P13060); Ham (hamster, U17362). The following sequences used to optimize the alignment are not shown: Ani (*Anacystis nidulans,* P18667); Ata (*Arabidopsis thaliana,* T43083); Atu, (*Agrobacter tumefaciens,* X99673); Bbu (*Borrelia burgdorferi,* AF021260); Bho (*Blastocystis hominis,* Q17152); Bsu (*Bacillus subtilis,* P80868); Bvu (*Beta vulgaris,* Z97178); Cel (*Caenorhabditis elegans,* P29691); Cke (*Chlorella kessleri,* P28996); Cpr (*Cryptosporidium parvum,* U21667); Ddi (*Dictyostelium discoideum,* P15112); Dmo (*Desulfurococcus mobilis,* P33159); Ecr (*Eikenella corrodens,* Z12610); Gga (*Gallus gallus,* Q90705); Gla (*Giardia lamblia.* D29835); Gpe (*Glugea plecoglossi,* D79220); Hha (*Halobacterium halobium,* P14823); Hin (*Haemophilus influenzae,* P43925); *Homo* (Hsa, X51446); Hpy (*Helicobacter pylori,* P56002); Mca (*Mycoplasma capricolum,* M96588); Mge (*Mycoplasma genitalium,* P47335); Mle (*Mycobacterium leprae,* P30767); Mpn (*Mycoplasma pneumoniae,* P75544); Mja (*Methanococcus jannaschii,* Q58448); Mmu (*Mus musculus,* J03200); Mtu (*Mycobacterium tuberculosis,* Z84395); Ngo (*Neisseria gonorrhaeae,* L36380); Osa (*Oryza sativa,* C26224); Pfa (*Plasmodium falciparum,* T02597); Pro (*Planobispora rosea,* P72230); rat (Q0780); Rpr (*Rickettsia prowazecki,* P41084); Sce (*Saccharomyces cerevisiae,* P32324); Spl (*Spirulina platensis,* P13550); Sra (*Streptomyces racemosissimus,* X67057); Sty (*Salmonella tiphymurium,* P26229); Tcr (*Trypanosoma cruzi,* D50806); Sso (*Sulfolobus solfataricus,* P30925); Syn (*Synechocystis* sp., PCC6803, P74228); Tma (*Thermotoga maritima,* P38525).

The two sets of factors typically harbor the three consensus motifs ([G,A]XXXGK[T,S], DXXG, NKXD) characteristic of the G superfamily (Dever et al. 1987), a consensus sequence RGITI (situated between motif I and motif II), a functionally important "fifth" consensus motif (GSA[L,K]) which is C-terminal to motif III (Bourne et al. 1990, 1991; Kjeldgaard and Nyborg 1992), and a consensus motif for domain II (Ævarsson, 1995).

According to a structure-guided alignment of EF-G(2) with other GTPases involved in translation (Ævarsson, 1995), EF-G and archaeal and eucaryal EF2s differ strikingly in the extension of two insertion regions (termed G′ and G″ subdomains) that are immediately N-terminal and C-terminal, respectively, to the fifth consensus element (GSA[L,K]). The G′ subdomain insertion spans up to 120 residues in Bacteria and only about 30 residues in Archaea and Eucarya. In contrast, the G″ subdomain is unique to Eucarya (up to 110 residues) and Archaea (only 35 residues). As one would expect for putative insertions, no recognizable similarity exists between the archaeal–eucaryal and the bacterial sequences in the G′ subdomain or between the archaeal and the eucaryal sequences in the G″ subdomain. Because these sequences are not common to all three major taxa, and are not ancestrally related entities, they are not eligible in principle for the construction of global phylogenies.

Most important, however, multiple alignment algorithms consistently associate the large insert of EF-G, the G′ subdomain, with the (unrelated) G′ subdomain sequences of Archaea and Eucarya as well as with elements of the G″ insertion (Ævarsson, 1995). Because positions comprising this block of unrelated sequences were selected for the construction of EF-G(2)-based phylogenies, their effect on the topology of the inferred trees has been analyzed with regard to Archaea being a monophyletic (Cammarano et al. 1992; Creti et al. 1994) or a paraphyletic (Rivera and Lake 1992; Hashimoto and Hasegawa 1996; Baldauf et al. 1997) grouping, and evidence has been sought for blocks of spurious homology beyond the G′ and G″ subdomains that could affect the "archaeal branch" of the EF-G(2) tree.

Here we report the detection of additional blocks of spuriously aligned bacterial and archaeal-eucaryal EF sequences, and demonstrate that deselecting positions corresponding to these blocks and to the G′ and G″ subdomains (Ævarsson, 1995) affects the robustness of the archaeal branch of the tree. The new alignment does not give any significant preference to either monophyly or paraphyly of the Archaea, as the two alternatives cannot be significantly discriminated by maximum-likelihood and maximum-parsimony methods.

Methods

Databank sequence retrievals and BLAST (Altschul et al. 1990) and FASTA (Pearson et al. 1988) probing of the DNA and protein databases were performed with the tBLASTN and FASTAp programs using the GCG program suite (Genetic Computer Group) (Deveraux et al. 1984) of the UK MRC Human Genome Mapping Project (HGMP) Resource Centre (Cambridge University, Cambridge, UK); FASTAp searches of the protein databases used gap creation and gap extension penalties of 12.0 and 4.0, respectively. Preliminary multiple alignments of amino acid sequences were generated with the programs Multalin (Corpet, 1988) and Clustal W (Thompson et al. 1994) using default gap penalties. Conversion of aminoacid sequence alignments into colinear alignments of nucleotide sequences (first *plus* second codon positions) used programs compiled by P. Boccardi (unpublished). Unrooted phylogenetic trees were constructed using the programs CONSENSE, DNADIST, DNAML, DNAPARS, FITCH, KITSCH, PROTDIST, PROTPARS, and SEQBOOT implemented in the Phylogeny Inference Package (PHYLIP), version 3.57 c (Felsenstein, 1989). Transition (TI)-to-transversion (TV) rate ratios for all pairs of nucleotide sequences compared [*R* parameter (Kumar et al. 1993)] were calculated by the program implemented in the package MEGA (Molecular Evolutionary Genetic Analysis), version 1.01 (Kumar et al. 1993), assuming a Kimura two-parameter model of nucleotide substitutions. Maximum-likelihood analyses utilized the NucML and ProtML programs of the MOLPHY (Molecular Phylogenetics) package, version 2.2 (Adachi and Hasegawa 1992). All ProtML analyses used the Jones–Taylor–Thornton (JTT) substitution matrix and the NucML analyses used the *R* parameter calculated as specified above; in all cases 1000 candidate topologies (of 2,027,025) were selected by the approximate log-likelihood criterion (Adachi 1995; Waddel 1995) from an exhaustive search of a partially constrained starting tree comprising 20 OTUs (operational taxonomic units) organized in 10 topological elements. The retained 1000 top-ranking topologies were then analyzed for the best tree by the RELL (resampling of estimated log-likelihood) bootstrap method with the "users" option of the NucML and ProtML programs (Kishino and Hasegawa 1989; Kishino et al. 1990).

Results and Discussion

Sequence Alignment

Figure 1 shows an updated alignment of EF-G(2) sequences initially obtained by standard algorithms and progressively optimized by addition of new species and incorporation of structural information (Ævarsson et al. 1994; Ævarsson, 1995). Up to residue 400 of the *Thermus thermophilus* EF-G sequence (domains dI and dII), the alignment in Fig. 1 is identical to the structure-guided alignment of Ævarsson (1995) except for a single position (see Fig. 1 legend), while beyond residue 400 (EF-G domains dIII–dV) the sequences were aligned by visually matching obvious signature sequences constraining the alignment topology (boldface characters in Fig. 1).

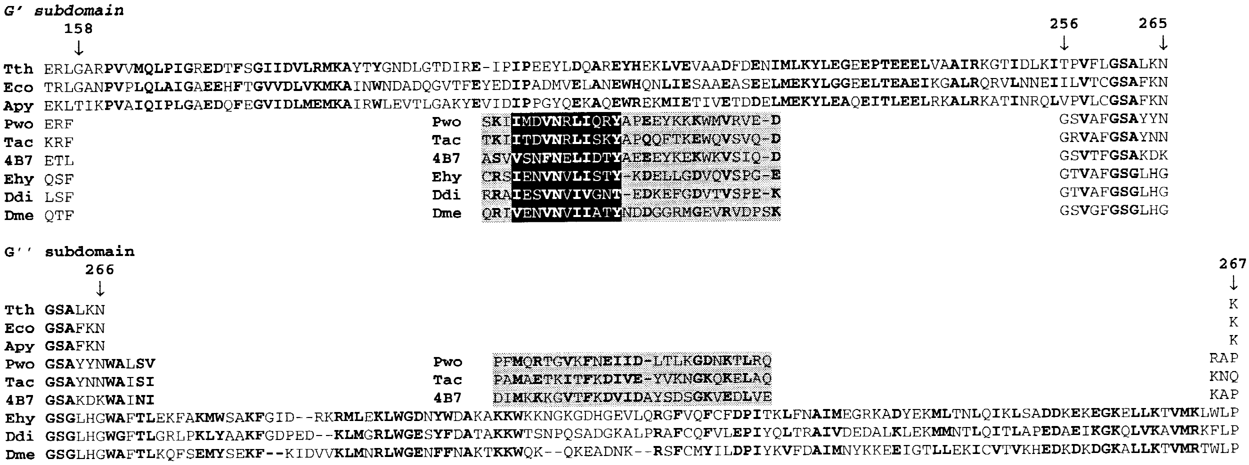
Displayed separately (Fig. 2) are the region of the multiple alignment encompassing the G′ subdomain, which is basically unique to Bacteria (97 residues in *T. thermophilus* and only 28–31 residues in Archaea and Eucarya), and the G″ subdomain, which is essentially unique to Eucarya (up to 123 residues in mammals but only 24–27 residues in Archaea) (see Ævarsson 1995). The archaeal and eucaryal sequences comprising the G′ subdomain are unrelated (by obvious signatures) to any regions of the bacterial sequences spanning the same structural space; however, they have similar lengths and are linked to one another by an obvious consensus motif, [I,V]XXVNX[I,L][I,V]XX[Y,M] (highlighted region in Fig. 2). This would be expected if the sequences constituting the G′ subdomain arose by insertion after the divergence of the bacterial and the archaeal–eucaryal lineages. In contrast, no apparent relatedness exists between the short archaeal G″ subdomain and any sequences of the longer G″ subdomain of Eucarya.

