J Mol Evol (2000) 50:366–380
DOI: 10.1007/s002399910040

JOURNAL OF MOLECULAR EVOLUTION

Phylogenetic Depth of the Bacterial Genera *Aquifex* and *Thermotoga* Inferred from Analysis of Ribosomal Protein, Elongation Factor, and RNA Polymerase Subunit Sequences

Maurizio Bocchetta,[1] Simonetta Gribaldo,[1] Annamaria Sanangelantoni,[2] Piero Cammarano[1]

[1] Istituto Pasteur Fondazione Cenci Bolognetti, Dipartimento di Biotecnologie Cellulari ed Ematologia, Sezione di Genetica Molecolare, Università di Roma I "La Sapienza," Policlinico Umberto I°, Viale Regina Elena 324, 00161 Roma, Italy
[2] Dipartimento di Scienze Ambientali, Università di Parma, Area delle Scienze, 43100 Parma, Italy

Received: 22 September 1999 / Accepted: 11 January 2000

Abstract. The phylogenetic placement of the *Aquifex* and *Thermotoga* lineages has been inferred from (i) the concatenated ribosomal proteins S10, L3, L4, L23, L2, S19, L22, and S3 encoded in the S10 operon (833 aa positions); (ii) the joint sequences of the elongation factors Tu(1α) and G(2) coded by the *str* operon *tuf* and *fus* genes (733 aa positions); and (iii) the joint RNA polymerase β- and β'-type subunits encoded in the *rpoBC* operon (1130 aa positions). Phylogenies of *r*-protein and EF sequences support with moderate (*r*-proteins) to high statistical confidence (EFs) the placement of the two hyperthermophiles at the base of the bacterial clade in agreement with phylogenies of rRNA sequences. In the more robust EF-based phylogenies, the branching of *Aquifex* and *Thermotoga* below the successive bacterial lineages is given at bootstrap proportions of 82% (maximum likelihood; ML) and 85% (maximum parsimony; MP), in contrast to the trees inferred from the separate EF-Tu(1α) and EF-G(2) data sets, which lack both resolution and statistical robustness. In the EF analysis MP outperforms ML in discriminating (at the 0.05 level) trees having *A. pyrophilus* and *T. maritima* as the most basal lineages from competing alternatives that have (i) mesophiles, or the *Thermus* genus, as the deepest bacterial radiation and (ii) a monophyletic *A. pyrophilus–T. maritima* cluster situated at the base of the bacterial clade. RNAP-based phylogenies are equivocal with respect to the *Aquifex* and *Thermotoga* placements. The two hyperthermophiles fall basal to all other bacterial phyla when potential artifacts contributed by the compositionally biased and fast-evolving *Mycoplasma genitalium* and *Mycoplasma pneumoniae* sequences are eschewed. However, the branching order of the phyla is tenuously supported in ML trees inferred by the exhaustive search method and is unresolved in ML trees inferred by the quartet puzzling algorithm. A rooting of the RNA polymerase-subunit tree at the mycoplasma level seen in both the MP trees and the ML trees reconstructed with suboptimal amino acid substitution models is not supported by the EF-based phylogenies which robustly affiliate mycoplasmas with low-G+C gram-positives and, most probably, reflects a "long branch attraction" artifact.

Key words: Bacterial rooting — Hyperthermophily — Ribosomal proteins — Elongation factors — RNA polymerase — *Mycoplasmatales* — S10 operon — Streptomycin operon — *rpoBC* operon

Introduction

In phylogenies of 16S rRNA sequences the hyperthermophilic Bacteria *Aquifex pyrophilus* and *Thermotoga maritima* branch off—in that order—from the main trunk of the bacterial tree before moderately thermophilic and