Detection of Alignment Artifacts

In multiple EF-G(2) alignments generated by standard methods (MULTALIN, CLUSTAL W) the eucaryal–archaeal G′ subdomains and some of the ensuing elements of their G″ subdomains were artifactually aligned just underneath the large bacterial G′ subdomain.

In sharp contrast, no archaeal or eucaryal EF sequences matching the bacterial G′ subdomain could be identified among the gap-free binary alignments obtained by BLAST probing the DNA and protein databases with *Aquifex pyrophilus* and *Thermotoga maritima* EF-Gs as the query sequences. And conversely, no bacterial sequences matching the G″ subdomain were retrieved by probing the databases with a variety of eucaryal and archaeal EF-2 species. This result was confirmed by BLAST and FASTA probing of the databases with the limited sequence segments comprising the G′ and G″ subdomains of archaeal, bacterial, and eucaryal EFs. As Table 1 shows, the three sets of query sequences retrieved only homologues of their own domains; the lack of mutual retrieval between the archaeal and the eucaryal G′ subdomains (Table 1) is unexpected, however, possibly reflecting excessive divergence of the sequence elements flanking the archaeal–eucaryal consensus region highlighted in Fig. 2.

In principle, therefore, regions of the multiple alignment arising from artifactual matching of sequence ele-



```
G' subdomain
      158                                                                                                              256       265
       ↓                                                                                                                ↓         ↓
Tth ERLGARPVVMQLPIGREDTFSGIIDVLRMKAYTYGNDLGTDIRE-IPIPEEYLDQAREYHEKLVEVAADFDENIMLKYLEGEEPTEEELVAAIRKGTIDLKITPVFLGSALKN
Eco TRLGANPVPLQLAIGAEEHFTGVVDLVKMKAINWNDADQGVTFEYEDIPADMVELANEWHQNLIESAAEASEELMEKYLGGEELTEAEIKGALRQRVLNNEIILVTCGSAFKN
Apy EKLTIKPVAIQIPLGAEDQFEGVIDLMEMKAIRWLEVTLGAKYEVIDIPPGYQEKAQEWREKMIETIVETDDELMEKYLEAQEITLEELRKALRKATINRQLVPVLCGSAFKN
Pwo ERF                                            Pwo  SKIIMDVNRLIQRYAPEEYKKKWMVRVE-D                              GSVAFGSAYYN
Tac KRF                                            Tac  TKIITDVNRLISKYAPQQFTKEWQVSVQ-D                              GRVAFGSAYNN
4B7 ETL                                            4B7  ASVVSNFNELIDTYAEEEYKEKWKVSIQ-D                              GSVTFGSAKDK
Ehy QSF                                            Ehy  CRSIENVNVLISTY-KDELLGDVQVSPG-E                              GTVAFGSGLHG
Ddi LSF                                            Ddi  RRAIESVNVIVGNT-EDKEFGDVTVSPE-K                              GTVAFGSGLHG
Dme QTF                                            Dme  QRIVENVNVIIATYNDDGGRMGEVRVDPSK                              GSVGFGSGLHG

G'' subdomain
      266                                                                                                                       267
       ↓                                                                                                                         ↓
Tth GSALKN                                                                                                                       K
Eco GSAFKN                                                                                                                       K
Apy GSAFKN                                                                                                                       K
Pwo GSAYYNWALSV                                    Pwo   PFMQRTGVKFNEIID-LTLKGDNKTLRQ                                         RAP
Tac GSAYNNWAISI                                    Tac   PAMAETKITFKDIVE-YVKNGKQKELAQ                                         KNQ
4B7 GSAKDKWAINI                                    4B7   DIMKKKGVTFKDVIDAYSDSGKVEDLVE                                         KAP
Ehy GSGLHGWAFTLEKFAKMWSAKFGID--RKRMLEKLWGDNYWDAKAKKWKKNGKGDHGEVLQRGFVQFCFDPITKLFNAIMEGRKADYEKMLTNLQIKLSADDKEKEGKELLKTVMKLWLP
Ddi GSGLHGWGFTLGRLPKLYAAKFGDPED--KLMGRLWGESYFDATAKKWTSNPQSADGKALPRAFCQFVLEPIYQLTRAIVDEDALKLEKMMNTLQITLAPEDAEIKGKQLVKAVMRKFLP
Dme GSGLHGWAFTLKQFSEMYSEKF--KIDVVKLMNRLWGENFFNAKTKKWQK--QKEADNK--RSFCMYILDPIYKVFDAIMNYKKEEIGTLLEKICVTVKHEDKDKDGKALLKTVMRTWLP
```

Fig. 2. Alignment of sequences situated immediately ahead (G′ subdomain; Tth residues 158–255; *top row*) and immediately beyond (G″ subdomain; Dme residues 224–328; *bottom row*) the fifth consensus element VXXGS[A,G], based on Ævarsson's structure-guided alignment of EF-G(2) sequences. Only representative organisms are shown. Species abbreviations are as in the legend to Fig. 1. *Shaded areas* delimit the sequence elements that are not alignable with any regions of their longer counterparts. The *black area* highlights the putative consensus [I,V]XXVNX[I,L][I,V]XX[Y,M] shared by nine archaeal and nine eucaryal G′ subdomain sequences.

Table 1. BLAST and FASTA retrieval of sequences spanning the **G′** and **G″** subdomains with Archaeal **(A),** Bacterial **(B),** and Eucaryal **(E)** query sequences[a]

G′ subdomain												G″ subdomain			
Query: **Aquifex (B)** [$_{158}$*TIKPV. . .INRQL*$_{257}$]				Query: **Entamoeba (E)** [$_{177}$*CRSIEN. . .SPGE*$_{204}$]				Query: **4B7 (A)** [$_{168}$*ASVV. . .SIQD*$_{196}$]				Query: **Entamoeba (E)** [$_{221}$*EKFAK. . .KTVML*$_{324}$]			
	p(*N*)	Res	id%		*p*(*N*)	Res	id%		*p*(*N*)	Res	id%		*p*(*N*)	Res	id%
B ***Apy***	***7.2e–60***	***99***	***100***	**E** ***Ehy***	***3.1e–11***	***28***	***100***	**A** ***4B7***	***1.1e–11***	***29***	***100***	**E** ***Ehy***	***2.4e–31***	***104***	***100***
B Hpy	9.4e–29	*97*	*51.5*	**E** Cpr	1.2e–05	nr		**A** Pwo	0.012	*29*	*44.8*	**E** Osa	1.5e–28	nr	
B Sty	6.7e–27	*97*	*44.3*	**E** Sce	8.2e–05	*26*	*65.4*	**A** Sso	0.13	*26*	*46.2*	**E** Dme	1.5e–25	nr	
B Mle	1.4e–23	*95*	*49.5*	**E** Osa	0.00037	nr		**A** Tac	0.17	*29*	*34.5*	**E** Cpr	4.5e–25	nr	
B Atu	1.6e–23	*96*	*44.8*	**E** Bho	0.00054	*27*	*55.6*	**A** Mja	0.39	nr		**E** Ddi	4.7e–25	*102*	*44.1*
B Eco	2.0e–23	*97*	*43.3*	**E** Ata	0.00062	nr		**A** Sac	0.46	nr		**E** Sce	1.3e–24	nr	
B Tma	2.4e–23	*96*	*49.0*	**E** Tcr	0.0014	nr						**E** Cel	1.5e–24	*102*	*38.2*
B Ecr	3.0e–23	nr		**E** Bvu	0.0019	nr						**E** Bho	2.1e–24	*106*	*44.3*
B Rpr	5.7e–23	*97*	*41.2*	**E** Cke	0.017	*25*	*60.0*					**E** Gla	1.1e–22	nr	
B Hin	4.3e–23	*97*	*38.1*	**E** Ddi	0.32	nr						**E** Dme	5.0e–22	*102*	*41.2*
B Bsu	1.5e–21	*94*	*47.9*									**E** Bvu	1.9e–21	nr	
B Syn	1.1e–20	*93*	*43.0*									**E** Gga	2.7e–21	*86*	*47.7*
B Mlu	3.1e–20	*99*	*43.4*									**E** Cke	1.6e–16	*103*	*44.7*
B Ani	3.3e–20	*93*	*45.2*									**E** Ham	1.3e–14	*86*	*46.5*
B Tth	1.9e–18	*94*	*43.6*									**E** Hsa	7.1e–14	*86*	*46.5*
B Mca	5.3e–16	nr										**E** Mmu	2.4e–14	nr	
B Mpn	2.1e–13	*98*	*34.7*												
B Bbu	1.1e–11	nr													
B Spl	2.3e–11	*99*	*42.4*												
B Mge	nr	*95*	*29.5*												

[a] Res and id% indicate the number of overlapping residues and the percentage identical residues, respectively, in the overlapping fragments given by a FASTAp search with the indicated query sequences (*italics*); *p*(*N*) is the Poisson probability of random homology given by a tBLASTn search of the databanks (Gish et al. 1993). Sequences are ranked in order of decreasing similarity to the query sequences. Species abbreviations are listed in the legend to Fig. 1. nr, not retrieved.

ments that are unique to different phylogenetic domains can be identified by searching the binary alignments given by BLAST (and FASTA) for the presence or absence of the alignment schemes generated by the multialignment algorithms (or visually inferred).

In three sections of the multiple alignment (regions A, B, and C; boxed in Fig. 1), the matching of the bacterial sequences with those of Archaea and Eucarya was not retrieved by scrutiny of four inventories of gap-free binary alignments obtained by BLAST probing of the protein databases with EF sequences representative of Bacteria (*Aquifex pyrophilus*), euryarchaeotes (*Pyrococcus*

Table 2. BLAST and FASTA retrieval of Archaeal **(A),** Bacterial **(B),** and Eucaryal **(E)** sequences with segments spanning regions **A, B,** and **C** of the global EF-G (2) alignment[a]

Region A

Query: **Entamoeba (E)**
[$_{448}$*KYRTS. . .AMANC*$_{470}$]

	p(*N*)	Res	id%
E ***Ehy***	***6.2e–09***	***23***	***100***
E Cpr	6.3e–09	nr	
E Gla	0.0027	nr	
E Ddi	0.0027	*23*	*65.2*
E Bvu	0.013	nr	
E Bho	0.11	*23*	60.9
E Dme	0.11	*23*	*60.9*
E Cel	0.11	*23*	*60.9*
E Gga	0.11	*23*	*60.9*
E Cke	0.15	*23*	*56.5*
E Dme	0.11	*23*	*60.9*
E Hsa	0.27	23	56.5
E Ham	0.28	*23*	*56.5*
A Mja	0.46	*23*	*52.2*
A Tac	0.995	*22*	*45.5*
A Sso	0.995	*20*	*50.0*
A Sac	0.0004	*22*	*50.0*
A Mva	0.9995	*23*	*43.5*
A Pwo	nr	*23*	*31.1*

Query: **Aquifex (B)**
[*IDLPPVKGTNPNTGEEEERRPLD*]

	p(*N*)	Res	id%
B ***Apy***	***1.2e–05***	***19***	***100***
B Ecr	0.99	nr	
B Bsu	nr	*19*	*52.6*
B Tma	nr	*19*	*52.6*

Query: **Escherichia (B)**
[*VPAINGILDDGKDTPAERH*]

	p(*N*)	Res	id%
B ***Eco***	***8.4e–06***	***19***	**100**
B Sty	8.6e–06	*19*	*100*
B Spl	nr	*13*	*53.8*

Query: **Micrococcus (B)**
[*DAGPVKGHAVNDEEVVLEREV*]

	p(*N*)	Res	id%
B ***Mlu***	***3.7e–06***	***21***	**100**
B Tma	nr	*19*	*52.6*

Region B

Query: **Giardia (E)**
[$_{598}$*VMAK. . .NLIL*$_{674}$]

	p(*N*)	Res	id%
E ***Gla***	***1.4e–45***	***77***	***100***
E Ata	4.9e–21	nr	
E Sce	3.6e–19	*73*	*50.7*
E Bvu	3.6e–19	nr	
E Rat	9.7e–19	*74*	*45.9*
E Gga	2.1e–17	*74*	*45.9*
E Hsa	2.9e–17	*74*	*45.9*
E Cke	3.9e–17	*75*	*45.5*
E Ehy	4.6e–16	*63*	*52.4*
E Cpr	5.7e–17	nr	
E Bho	2.7e–14	*68*	*44.1*
E Dme	5.2e–11	*68*	*44.1*
E Tcr	1.5e–13	nr	
A Sso	1.6e–12	*57*	*49.1*
A Sac	1.0e–11	*57*	*47.4*
A Dmo	1.4e–11	*57*	*42.1*
E Pfa	7.2e–13	nr	
E Ddi	1.5e–0.8	*61*	*42.6*
A Tac	7.9e–0.7	*60*	*50.0*
A Mja	5.8e–0.6	*62*	*41.9*
A Pwo	1.2e–0.5	nr	
E Dme	0.00011	*57*	*35.1*
A Mva	0.0018	*75*	*33.3*
E Tcr	0.043	nr	

Query: **Sulfolobus (A)**
[$_{482}$*EGK. . .FVDLT*$_{554}$]

	p(*N*)	Res	id%
A ***Sso***	***1.1e–41***	***72***	***100***
A Sac	2.4e–45	*72*	*79.2*
A Dmo	1.8e–28	*72*	*68.1*
E Gla	2.2e–12	nr	
E Sce	1.5e–0.9	*57*	*42.1*
A 4B7	2.7e–0.9	nr	
E Cke	1.5e–0.9	*63*	*41.3*
E Ata	2.8e–0.9	nr	
E Mmu	1.2e–0.8	nr	
E Gga	3.0e–0.7	*57*	*36.8*
A Hha	3.1e–0.7	*70*	*40.0*
E Ddi	3.4e–0.7	nr	
E Hsa	3.7e–0.7	*57*	*36.8*
E Cel	4.5e–0.7	*57*	*36.1*
E Cpr	7.6e–0.7	nr	
E Rat	8.4e–0.7	*57*	*36.9*
A Pwo	9.1e–0.6	*72*	*47.2*
E Ham	1.0e–0.6	*57*	*36.8*
E Ehy	1.9e–0.6	*74*	*37.8*
A Tac	2.6e–0.6	*71*	*39.4*
E Pfa	1.4e–0.5	nr	
A Mja	5.5e–0.5	*75*	*41.3*
E Gpe	0.0046	nr	

Query: **Pyrococcus (A)**
[$_{479}$*EGK. . . .FLDNT*$_{551}$]

	p(*N*)	Res	id%
A ***Pwo***	***8.2e–44***	***72***	***100***
A Tac	2.0e–15	*73*	*52.1*
A Mva	8.5e–12	*74*	*43.2*
A Mja	4.1e–11	*74*	*43.3*
A Hha	3.1e–08	*72*	*34.7*
A Dmo	7.4e–08	*74*	*44.6*
E Cke	2.8e–06	*62*	*43.5*
A Sso	3.8e–06	*72*	*47.2*
A Sac	9.0e–05	*74*	*43.2*
E Cpr	0.0012	nr	
E Sce	0.00013	*30*	*56.7*
E Gla	0.0027	nr	
E Cel	0.022	*31*	*45.2*
E Rat	0.057	*47*	*44.7*
E Mmu	0.058	nr	
E Hsa	0.061	*31*	*45.2*
A 4B7	0.063	nr	
E Ham	0.064	*47*	*44.7*
E Gga	0.065	nr	
E Dme	0.69	*55*	*40.0*
E Bho		*59*	*42.4*

woesei), crenarchaeotes (*Sulfolobus acidocaldarius*), and Eucarya (*Entamoeba histolytica*). And the lack of relatedness of the bacterial and archaeal–eucaryal sequences forming these three sections of the global alignment was further confirmed by probing the databases with the limited sequence elements spanning the A, B, and C boxes; a selection of the results obtained in this way is given in Table 2.

Table 2. Continued

Region C											
Query: **Entamoeba (E)**				Query: **Sulfolobus (A)**				Query: **Thermotoga (B)**			
[*DAIHRGGAQMIPCARRCCFACVLTG*]				[*DPAHRGPAQLYPAVRNAIFAGILTS*]				[*DSSEMAFKIAASMAFKEAMKKA*]			
	p(*N*)	Res	id%		*p*(*N*)	Res	id%		*p*(*N*)	Res	id%
E *Ehy*	***3.3e–13***	***25***	***100***	**A *Sac***	***4.2e–10***	***25***	***100***	**B *Tma***	***1.0e–05***	***22***	***100***
E Gga	1.2e–07	*25*	*76.0*	**A** Sso	7.7e–09	*25*	*88.0*	**B** Pro	0.00059	*22*	*86.4*
E Hsa	3.7e–06	*25*	*72.0*	**A** Dmo	2.8e–08	*25*	*84.0*	**B** Ani	0.0022	*22*	*81.8*
E Ham	3.9e–06	*25*	*72.0*	**A** Pwo	0.012	*25*	*52.0*	**B** Eco	0.0027	*22*	*77.3*
E Tcr	3.1e–05	nr		**A** Tac	0.25	*22*	*54.5*	**B** Hpy	0.0042	*22*	*81.8*
E Dme	8.9e–05	*25*	*60.0*	**A** Hha	0.43	nr		**B** Mge	0.0052	*21*	*81.0*
E Rat	9.6e–05	*25*	*72.0*	**E** Mmu	0.77	nr		**B** Sra	0.0071	*22*	*77.3*
E Cel	0.00018	*25*	*68.0*	**E** Hsa	0.78	*25*	*40.0*	**B** Tth	0.0093	*22*	*77.3*
E Ddi	0.00071	*25*	*64.0*	**E** Rat	0.84	*25*	*40.0*	**B** Ecr	0.013	nr	
E Cke	0.00072	*25*	*64.0*	**E** Ehy	0.87	*25*	*48.0*	**B** Mlu	0.014	*22*	*72.7*
E Sce	0.032	*23*	*60.9*	**E** Ham	0.87	*25*	*40.0*	**B** Hin	0.019	*22*	*72.7*
E Cpr	0.062	nr		**E** Cel	0.92	*25*	*40.0*	**B** Mge	0.027	*21*	*81.0*
A Mja	0.17	*15*	*86.7*	**E** Sce	0.94	*23*	*47.8*	**B** Mpn	0.036	*20*	*85.0*
A Mva	0.30	*15*	*66.7*	**E** Gga	0.995	*25*	*40.0*	**B** Ngo	0.060	*22*	*63.6*
A Gla	0.40	nr		**A** 4B7	0.9993	nr		**B** Bsu	0.065	*21*	*71.4*
A Sac	0.52	*25*	*48.0*	**E** Gla	0.9994	nr		**B** Syn	0.067	*22*	*72.7*
A 4B7	0.87	nr		**A** Mja	0.9995	*23*	*47.8*	**B** Spl	0.086	*22*	*72.7*
A Dmo	0.94	*25*	*48.0*	**E** Ddi	nr	*25*	*40.0*	**B** Bbu	0.93	nr	
A Hha	0.95	*15*	*73.3*					**B** Rpr	0.97	*22*	*59.1*
A Sso	nr	*25*	*48.0*					**B** Atu	0.97	*22*	*59.1*
E Bho	nr	*25*	*44.0*					**B** Mtu	0.995	nr	
								B Mle	0.999	*22*	*59.0*

[a] See Table 1, footnote a.