Correspondence to: Piero Cammarano; *e-mail:* cammarano@bce.med.uniroma1.it

mesophilic phyla. Nonetheless, the phylogenetic depth of the *Aquifex–Thermotoga* pair remains controversial, as protein-coding genes either tenuously support the rooting of Bacteria at the *A. pyrophilus* and *T. maritima* level or predict alternative placements of the two hyperthermophiles. Phylogenies of elongation factor (EF) G(2) and *fus* gene sequences confirm the *A. pyrophilus* and *T. maritima* placements inferred from analysis of 16S rRNA sequences, but with low statistical support (Bocchetta et al. 1995). In contrast, trees of RNA polymerase (RNAP) β- and β′-type subunits show *Mycoplasmatales* as the deepest bacterial grouping, with *Aquifex pyrophilus* and *T. maritima* branching off in the proximity of *Spirochaetes* and *Proteobacteria* (Klenk et al. 1999). Furthermore, an affiliation of *T. maritima* with gram-positive Bacteria has been invoked to explain (i) the anomalous clustering of Euryarchaeotes, *T. maritima*, and low-G+C gram-positives in phylogenetic trees of glutamine synthetase1 (GSI) sequences (Tiboni et al. 1993) and (ii) the observation (Gribaldo et al. 1999) that the *T. maritima* 70-kDa heat shock protein (Hsp70) resembles that of gram-positive Bacteria and Archaea in lacking a relatively conserved insert (22–23 residues) that occurs in the same position in the Hsp70s of all other organisms (Macario et al. 1991; Gupta et al. 1997, 1998).

In a continuing effort to clarify the evolutionary placement of hyperthermophilic Bacteria, we have reconstructed outgroup-rooted bacterial phylogenies by using three sets of concatenated proteins. One spans a subset of ribosomal proteins (*r*-proteins) encoded in the "S10 operon"-equivalent gene cluster. The second encompasses the two EFs encoded by the *fus* [for EF-G(2)] and *tuf* [for EF-Tu(1α)] genes of the streptomycin (*str*) operon, which lies immediately upstream from (or is fused with) the S10 operon gene cluster. The last comprises the RNAP β- and β′-type subunit sequences encoded in the *rpoBC* operon, which, in most Archaea and Bacteria, is situated shortly upstream from the *str* operon. Phylogenetic evaluation of EF and *r*-protein data sets resulted in gene trees which are in moderate (*r*-proteins) and excellent (EFs) overall agreement with 16 S rRNA trees with respect to the *Aquifex* and *Thermotoga* placements. On the other hand, phylogenies of RNAP-subunit sequences give equivocal results concerning the branching order of the bacterial phyla.

Methods

Sequence Retrieval and Selection of Alignment Positions. Protein sequences were retrieved by BLAST (Altschul et al. 1990) and FASTA (Pearson et al. 1988), probing the DNA and protein databases with the tBLASTN program using the GCG (Genetic Computer Group) program suite (Deveraux et al. 1984) of the MRC Human Genome Mapping Project (HGMP) Resource Centre (Cambridge University, Cambridge, UK).

Alignments. Multiple alignments of amino acid sequences were performed using the programs CLUSTAL W (Thompson et al. 1994) and MULTALIN (Corpet 1988) with default gap penalties. Spurious matches of bacterial sequences with unrelated segments of archaeal and eucaryal homologues were identified by a BLAST-guided scrutiny of all regions of the CLUSTAL and MULTALIN alignments as detailed elsewhere (Cammarano et al. 1999). In essence, the matching schemes of bacterial and archaeal–eukaryal sequences generated by the automatic multialignment programs were sought in three inventories of significant binary alignments obtained by BLAST probing the protein and DNA databases with query sequences representative of the three domains of life. Alignment schemes that were not retrieved in any of the three inventories were assumed to represent artifacts produced by the multialignment algorithms (Ævarsson, 1995) and the positions comprising the spurious matches were excluded from the final data sets.

The final multiple alignments, the global alignments of the individual proteins, and the accession numbers of all the sequences used in this work are retrievable upon request to P. Cammarano. The *A. pyrophilus* S10 operon and spectinomycin (*spc*) operon sequences have been deposited in the EMBL/GenBank and have been assigned accession numbers AF040100 and AF040101, respectively. The corresponding *T. maritima* S10 and *spc* operon sequences have been assigned accession number Z21677.

Tree-Making Algorithms. Phylogenetic trees were constructed using maximum-likelihood (ML) and maximum-parsimony (MP) methods. ML analyses utilized the ProtML program of the MOLPHY (Molecular Phylogenetics) software package version 2.2 (Adachi and Hasegawa, 1992) and the quartet puzzling (QP) algorithm implemented in the program PUZZLE version 4.0 (Strimmer and von Haeseler 1996, 1997).

When ProtML was used, candidate topologies, selected by the approximate log-likelihood criterion (Adachi 1995; Wadell 1995) from the exhaustive search (option −e) of a partially constrained starting tree, were examined for the best tree by the exact likelihood method with the "user" (-u) option of ProtML. The following models of amino acid substitution were used in the ProtML analyses: (1) the Jones–Taylor–Thornton (JTT) model (default option), (2) the JTT-F model (option -jf), (3) the Dayhoff model (option -d), (4) the Poisson model (option -p), and (5) the proportional model (option -pf), which corresponds to the F option of the Poisson model. When the F option was invoked, the models were reconstructed by using the actual amino acid compositions of the proteins under analysis as the equilibrium frequencies. The adequacy of different models was evaluated by the Akaike information criterion (AIC) defined as AIC $= -2$ (log-likelihood)+2N, where N is the number of free parameters (Hashimoto and Hasegawa 1996). The model that minimized AIC was considered to be the most appropriate (Hashimoto and Hasegawa 1996). The relative bootstrap probabilities of alternative topologies were computed with the RELL (resampling of estimated log-likelihood) bootstrap method (Kishino and Hasegawa 1989; Kishino et al. 1990) with the user option of ProtML. Standard errors of log-likelihood differences were estimated by Eq. (12) of Kishino and Hasegawa (1989). Bootstrap values for internal tree branches were calculated as the sum of the bootstrap probabilities of all the trees showing the node being analyzed among the alternatives (945 trees for the *r*-protein data set and 1000 top-ranking trees for the EF and RNAP data sets).

When the QP algorithm was used, phylogenetic trees were reconstructed with eight categories of site-by-site rate variation using the JTT-F, Dayhoff-F, and Blosum 62-F models of amino acid substitution.

MP analyses used the PROTPARS program of the Phylogeny Inference Package (PHYLIP) version 3.57c (Felsenstein, 1993). The PHYLIP programs SEQBOOT, PROTPARS, and CONSENSE were used (in that order) to derive a MP tree which was replicated in 100 bootstraps. Alternative trees were declared significantly "worse" than the MP tree, at the 0.05 level, when the mean of Δ_{sbst} (the difference in number of inferred substitutions) was ≥1.96 SD (Felsenstein 1993).

Estimation of Compositional Biases and Invariant Sites. To identify compositionally biased taxa that might distort the tree topologies (Lake 1994; Lento et al. 1995; Foster and Hickey 1999), the fitting of the amino acid compositions of individual sequences to the frequency distribution in the assumed ML models was evaluated by use of a 5% χ^2 analysis implemented in PUZZLE Version 4.0.

Fractions of constant and invariant sites (Lockhart et al. 1996) were inferred by using a two-site rate heterogeneity (either variable or invariant) ML model, with the JTT-F or the Blosum 62 matrixes in PUZZLE version 4.0, and the effect of their exclusion on the tree topology was evaluated with both the ProtML and the QP methods.

Species Abbreviations. Aae (*Aquifex aeolicus*), Apy (*Aquifex pyrophilus*), Ani (*Anacystis nidulans*), Atu (*Agrobacter tumefaciens*), Bbu (*Borrelia burgdorferi*), Bst (*Bacillus stearothermophilus*), Bsu (*Bacillus subtilis*), Cpa (*Cyanophora paradoxa* cyanelles), chl (chloroplast), Ctr (*Chlamydia trachomatis*), Eco (*Escherichia coli*), Hma (*Halobacterium marismortui*), Hha (*Halobacterium halobium*), Hin (*Haemophilus influenzae*), Mja (*Mehanococcus jannaschii*), Mca (*Mycoplasma capricolum*), Mge (*Mycoplasma genitalium*), Mpn (*Mycoplasma pneumoniae*), Mle (*Mycobacterium leprae*), Mtu (*Mycobacterium tubercolosis*), Mva (*Methanococcus vannielii*), Ppu (*Pseudomonas putida*), Pho (*Pyrococcus horikoshi*), Pwo (*Pyrococcus woesei*), Rpr (*Rickettsia prowazeckii*), Sau (*Staphylococcus aureus*), Spl (*Spirulina platensis*), Sac (*Sulfolobus acidocaldarius*), Sso (*Sulfolobus solfataricus*), Syn (*Synechocystis* sp.), Tac (*Thermoplasma acidophilum*), Tce (*Thermococcus celer*), Tma (*Thermotoga maritima*), Tth (*Thermus thermophilus*).