The first of the unretrieved regions (box A, highlighted in Fig. 3A) is situated at the very beginning of the structural domain II. The sequence elements spanning this structural space (*T. thermophilus* residues 290–303) are well conserved among Archaea–Eucarya but exhibit considerable primary structural and length heterogeneity and are not unambiguously alignable among Bacteria. Notably, a BLAST/FASTA search with the eucaryal EF-2 residues comprising the A region (Table 2) retrieved only the archaeal counterparts, indicating that these two groups of sequences are variants of the same ancestral theme; in contrast, a search using the corresponding EF-G segments retrieved only one to three of all available bacterial homologues (no fewer than 25 EF-G sequences) (see Table 2). Whether the lack of mutual retrieval between the EF-2 and the EF-G sequences in the A box reflects a peculiar instability of the bacterial sequences spanning this region or a unique insertion–substitution event occurring during EF-2 evolution is an unresolvable issue. If an insertion occurred in this section of the EF-2 sequence, this should be placed between the motif IPPI (residues 286–289 of *T. thermophilus*, variants of which are ostensibly present in archaeal EF-2) and the motif PDPNG (residues 303–307 of *T. thermophilus;* see Fig. 1 legend).

The global alignment region comprising the second unretrieved alignment scheme (Fig. 1, box B; highlighted in Fig. 3B) spans a conserved lysine (*T. thermophilus* residue 496) and the relatively conserved sequence [F,I,L,M,V]X(N,D)X[I,T]XG (*T. thermophilus* residues 525–537) with a spacing of only 30 or 31 residues in Bacteria but up to 71 residues in Archaea and Eucarya. The archaeal and eucaryal sequences spanning the B region exhibit a high degree of similarity (strikingly so in the vicinity of the N-terminal lysine) and appear to be unrelated (by recognizable signatures) to their shorter bacterial counterparts. As expected from this lack of similarity, no bacterial homologue of the archaeal–eucaryal sequences was retrieved by BLAST/FASTA probing of the databases with limited EF-2 segments. In contrast, the archaeal and eucaryal sequences were mutually retrieved, and in most cases, the fragment overlap (FASTA results) covered the the full length (72–77 residues) of the query sequence (Table 2). The simplest explanation of this is that the bacterial and the archaeal–eucaryal sequences comprising this section of the global alignment are (ancestrally) unrelated entities resulting from genetic (insertion–substitution) events after the divergence of the bacterial and the archaeal–eucaryal lineages. According to the proposed alignment, this should have occurred between the two motifs bordering the boxed B region (underlined in Fig. 1). As Fig. 3B shows, however, shifting the archaeal–eucaryal insert to the right generates a new putative consensus element that has a conserved proline in all but one (*Halobacterium halobium*) of 20 EF-G(2) sequences. This suggests that