Results

Order and Linkage of the A. pyrophilus *and* T. maritima *S10- and* str-*Operon Genes*

We have previously cloned and sequenced the *T. maritima* DNA region comprising the S10 operon *r*-proteins gene cluster (Sanangelantoni et al. 1994). Sequencing of the corresponding region of *A. pyrophilus* DNA (4857 bp) revealed the same conserved gene order as *T. maritima*, gram-positive bacteria, Proteobacteria, and certain Archaea (Sanangelantoni et al. 1994), i.e., eight tightly linked genes encoding ribosomal proteins S10 (104 aa), L3 (244 aa), L4 (200 aa), L23 (109 aa), L2 (280 aa), S19 (219 aa), L22 (133 aa), and S3 (231 aa), in that order, with a four bases overlap (ATGA) at the L3/L4 genes junction. The lengths of the *r*-proteins were within the bacterial range (Sanangelantoni et al. 1994) with the exception of protein S19, which was about twice that of other Bacteria (87–95 aa) due to an N-terminal accretion (128 aa) also observed in the homologous sequence of *A. aeolicus* (Deckert et al. 1998). The *A. pyrophilus r*-protein gene cluster was situated immediately downstream (12-bp distance) from the stop codon of the second gene (*tuf*) of an amputated streptomycin (*str*) operon comprising only the genes *fus* (for EF-G), and *tuf* (for EF-Tu) instead of the canonical gene sequence 5′-*rps*12–*rps*7–*fus*–*tuf*-3′ (Bocchetta et al. 1995). A potential promoter for S10 operon was apparent shortly 5′ to the *fus* gene stop codon (Bocchetta et al. 1995) , at a suitable distance from the downstream located *rps*10 gene.

To assess the phylogenetic placement of hyperthermophilic Bacteria, the *A. pyrophilus* and *T. maritima* proteins encoded in the contiguous S10 and *str* operons were included into two global alignments: one spanning the concatenated S10 operon proteins and the other encompassing the EF-G(2) and EF-Tu(1α) sequences. A third alignment inclusive of *Aquificales* (*Aquifex aeolicus*) and *T. maritima* was generated from the RNA polymerase β- and β′-subunits (and their archaeal B- and A-type homologues) encoded in the *rpoBC* gene cluster, which, in most Archaea and Bacteria, lies shortly upstream from the *str* operon (Zillig et al. 1993) with the notable exception of the *Aquifex* genus (Deckert et al. 1998).