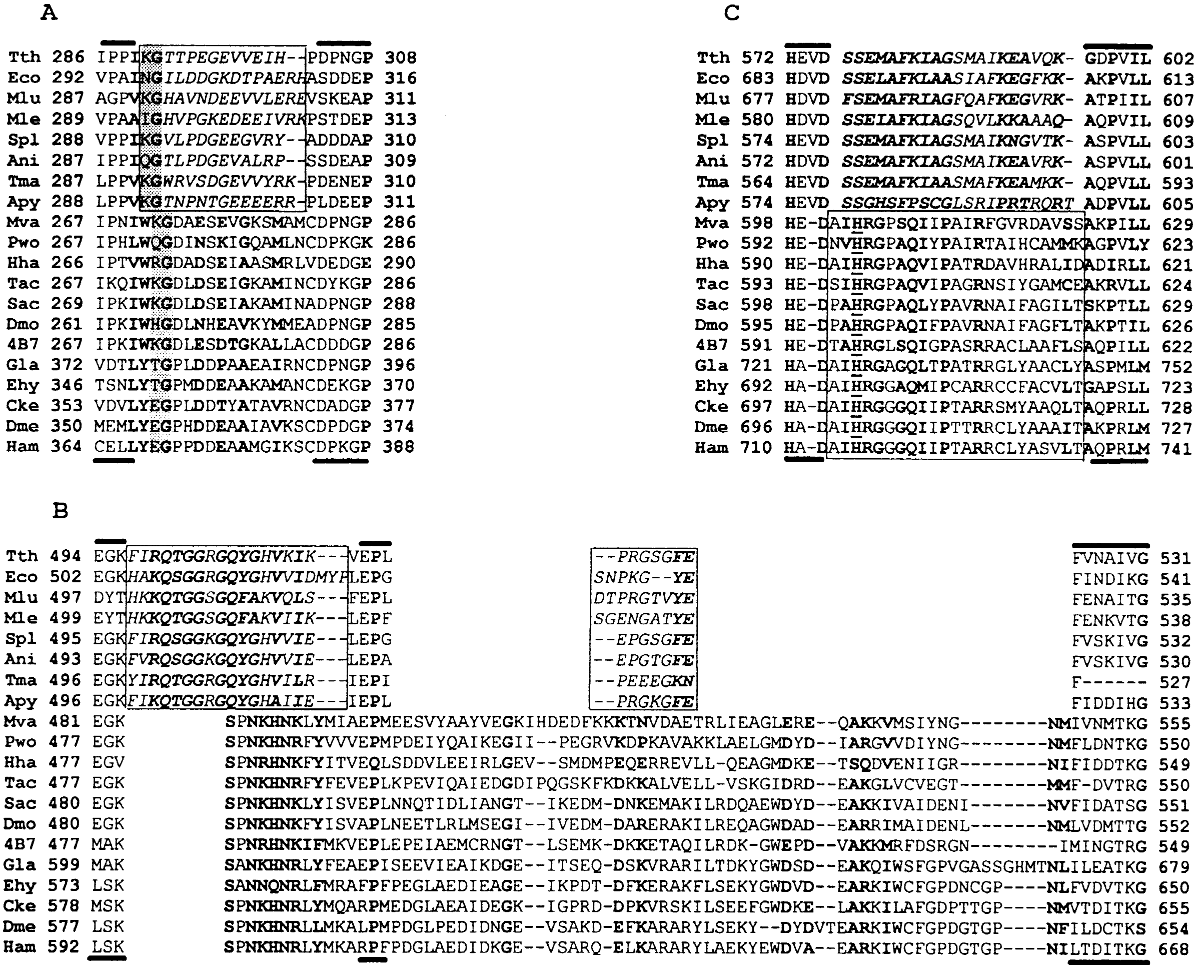

A
Tth 286 IPPIKGTTPEGEVVEIH--PDPNGP 308
Eco 292 VPAINGILDDGKDTPAERHASDDEP 316
Mlu 287 AGPVKGHAVNDEEVVLEREVSKEAP 311
Mle 289 VPAAIGHVPGKEDEEIVRKPSTDEP 313
Spl 288 VPPIKGVLPDGEEGVRY--ADDDAP 310
Ani 287 IPPIQGTLPDGEVALRP--SSDEAP 309
Tma 287 LPPVKGWRVSDGEVVYRK-PDENEP 310
Apy 288 LPPVKGTNPNTGEEEERR-PLDEEP 311
Mva 267 IPNIWKGDAESEVGKSMAMCDPNGP 286
Pwo 267 IPHLWQGDINSKIGQAMLNCDPKGK 286
Hha 266 IPTVWRGDADSEIAASMRLVDEDGE 290
Tac 267 IKQIWKGDLDSEIGKAMINCDYKGP 286
Sac 269 IPKIWKGDLDSEIAKAMINADPNGP 288
Dmo 261 IPKIWHGDLNHEAVKYMMEADPNGP 285
4B7 267 IPKIWKGDLESDTGKALLACDDDGP 286
Gla 372 VDTLYTGPLDDPAAEAIRNCDPNGP 396
Ehy 346 TSNLYTGPMDDEAAKAMANCDEKGP 370
Cke 353 VDVLYEGPLDDTYATAVRNCDADGP 377
Dme 350 MEMLYEGPHDDEAAIAVKSCDPDGP 374
Ham 364 CELLYEGPPDDEAAMGIKSCDPKGP 388
B
Tth 494 EGKFIRQTGGRGQYGHVKIK---VEPL --PRGSGFE FVNAIVG 531
Eco 502 EGKHAKQSGGRGQYGHVVIDMYPLEPG SNPKG--YE FINDIKG 541
Mlu 497 DYTHKKQTGGSGQFAKVQLS---FEPL DTPRGTVYE FENAITG 535
Mle 499 EYTHKKQTGGSGQFAKVIIK---LEPF SGENGATYE FENKVTG 538
Spl 495 EGKFIRQSGGKGQYGHVVIE---LEPG --EPGSGFE FVSKIVG 532
Ani 493 EGKFVRQSGGKGQYGHVVIE---LEPA --EPGTGFE FVSKIVG 530
Tma 496 EGKYIRQTGGRGQYGHVILR---IEPI --PEEEGKN F------ 527
Apy 496 EGKFIKQTGGRGQYGHAIIE---IEPL --PRGKGFE FIDDIHG 533
Mva 481 EGK SPNKHNKLYMIAEPMEESVYAAYVEGKIHDEDFKKKTNVDAETRLIEAGLERE--QAKKVMSIYNG--------NMIVNMTKG 555
Pwo 477 EGK SPNKHNRFYVVVEPMPDEIYQAIKEGII--PEGRVKDPKAVAKKLAELGMDYD--IARGVVDIYNG--------NMFLDNTKG 550
Hha 477 EGV SPNRHNKFYITVEQLSDDVLEEIRLGEV--SMDMPEQERREVLL-QEAGMDKE--TSQDVENIIGR--------NIFIDDTKG 549
Tac 477 EGK SPNKHNRFYFEVEPLKPEVIQAIEDGDIPQGSKFKDKKALVELL-VSKGIDRD--EAKGLVCVEGT--------MMF-DVTRG 550
Sac 480 EGK SPNKHNKLYISVEPLNNQTIDLIANGT--IKEDM-DNKEMAKILRDQAEWDYD--EAKKIVAIDENI-------NVFIDATSG 551
Dmo 480 EGK SPNKHNKFYISVAPLNEETLRLMSEGI--IVEDM-DARERAKILREQAGWDAD--EARRIMAIDENL-------NMLVDMTTG 552
4B7 477 MAK SPNRHNKIFMKVEPLEPEIAEMCRNGT--LSEMK-DKKETAQILRDK-GWEPD--VAKKMRFDSRGN--------IMINGTRG 549
Gla 599 MAK SANKHNRLYFEAEPISEEVIEAIKDGE--ITSEQ-DSKVRARILTDKYGWDSD--EAKQIWSFGPVGASSGHMTNLILEATKG 679
Ehy 573 LSK SANNQNRLFMRAFPFPEGLAEDIEAGE--IKPDT-DFKERAKFLSEKYGWDVD--EARKIWCFGPDNCGP----NLFVDVTKG 650
Cke 578 MSK SPNKHNRLYMQARPMEDGLAEAIDEGK--IGPRD-DPKVRSKILSEEFGWDKE--LAKKILAFGDPTTGP----NMVTDITKG 655
Dme 577 LSK SPNKHNRLLMKALPMPDGLPEDIDNGE--VSAKD-EFKARARYLSEKY--DYDVTEARKIWCFGPDGTGP----NFILDCTKS 654
Ham 592 LSK SPNKHNRLYMKARPFPDGLAEDIDKGE--VSARQ-ELKARARYLAEKYEWDVA--EARKIWCFGPDGTGP----NILTDITKG 668
C
Tth 572 HEVD SSEMAFKIAGSMAIKEAVQK- GDPVIL 602
Eco 683 HDVD SSELAFKLAASIAFKEGFKK- AKPVLL 613
Mlu 677 HDVD FSEMAFRIAGFQAFKEGVRK- ATPIIL 607
Mle 580 HDVD SSEIAFKIAGSQVLKKAAAQ- AQPVIL 609
Spl 574 HEVD SSEMAFKIAGSMAIKNGVTK- ASPVLL 603
Ani 572 HDVD SSEMAFKIAGSMAIKEAVRK- ASPVLL 601
Tma 564 HEVD SSEMAFKIAASMAFKEAMKK- AQPVLL 593
Apy 574 HEVD SSGHSFPSCGLSRIPRTRQRT ADPVLL 605
Mva 598 HE-DAIHRGPSQIIPAIRFGVRDAVSSAKPILL 629
Pwo 592 HE-DNVHRGPAQIYPAIRTAIHCAMMKAGPVLY 623
Hha 590 HE-DAIHRGPAQVIPATRDAVHRALIDADIRLL 621
Tac 593 HE-DSIHRGPAQVIPAGRNSIYGAMCEAKRVLL 624
Sac 598 HE-DPAHRGPAQLYPAVRNAIFAGILTSKPTLL 629
Dmo 595 HE-DPAHRGPAQIFPAVRNAIFAGFLTAKPTIL 626
4B7 591 HE-DTAHRGLSQIGPASRRACLAAFLSAQPILL 622
Gla 721 HA-DAIHRGAGQLTPATRRGLYAACLYASPMLM 752
Ehy 692 HA-DAIHRGGAQMIPCARRCCFACVLTGAPSLL 723
Cke 697 HA-DAIHRGGGQIIPTARRSMYAAQLTAQPRLL 728
Dme 696 HA-DAIHRGGGQIIPTTRRCLYAAAITAKPRLM 727
Ham 710 HA-DAIHRGGGQIIPTARRCLYASVLTAQPRLM 741

Fig. 3. **A** Magnification of the alignment region in box A in Fig. 1 with additional species; the *shading* indicates that Ævarsson's alignment in this particular region is probably inaccurate, as a universally conserved XG diplet (KG in most species) and an archaeal–eucaryal [Y,W] insertion become apparent by the introduction of a single gap in all of the EF-G sequences immediately after *Thermus* residue 289. **B** Interpretation of the alignment regions in box B in Fig. 1; a putative motif having a proline as the second element can be generated by shifting to the right the characters underlined in Fig. 1. **C** Magnification of the sequence alignment in box C with additional species. *Frames* delimit bacterial sequences that are not alignable with their archaeal–eucaryal counterparts.

insertions events may have occurred both N-terminally and C-terminally to this site.

The region of the global alignment comprising the third unretrieved alignment scheme (box C; highlighted in Fig. 3C) encompasses the motifs (H[D,E,A][V,*del*]D and [A,S,G]X[P,I,R]X [I,L,M][L,M]EP) and harbors the histidine that is ADP-ribosylatable by diphtheria toxin in Eucarya and Archaea (Kessel and Klink, 1980; Lechner et al. 1988). The archaeal–eucaryal sequences spanning this region (23 residues) are strikingly similar to each other (65–70% identity) and share no apparent similarity to their bacterial counterparts (20–21 residues). As expected from this lack of similarity, no relatedness between the archaeal–eucaryal and the bacterial sequences could be inferred by probing the DataBanks with the sequence elements spanning the C region (Table 2). Both the lack of relatedness of the EF-2 and EF-G segments and the remarkable conservation of the sequence elements within each of the two groups of EF sequences strongly suggest that the archaeal–eucaryal and the bacterial sequences nested between the two (highly conserved) flanking motifs are ancestrally urelated entities probably resulting from genetic insertion events.

Finally, the EF-2 sequences exhibit an N-terminal accretion (framed N-terminal region in Fig. 1) having no counterpart in EF-G and are ostensibly related by a an obvious signature element (RXRKGL).

Phylogenetic Trees

Figure 4–6 show the phylogenetic trees inferred from the alignment positions overlined in Fig. 1 (503 sites) and from a colinear alignment of 1006 first *plus* second codon positions. Compared to previous analyses (Cammarano 1992; Creti et al. 1994), the two data sets do not include (i) spurious characters generated by the misalignment of the G′ and G″ subdomain sequences (Ævarsson,

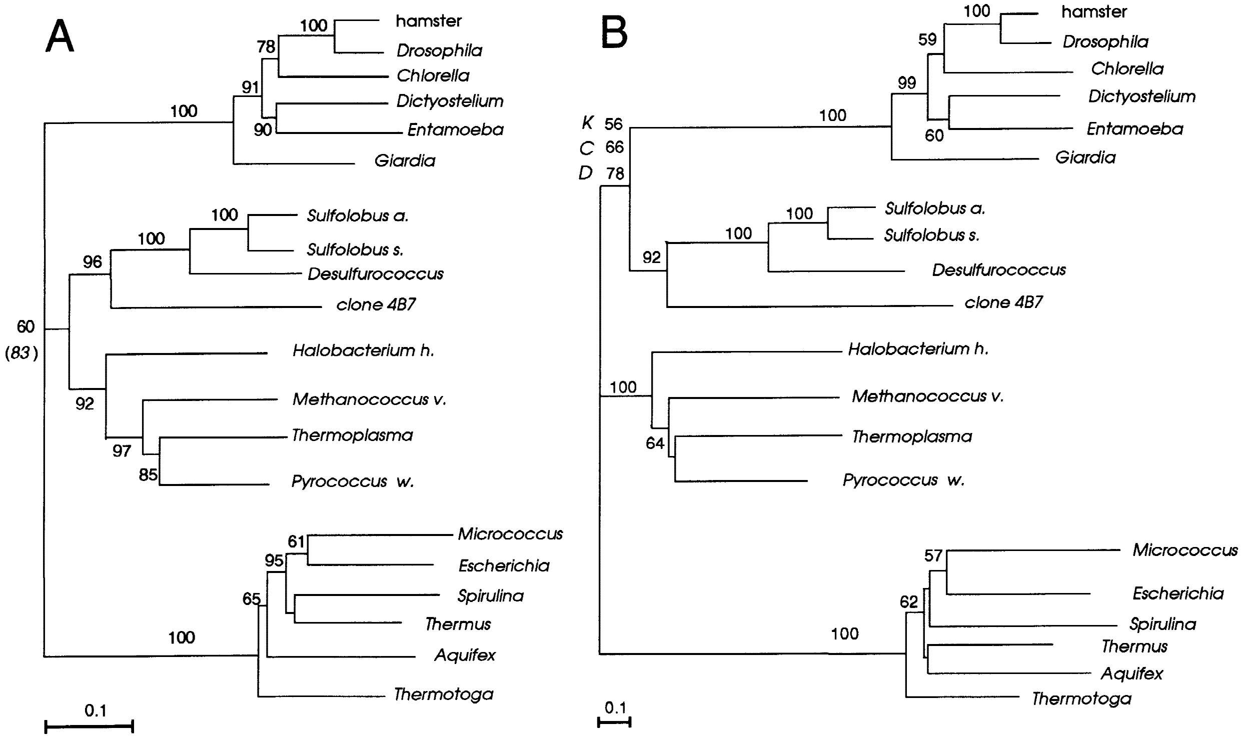


Fig. 4. Evolutionary trees of EF-G(2) sequences inferred from the 503 sites *overlined* (plus signs) in Fig. 1. The numbers shown are percentages of 100 boostrap replicates in which the same internal branch was recovered. **A** Distance-matrix tree constructed from the first and second codon-position data set (1006 sites) by the least-squares method (program FITCH); the evolutionary distances were calculated by the Kimura two-parameter model of nucleotide substitution (program DNADIST) with an estimator, $R \cong 0.8$, of the TI/TV rate ratios (Wakeley 1996; Kumar et al. 1993) (bootstrap analysis of 100 resamplings). The italic number in parentheses *below* the archaeal branch is the BCL of a tree inferred after deselection of the 4B7 sequence. **B** Distance-matrix tree inferred from the protein data set (503 sites) by the program FITCH with evolutionary distances calculated by the category method with the George–Hunt–Barker categorization of amino acids (program PROTDIST); numbers *above* the branch supporting the crenarchacal–eucaryal clade are BCLs of least-squares trees based on the Kimura (K), category (C), and Dayhoff (D) corrections (bootstrap analyses of 100 resamplings); deselection of 4B7 resulted in monophyletic Archaea (BCL, 60–65%) in the phylogenies based on the C and the K corrections and substantially reduced the bootstrap support for paraphyletic Archaea of the tree based on the D correction (BCL, 55%). Scale lengths represent 0.1 substitution per site.

1995) or (ii) positions comprising the unalignable bacterial sequences of region A and the artifactually aligned (putative) insertions of regions B and C in Fig. 1. Also, new sequences representing deep-branching lineages [notably the uncultivated planktonic marine Archaeon represented by the 4B7 clone (Stein et al. 1996)] have been used in the present analysis. The alignments used for phylogeny treeing are available (file EF-G.aln) *via* anonymous ftp at ftp.bce.med.uniromal.it, dir/cammara.

Distance-Matrix Analysis. Unlike previous analyses (Creti et al. 1994) the nucleotide and amino acid data sets (Figs. 4A and B, respectively) support, albeit weakly, alternative phylogenetic placements of the crenarchaotes [bootstrap confidence levels (BCL), <90%)]: while nucleotide sequence analysis gives a monophyletic Archaea (BCL, 60%), analysis of amino acid sequences gives a paraphyletic association of the crenarchaeotes with Eucarya with weak to moderate bootstrap support (BCL, 56–78%), depending on the correction method used (see Fig. 4B legend). The robustness of the two trees in Fig. 4 was critically affected by the 4B7 sequences. Deselecting 4B7 resulted in increased support for monophyletic Archaea in the DNA-based tree (BCL of 83% instead of only 60%), and gave a monophyletic-Archaea tree (BCL, 55–65%) in the case of protein-based phylogenies inferred by use of the "Kimura" and "Category" correction methods (Fig. 4 legend).

Maximum-Likelihood (ML) Analysis. Figures 5A and B show the single best trees inferred by exhaustive search of a partially constrained starting tree from the nucleotide (NucML) and amino acid sequences (ProtML). The two data sets support alternative topologies, albeit modestly. Whereas the NucML analysis weakly supports (66% confidence) archaeal monophyly, ProtML moderately supports (78% confidence) a paraphyletic Archaea, with the crenarchaeotes forming a monophyletic clade with the Eucarya; the crenarchaeal–eucaryal clade was also supported, albeit more weakly (BCL, 65%), by a parallel analysis in which 100 bootstrap samples of the protein data set (generated with

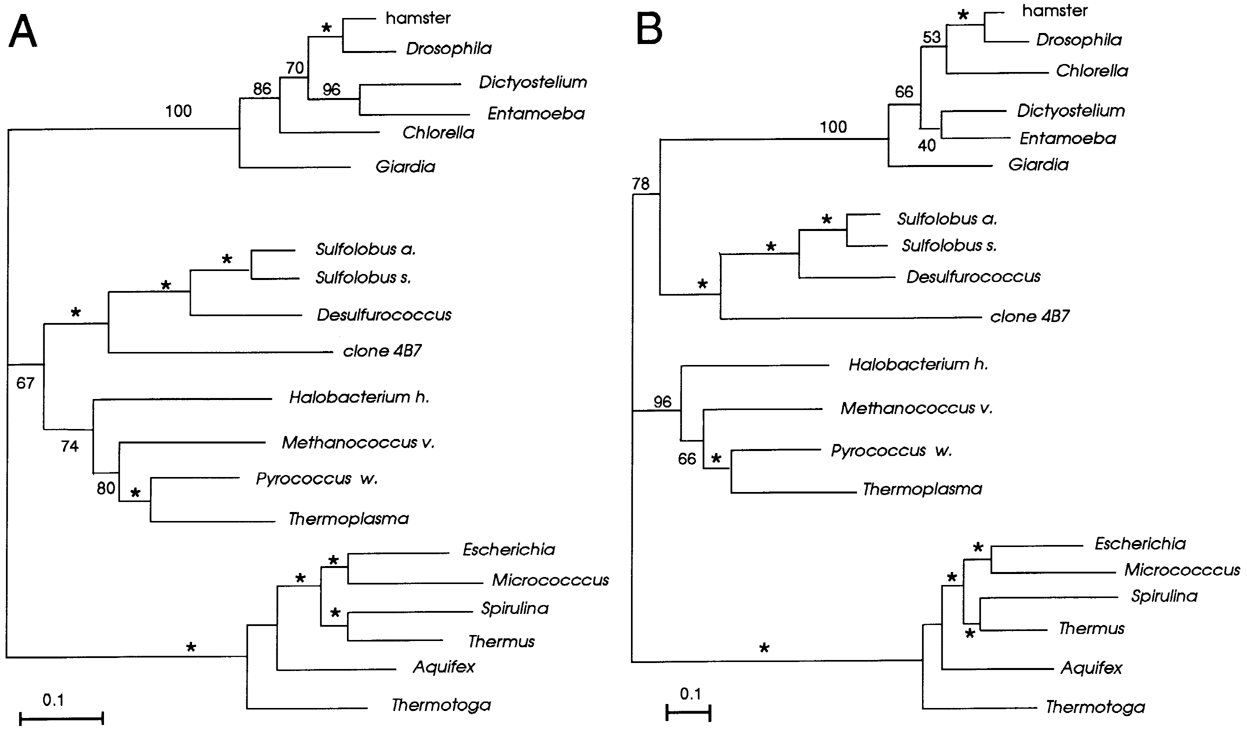


Fig. 5. **A** Maximum-likelihood analysis (program NucML) of a first plus second codon-position data set (1006 sites). An identical tree was obtained with the DNAML program; in the latter case, however, only 15 OTUs could be used for bootstrap analysis, and these gave a monophyletic Archaea in 65 of 100 resamplings **B** Maximum-likelihood analysis (program ProtML) of the 503-amino acid data set corresponding to the nucleotide data set used to infer tree A. The two trees shown are the single best trees obtained by analysis of the top-ranking 1000 topologies (of 2,027,025) selected by an exhaustive search of the partially constrained starting tree (((Eco,Mlu), (Tth,Spl)), Apy, Tma, {(Ham,Dme), (Pwo,Tac), (((Sso, Sac), Dmo), 4B7), Cke, Ddi, Hha, Mva, Gla, Ehy}) in which the 20 OTUs used for phylogenetic analysis were organized in 10 topological elements based on a preliminary analysis done with DNAML. *Asterisks* indicate constrained nodes. Numbers attached to unconstrained nodes represent local bootstrap probabilities for individual topological elements calculated by summation of the bootstrap probabilities of all the trees showing that element among the 1000 trees retained by the approximate log-likelihood method. Scale lengths represent 0.1 substitution per site.

SEQBOOT) were analyzed by the "star decomposition" algorithm of ProtML (Adachi and Hasegawa 1992).

The differences in the log-likelihoods of alternative trees from those of the ML trees are shown in Table 3 along with their SEs (Kishino et. al. 1990) and with the bootstrap probabilities for tree *i*, being the ML tree among the alternatives. Of 15 possible trees generated by five topological elements (Bacteria, Eucarya, crenarchaeotes, halophiles, and euryarchaeotes except halophiles), the Archaea–paraphyletic tree favored by the amino acid sequence analysis (tree 1 in Table 3) could be confidently discriminated from most alternatives (trees 4 through 15) by the criterion of more than 2 SE of log-likelihood difference; tree 1, however, was not significantly favored over an otherwise identical tree (tree 2) showing monophyletic Archaea by the criterion of only 0.68 SE of log-likelihood difference (Δ*l*, −3.4 ± 5.0), and was also poorly discriminated (1.4 SE of Δ*l*) from a paraphyletic Archaea tree (tree 3) showing the euryarchaeotes as the sister group to Eucarya. And conversely, the Archaea monophyletic tree favored by nucleotide sequence analysis (tree 2 in Table 3) was not significantly better than the alternative tree 1 showing paraphyletic Archaea by the criterion of 0.5 SE of log-likelihood difference (Δ*l*, −2.3 ± 4.5).

The two trees in Fig. 5 (tree topologies 1 and 2 in Table 3) were also contrasted with otherwise identical trees in which the 4B7 clone was individually affiliated to the Eucarya, and the possibility of a monophyletic euryarchaeal–crenarchaeal clade excluding 4B7 could be strongly rejected by the criterion of more than 2 SE of log-likelihood difference. Similarly there was strong discrimination against the deconstruction of Archaea into three monophyletic taxa (the euryarchaeotes, the crenarchaeotes, and the 4B7 lineage), with 4B7 sharing a more recent common ancestor with Eucarya (results not shown).

The discrimination between tree 1 and tree 2 in Table 3 is comparable to that borne by a recent analysis of Baldauf et al. (1996) (382 amino acid positions) showing 0.935 SE of log-likelihood difference between a ML EF-G(2) tree with paraphyletic Archaea and an otherwise identical tree showing monophyletic Archaea. In that the log-likelihood differences between alternative trees are smaller than their SEs, neither analysis convincingly supports archaeal paraphyly (Kishino et al. 1990).