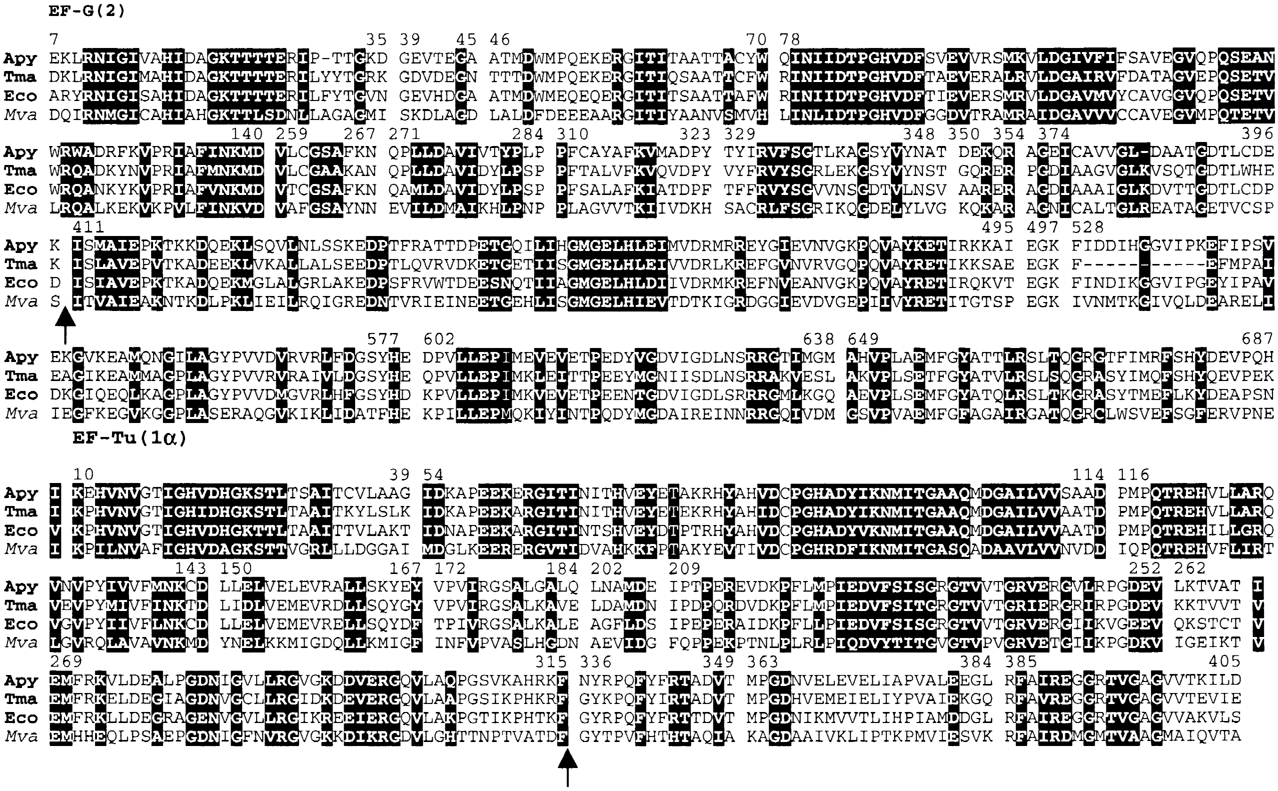
Protein Data Sets

The "S-10 operon" *r*-protein data set was assembled from preliminary alignments of the individual *A. pyrophilus* and *T. maritima r*-proteins with all available homologues. Sections of the global alignments comprising artifactual matches of bacterial sequences with unrelated segments of the (generally longer) archaeal–eukaryal homologues were identified by the BLAST-guided scrutiny (see Methods) and the remaining regions (totaling 830 positions) were linked together. Exclusion of species having an incomplete repertory of sequenced S10-operon genes left 16 taxa, of which 3 were from Archaea (O for outgroup) and 13 were from Bacteria representing Aquificales (A), Thermotogales (T), Cyanobacteria (C), low-G+C gram-positives (lG+), high-G+C gram-positives (hG+), and Proteobacteria of the γ subdivision (Pγ). An abridged version of the final alignment is shown in Fig. 1.

A data set of concatenated EF sequences was constructed by adapting the *T. maritima* and *A. pyrophilus* EF-G (682–700 aa) and EF-Tu sequences (400–405 aa) to existing global alignments generated by using tertiary structural data (Ævarsson 1994, 1995; Baldauf et al.

Fig. 1. Abridged multiple alignment of ribosomal proteins S3, L22, S19, L2, L23, L4, L3, and S10, in that order. *Uppercase roman numerals* designate consecutive alignment blocks belonging to the same protein as selected by the BLAST-guided exclusion of regions comprising misaligned segments (Cammarano et al. 1999). *Numbers in parentheses* indicate the sequence positions (*A. pyrophilus* numbering) comprising each block. The protein L3 blocks II* and IV** span *A. pyrophilus* residues 106–110 and 196–202, respectively. Species abbreviations are as listed under Methods. *Highlighting on black background* delimits sites occupied by identical or similar amino acids (ILVM; DEKRH; FWY, ST GA, NQ) in no fewer than 14 of 16 sequences. *Gray shading* highlights bacterial signatures. The chloroplast sequence was compounded from *Euglena gracilis* and *Marchantia polymorpha* chloroplasts.

```
    S3        I(27-59)                II (63-94)                       III (110-175)
Apy KDYAKLLHEDLKIKNYIKKRYKVAGVSKVEIER KVRIKIHTAKPAIVIGRRGQEVDRLKKTIERM EVKVPELDAQLVAEDIALQIERRVSHRRAMKRA
Tma KNYKEWLLEDEEIRKIIKNKYYHAGISEIYVER RINITVKTARPGIIIGRKGSEITSLREELERK EIKTPELDAQLVAESIASRIEKRASYKVAMKRA
Bsu KDYADFLHEDLKIREYISKRLSDASVSKVEIER RVNITIHTAKPGMVIGKGGSEVEALRKALNSL EIKRADLDAQLVADNIARQLENRVSFRRAQKQQ
Mtu KQYAEYVKEDVAIRRLLSSGLERAGIADVEIER RVRVDIHTARPGIVIGRRGTEADRIRADLEKL EVKNPESQAQLVAQGVAEQLSNRVAFRRAMRKA
Eco KEFADNLDSDFKVRQYLTKELAKASVSRIVIER SIRVTIHTARPGIVIGKKGEDVEKLRKVVADI EVRKPELDAKLVADSITSQLERRVMFRRAMKRA
Syn KRYPELLQEDHKIRQYIEKTLNNAGISDIRIER QIELGIHTARPGVVVGRGGSGIEQLREGLQKL EVPNADADAALMAEYIGQQLERRVSFRRVVRQA
Hma QQFIEDGLQRTQIDEFFAEELGRAGYGGMDVAK GTQIVLKAEKPGMVIGKGGKNIRKITTELEDR EVDEPDLNARIVADRLANALERGWYFRKAGHTT
Mja RTFVKENVKRLLIDEYFKKELSKAGYSHCDIRK GTKIIIYAEKPGFVIGRRGSRIRELTETLAKE PVENPDLDAQVVAQKVAQSLERGLHFRRVGHTA

                                         IV (179-205)                L22   I (20-44)            II(55-74)
Apy IDNALKAGQKGVKVQVKGRIGGAARARKEWFLVGRM TLRADIDYGFATAYTKYGILSVKVWIY AILRYAHISPLKARLVLREIHGKDV PKRAARIAE
Tma IMNAMRKGAQGIKVMVAGRLGGAEIARREWYLRGRL KIKAIIDYGTATAWTKYGTIGIKVWIY AVAKYIRISPRKARAIANTIRGKSV PKKAARIME
Bsu IQRTMRAGAQGVKTMVSGRLGGADIARPEYYSEGTV TLRADIDYATSEADTTYGKLGVKVWIY AVARTVRIAPRKARLVMDLIRGKQV PRAASPIIE
Mtu IQSAMRQPNKGIRVQCSGRLGGAEMSRSEFYREGRV TLRADIDYGLYEAKTTFGRIGVKVWIY AKARFVRVSPRKARRVIDLVRGRSV PQAASGPVA
Eco VQNAMRLGAKGIKVEVSGRLGGAEIARTEWYREGRV TLRADIDYNTSEAHTTYGVIGVKVWIF AKHRHARSSAQKVRLVADLIRGKKV NKKAAVLVK
Syn LQRAERAEVKGIKIQVSGRLNGAEIARTEWVREGRV TLRADIDYAYRTALTTYGILGIKVWIF AIARYVRMSPLKVRRVLDQIRGRSY PYKACEPVL
Hma IDRIMESGALGAEIVLSGKVTG-ARSRVEKFNRGYV PAEEIVDSGVGVAVMKLGTIGVRVKII AMLRERQMSFKHSKAIAREIKGKTA PEKASKAFL