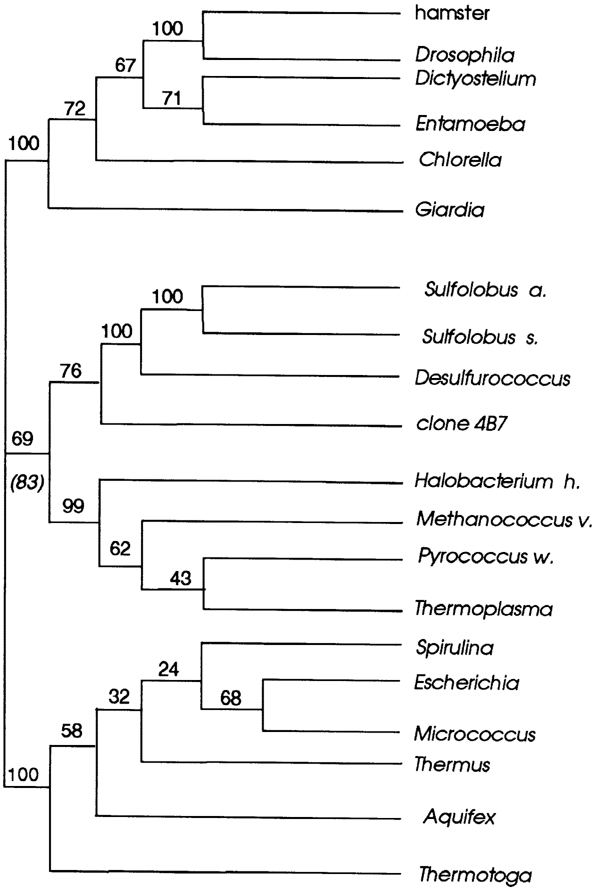


Fig. 6. Parsimony tree inferred from the protein data set with the program PROTPARS (bootstrap analysis of 100 resamplings). The tree requires 3467 substitutions (neglecting synonymous changes), which falls short of the maximum-parsimony tree (3484 steps) obtained by DNAPARS from the 1006 first plus second codon-position data set).

Stronger support for a monophyletic crenarchaeal–eucaryal clade comes from a ML analysis of an EF-G(2) alignment inferred by a maximum-likelihood method (Hashimoto and Hasegawa 1996). Based on the ML alignment (529 amino acid positions), the sisterhood of the crenarchaeotes with Eucarya was given at 99% bootstrap probability and the Archaea–paraphyletic tree could be confidently discriminated from an alternative tree showing monophyletic Archaea by the criterion of more than 2 SE of log-likelihood difference (Δl, −14.3 ± 6.5). This discrepancy with our results (showing Δl, −3.4 ± 5.0) is most probably accounted for by differences in character selection. Unlike the present report, the Hashimoto–Hasegawa data set includes (i) a section (15 residues) in which segments of the G′ subdomain of Bacteria are aligned with segments of the archaeal–eucaryal G′ subdomain; (ii) the entire region corresponding to our box A (residues 289–302 of *T. thermophilus* EF-G), which we have discarded for the reasons given above, and (iii) a large section (41 positions) of the alignment encompassing the region which is immediately N-terminal to the B box in Fig. 1 and the whole B box sequences (corresponding to sequences aligned with *T. thermophilus* residues 489–525). Also, unlike the present report, the Hashimoto–Hasegawa data set does not include 15 positions belonging to the domain II sequences that are immediately N-terminal to the start of domain III (residues 381–405 of *T. thermophilus* EF-G in our alignment) in which their alignment deviates from the Ævarsson's structure-guided alignment.

Maximum-Parsimony (MP) Analysis. A MP analysis of the protein data set with Felsenstein's protein parsimony algorithm (which neglects synonymous substitutions) showed monophyletic Archaea at BCL between 69% (with the full archaeal spectrum) and 83% (after deselecting 4B7) (Fig. 6). The extent to which the monophyletic Archaea tree in Fig. 6 is a significantly better representation of the "true" tree than trees showing a paraphyletic Archaea is given in Table 4, showing the differences in substitution number [Δ(sbst) and its SD] between the alternatives (Table 4). Similarly to ML, the monophyletic Archaea tree (tree 1) could not be confidently discriminated from tree 2 (supporting a crenarchaeal–eucaryal clade) and from tree 3 (supporting a euryarchaeal–eucaryal clade).

Conclusions

Objective criteria for circumventing ambiguities affecting multiple sequence alignments have been proposed in the recent past (Lake 1991; Zhu-Zy et al. 1992; Ellis and Morrison 1995; Gatesy et al. 1993; Wheeler 1994; Wheeler et al. 1995;). However, the possibility that certain alignment blocks may arise from the artifactual matching of insertion elements spanning the same structural space in the three domains of life has been overlooked. This situation is best exemplified by the matching of the archaeal–eucaryal and bacterial sequences nested between the two conserved motifs bordering the C region of the EF-G(2) alignment (Figs. 1 and 3 C). Some of the results in the present report provide an objective, generally applicable (albeit empyrical) criterion to identify blocks of spuriously matched sequences.

The phylogenetic results obtained from the new BLAST/FASTA-guided selection of the EF-G(2) alignment blocks render evidence for archaeal monophily less compelling (statistically) than previously thought (Creti et al. 1994). However, they do not convincingly support archaeal paraphyly as well. Alternative methods, and alternative (DNA and protein) data sets, support conflicting topologies, none of which is robust by bootstrap, and which cannot be discriminated by differences in log-likelihood (ML analysis) and number of inferred substitutions (MP analysis). Essentially identical conclusions are supported by phylogenetic trees of the two major components of the protein-targeting machinery [the 54-

Table 3. Phylogenetic relationships among Bacteria, Eucarya, crenarchaeotes, and euryarchaeotes by ML analysis of the EF-G(2) protein and nucleotide sequences[a]

	NucML (R = 0.8)		ProtML (JTT model)	
Tree topology	Δl_i	p_i	Δl_i	p_i
1.* (B,(MPT,H),(C,Ec))	−2.3 ± 4.5	.0215	(*−14230.1*)	.7350
2.* (B,(C,(MPT,H)),Ec)	(*−15085.8*)	.5230	−3.4 ± 5.0	.2320
3.* (B,C,((MPT,H),Ec))	−4.6 ± 3.3	.0040	−5.6 ± 4.0	.0100
4. (B,H,(MPT,(C,Ec)))	−9.1 ± 10.2	.0940	−15.5 ± 7.8	.0210
5. (B,MPT,(H,(C,Ec)))	−14.6 ± 9.2	.0400	−17.2 ± 7.2	.0010
6. (B,(H,(MPT,C)),Ec)	−8.9 ± 10.4	.1330	−35.4 ± 12.3	.0010
7. (B,(MPT,(C,H)),Ec)	−19.0 ± 7.5	.0000	−35.9 ± 12.1	.0000
8. (B,H,((C,MPT),Ec))	−14.8 ± 12.4	.0240	−41.7 ± 12.8	.0000
9. (B,H,(C,(MPT,Ec)))	−23.1 ± 11.0	.0000	−42.5 ± 12.4	.0000
10. (B,C,(MPT,(H,Ec)))	−25.1 ± 9.8	.0000	−42.5 ± 11.3	.0000
11. (B,C,(H,(MPT,Ec)))	−28.5 ± 8.7	.0000	−42.7 ± 11.1	.0000
12. (B,(C,MPT),(H,Ec))	−16.7 ± 12.0	.0020	−43.0 ± 12.7	.0000
13. (B,(C,H),(MPT,Ec))	−30.0 ± 9.4	.0000	−43.9 ± 12.4	.0000
14. (B,MPT,((C,H),Ec))	−27.1 ± 10.3	.0000	−43.9 ± 12.4	.0000
15. (B,MPT,(C,(H,Ec)))	−24.4 ± 11.0	.0010	−44.3 ± 12.3	.0000

[a] Δl_i is the difference of the log-likelihood of tree i from that of the maximum-likelihood tree (*italics* in parentheses) and ± is 1SE; p_i is the bootstrap probability for tree i being the ML tree among alternatives during bootstrap resampling estimated by RELL (Kishino et al. 1990). R is the transition/transversion rate parameter under the Kimura model of nucleotide substitution (Kumar et al. 1993; Wakeley 1996). B, Bacteria; C, crenarchaeotes, Ec, Eucarya; H, *Halobacterium;* MPT, *Methanococcus–Pyrococcus–Thermoplasma* cluster. The topologies shown are the 15 possible trees generated (MOLPHY) from the 20 OTUs organized into five topological elements [contrained tree {((((Mlu, Eco), (Tth,Spl)), Apy), Tma), (Gla, ((Ehy,Ddi), (Cke, (Ham,Dme)))), (((Sso,Sac), Dmo),4B7), ((Pwo,Tac), Mva), Hha}; ProtML analysis] and [constrained tree {((((Mlu,Eco), (Tth,Spl)), Apy), Tma), (Gla, (Cke, ((Ehy,Ddi), (Ham,Dme)))), (((Sso,Sac), Dmo), 4B7), ((Pwo,Tac), Mva), Hha} NucML analysis]. Asterisks indicate the three principal competing topologies: tree 1 is a paraphyletic Archaea tree showing a crenarchaeal–eucaryal clade [Eocyte tree of Rivera and Lake (1992)], tree 2 is the classical “archaebacterial tree, and tree 3 is a paraphyletic Archaea tree showing a euryarchaeal–eucaryal clade. The subtotal of the bootstrap probabilities of the trees supporting monophyletic Archaea (trees 2 and 6) in the NucML analysis is 0.67, while that of the trees supporting the sisterhood of crenarchaeotes and Eucarya in the ProtML analysis (trees 1, 4, 5) is 0.78.

kda signal recognition particle SRP54(Ffh) and the paralogous SRP-receptor protein SRα(Ftsy)] (Gribaldo and Cammarano 1998). Neither the monophily nor the paraphyly of Archaea with respect to Eucarya can be asserted with certainty from the SRα(Ftsy) and SRP54(Ffh) analyses. All the more important, neither the individual, nor the concatenated [SRP54(Ffh)-SRa(Ftsy)] paralogous proteins (totaling 440 positions) show the crenarchaeotes as a sister branch to Eucarya (Gribaldo and Cammarano, 1998); if anything, some of the results indicate the euryarchaeotes, instead of the crenarchaeotes, as a sister branch to Eucarya. The possibility should in fact be contemplated that the Archaea monophyly vs paraphyly issue is undecidable on the basis of single gene analyses.

Furthermore, the discovery of the Korarchaeota, a group of as yet uncultivated hyperthermophilic Archaea (likely) predating the bifurcation between euryarchaeotes and crenarchaeotes in the 16S RNA-based phylogenies (Barns et al. 1994, 1996), renders less likely the possibility that the Eucarya form a monophyletic grouping with (and are ancestrally related to) the crenarchaeotes, as this would require moving the bacterial branch (i.e., the root of the archaeal–eucaryal clade) across two nodes instead of only one (Barns et al. 1996).

Phylogenetically relevant to the question of archaeal monophily is the distribution of the EF-G(2) insertions among the three major taxa. All of these overwhelmingly support an archaeal–eucaryal clade, and none is found supporting a crenarchaeal–eucaryal clade. Based on the alignments in Figs. 1 and 2, only one major insertion (the 120-residue G″ subdomain) has been accreted to EF-2 in eucaryal evolution. All of the other discrete insertions that distinguish eucaryal EF-2 from EF-G are systematically common to Eucarya and Archaea of *all* known orders and genera. These include (Figs. 1–3) (i) the G′ subdomain of EF-2, in which the archaeal and eucaryal sequences are uniquely related by the motif [I,V]XXVNX[I,V]XX[Y,M]; (ii) a highly conserved archaeal–eucaryal insertion (AQKYR) immediately preceding the start of domain II (*Methanococcus* residues 262–266 in Fig. 1); (iii) the 72–77 EF-2 residues spanning the B region; (iv) the 23 EF-2 residues comprising the C regions and (v) the C-terminal accretion (boxed in Fig. 1) harboring the archaeal–eucaryal consensus element [I,T]RXRKGL. No discrete or short insertions or signatures unique to Eucarya and crenarchaeotes are detectable in the EF-G(2) sequence alignment, although these would be expected to occur if the two groupings were sister taxa, i.e., crenarchaeotes arose in evolution after the divergence of the methanogen–halophile (euryarchaeal) lineage. To our knowledge, the only element

Table 4. Phylogenetic relationships among Bacteria, Eucarya, crenarchaeotes, and euryarchaeotes by MP analysis of the EF-G(2) sequences[a]

Tree topology	Δ(*sbst*)	Significantly worse
1.* (B,(C,(MPT,H)),Ec)	(*3467*) ← best	
2.* (B,(MPT,H),(C,Ec))	+8.0 ± 6.0060	No
3.* (B,C,((MPT,H),Ec))	+10.0 ± 6.0059	No
4. (B,(H,(MPT,C)),Ec)	+18.0 ± 6.9350	Yes
5. (B,(MPT,(C,H)),Ec)	+25.0 ± 6.7147	Yes
6. (B,H,(MPT,(C,Ec)))	+33.0 ± 9.7563	Yes
7. (B,C,(MPT,(H,Ec)))	+34.0 ± 9.8075	Yes
8. (B,(C,MPT),(H,Ec))	+36.0 ± 10.4024	Yes
9. (B,MPT,(H,(C,Ec)))	+38.0 ± 9.1740	Yes
10. (B,C,(H,(MPT,Ec)))	+38.0 ± 9.3899	Yes
11. (B,H,((C,MPT),Ec))	+41.0 ± 10.2568	Yes
12. (B,MPT,(C,(H,Ec)))	+43.0 ± 9.9594	Yes
13. (B,H,(C,(MPT,Ec)))	+43.0 ± 10.0595	Yes
14. (B,(C,H),(MPT,Ec))	+45.0 ± 9.6528	Yes
15. (B,MPT,((C,H),Ec))	+49.0 ± 9.4429	Yes

[a] Δ(*sbst*) is the difference in substitution numbers of alternative trees from the MP tree in Fig. 6 (*italics* in parentheses) and ± is 1 SD; tree *i* is declared significantly different from the MP tree if Δ(*sbst*) is no less than 1.96 times greater than its SD (Felsenstein 1993). Alternative trees were generated by MOLPHY starting from the constrained treefile {(((((Mlu,Eco), Spl, Tth), Apy), Tma), (Gla, (Cke, ((Ehy,Ddi), (Ham, Dme)))), (((Sso,Sac), Dmo), 4B7), ((Pwo,Tac), Mva), Hha}, based on the groupings in Fig. 6. Abbreviations are as in Table 3, footnote a. Asterisks indicate the three principal conflicting topologies (see Table 3, footnote a).

specifically supporting a paraphyletic Archaea is a putative EF-1α insertion (EFEAGISKDG and variants thereof) linking specifically crenarchaeotes and Eucarya (Rivera and Lake 1992); it is not clear, however, whether this element could have been lost by the euryarchaeotes which harbor, in the same structural space, the sequence GE (*T. acidophilum*) and AKS (*M. vannielii*) (see Fig. 1 of Baldauf et al. 1996).

Acknowledgments. This research was supported by grants from MURST Project Nucleic acid–Protein interaction and from CNR Projecto f'inalizzato Biotecnologie. We are indebted to M. Hasegawa for generous guidance in using the MOLPHY programs.

References

Ævarsson A (1995) Structure based sequence alignment of elongation factors Tu and G with related GTPases involved in translation. J Mol Evol 41:1096–1104

Ævarsson A, Brazhnikov E, Garber M, et al. (1994) Three-dimensional structure of the ribosomal translocase: Elongation factor G from *Thermus thermophilus.* EMBO J 13:3669–3677

Adachi, J (1995) Modeling of molecular evolution and maximum likelihood inference of molecular phylogeny, PhD dissertation, Graduate University for Advanced Studies, Tokyo

Adachi J, Hasegawa M (1992) MOLPHY: Programs for molecular phylogenetics. I. PROTML: Maximum likelihood inference of protein phylogeny. Computer Science Monographs, No. 27, Institute of Statistical Mathematics, Tokyo, Vol 1

Altschul SF, Gish G, Miller W, Myers E, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215:403–410

Baldauf SL, Palmer JD, Doolittle WF (1996) The root of the universal tree and the origin of cucaryotes based on the elongation factor phylogeny. Proc Natl Acad Sci USA 93:7749–7754

Barns SM, Fundyga RE, Jeffries MW, Pace NR (1994) Remarkable archaeal diversity detected in a Yellowstone National Park hot spring environment. Proc Natl Acad Sci USA 91:1609–1613

Barns SM, Delwiche CF, Palmer JD, Pace NR (1996) Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. Proc Natl Acad Sci USA 93:9188–9183

Bourne HR, Sanders DA, McCormick F (1990) The GTPase superfamily: A conserved switch for diverse cell functions. Nature 348:125–132

Bourne HR, Sanders DA, McCormick F (1991) The GTPase superfamily: conserved structure and molecular mechanism. Nature 349: 117–127

Cammarano P, Palm P, Creti R, Ceccarelli E, Sanangelantoni AM, Tiboni O (1992) Early evolutionary relationship among known life forms inferred from elongation factor EF-2/EF-G sequences—Phylogenetic coherence and structure of the archaeal domain. J Mol Evol 34:396–405

Corpet F (1988) Multiple sequence alignment with hierarchical clustering. Nucleic Acids Res 16:10881–10890

Creti R, Ceccarelli E, Bocchetta M, et al. (1994) Evolution of translational elongation factor (EF) sequences. Reliability of global phylogenies inferred from EF-1α(Tu) and EF-G sequences. Proc Natl Acad Sci USA 91:3255–3259

Dever TE, Glynias MJ, Merrick WC (1987) GTP-binding domain: Three consensus sequence elements with distinct spacing. Proc Natl Acad Sci USA 84:1814–1818

Deveraux J, Haeberli P, Smithies O (1984) A comprehensive set of sequence analysis program for the VAX. Nucleic Acids Res 12: 387–395

Felsenstein J (1989) PHYLIP—phylogeny inference package. Cladistics 5:164–166

Felsenstein J (1993) PHYLIP (phylogeny inference package), version 3.5c. Department of Genetics, University of Washington, Seattle

Gish W, States DJ (1993) Identification of protein regions by database similarity search Nature Genet 3:226–272

Gribaldo S, Cammarano P (1998) The root of the universal tree of life inferred from aciently duplicated genes encoding components of the protein-targeting machinery. J Mol Evol 47:508–516

Hashimoto T, Hasegawa M (1996) Origin and early evolution of eucaryotes inferred from the aminoacid sequences of translation elongation factors 1α/Tu and 2/G. Adv Biophys 32:73–120

Iwabe N, Kuma K, Hasegawa M, Osawa S, Miyata T (1989) Evolutionary relationship of archaebacteria, eubacteria and eukaryotes inferred from phylogenetic trees of duplicated genes. Proc Natl Acad Sci USA 86:9355–9359

Kessel M, Klink F (1980) Archaebacterial elongation factor is ADP-ribosylated by Diphteria toxin. Nature 287:250–251

Kishino H, Hasegawa M (1989) Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. J Mol Evol 29:170–179

Kishino H, Myiata T, Hasegawa M (1990) Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. J Mol Evol 30:151–160

Kjelgaard M, Nyborg J (1992) Refined structure of elongation factor EF-Tu from *Escherichia coli.* J Mol Biol 223:721–742

Kumar S, Tamura K, Nei M (1993) MEGA: Molecular evolutionary genetic analysis, version 101, Pennsylvania State University, University Park

Lechner K, Heller G, Böck A (1988) Gene for the diphtheria toxin-susceptible elongation factor 2 from *Methanococcus vannielii.* Nucleic Acids Res 16:7817–782

Pearson WR Lipman DJ (1988) Improved tools for biological sequence comparisons. Proc Natl Acad Sci USA 85:2444–244

Rivera MC and Lake JA (1992) Evidence that Eukaryotes and Eocyte prokaryotes are immediate relatives. Science 257:74–76

Stein JL, Marsh TL, Wu KY, Shizuka H, DeLong EF (1996) Characterization of uncultivated prokaryotes: Isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. J Bact 178:591–599

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight-matrix choice. Nucleic Acids Res 22:4673–4680

Wadell, PJ (1995) Statistical methods of phylogenetic analysis: Including Hadamard conjugations, LogDet transforms, and maximum likelihood, PhD dissertation. Massey University, New Zealand

Wakeley J (1996) The excess of transitions among nucleotide substitutions: new methods of estimating transition bias underscore its significance. Tree 11:158–163