Mja VRRVMNAGAKGVIVIISGKLTG-ERARTEKFMAGYM PAEELVDKGRAIAKTKPGVIGVTVKIM AMGRNIPISRKHAREICKSINGMKL PQKATEEIL

                    III(79-124)                                S19   I(26-84)
Apy KLLKSAIANAE GLDLDRLYIKKAVADRGPILKKWIPRAHGRATMVRKRLSHITIVLE KVVRTYSRATTIIPEFVGHTIAVHNGKTFIPVYITQDMVGH
Tma KVLKSRVANAE GLSVENLYVSECYVNDGPRMKRIWPRGRGRADIIQKRMSHITVVVR KVIKTWSRASMIIPKWVGHGIAVYNGMKHIPVYITENMIGH
Bsu KVLKSAIANAE EMDANNLVISQAFVDEGPTLKRFRPRAMGRASQINKRTSHITIVVS QVVKTWSRRSTIFPQFIGHTIAVYDGRKHVPVFISEDMVGH
Mtu KVIASAAANAQ GLDPATLVVATVYADQGPTAKRIRPRAQGRAFRIRRRTSHITVVVE QVIKTWSRRSTIIPDFIGHTFAVHDGRKHVPVFVTESMVGH
Eco KVLESAIANAE GADIDDLKVTKIFVDEGPSMKRIMPRAKGRADRILKRTSHITVVVS KPLRTWSRRSTIFPNMIGLTIAVHNGRQHVPVFVTDEMVGH
Syn KVLRSAVANAE GLEPADLVVSQAFADGGPSLRRFRPRAQGRAYQIRKPTCHITVAVA QVVKTWSRASTILPQMVGHTIAVHNGRQHVPVFVSEQMVGH
Hma DLLENAVGNAD GFDGEAMTIKHVAAHKVGEQQGRKPRAMGRASAWNSPQVDVELILE DPIRTHLRDMPVVPEMVGLTLAVHDGQNFERVKVEPEMLGH
Mja KVLDNAKKNAE GLNTEKLRIKHISTNKGITIKRYMPRAFGRATPKFQETVHIQVILE RIIRTHCRDFVITPDMVGLTFGVYNGKEFVEVKVTPEMIGH

                       L2  I (64-99)                        II (101-120)         III (133- 179)
Apy KLGEFAPTRTFKGHPDKY RIIDFERDKSLVPAKVVSIEYDPFRSARICLLHYAD EKRYIIWPEGLKVGDTVMSI EIKPGNAMPLKYIPEGTIIHNIE
Tma RLGEFAPTRRFGGHADKY RIIDFKRDKAGIPAKVLAIEYDPNRSARIALLLYAD EKRYILAPKGVNVGDTLMSG EIRPGNALPLEKIPVGTLVHNVE
Bsu KLGEFAPTRTYKGHASDY RVIDFKRDKDGIPGRVATVEYDPNRSANIALINYAD EKRYILAPKGIQVGTEIMSG DIKVGNALPLINIPVGTVVHNIE
Mtu KLGEFAPTRTFKGHIKDY RMIDFRRDKDGVNAKVAHIEYDPNRTARIALLHYLD EKRYIIAPNGLSQGDVVESG DIKPGNNLPLRNIPAGTLIHAVE
Eco KLGEFAPTRTYRGHAADY RIVDFKRNKDGIPAVVERLEYDPNRSANIALVLYKD ERRYILAPKGLKAGDQIQSG AIKPGNTLPMRNIPVGSTVHNVE
Syn KLGEFAPTRTFRSHSKSY RIIDFKRNKQNIPARVAAIEYDPNRNARIALLFYTD EKRYILAPAGLQVGMTVIAG PFEIGNTLPLSRIPLGSEIHNVE
Hma YLGEFQLTRSSVEHGQAH RKVEDGD---VIAGTVVDIEHDPARSAPVAAVEFED DRRLILAPEGVGVGDELQVG EIAPGNTLPLAEIPEGVPVCNVE
Mja YLGEFSLTRKPVQHGAPY RRFDELEKKGKVLGKIVDILHDPGRSAPVAKVEYET EEGLLVVPEGVKVGDIIECG EIKPGNILPLGAIPEGIPVFNIE

                             IV (186-223)                           V (230-250)           L23  I (8-24)
Apy FMPGKGGQIARAAGTWAQVLGRST VRMPSGEVRMIHERCMATIGRVGLAEHELVNVGKAGRA PHTRGTAMNPVDHPHGGGEGR EILIRPIITEKSNR
Tma FTPGKGGQIARAAGTYCQIMANEG LRMPSGELRKVHIKCYATVGVVGNEDHKKEVHGKAGRV PHVRGVAMNPVDHPHGGGEGR DVLIRPIITEKALI
Bsu LKPGKGGQLVRSAGTSAQVLGKEG VRLNSGEVRMILSACRASIGQVGNEQHELINIGKAGRS PTVRGSVMNPNDHPHGGGEGR DVLKRPVITERSAD
Mtu LRPGGGAKLARSAGSSIQLLGKEA LRMPSGEIRRVDVRCRATVGEVGNAEQANINWGKAGRM PSVRGVVMNPVDHPHGGGEGK DIILAPVISEKSYG
Eco MKPGKGGQLARSAGTYVQIVARDG LRLRSGEMRKVEADCRATLGEVGNAEHMLRVLGKAGAA PTVRGTVMNPVDHPHGGGEGR KVLRAPHVSEKAST
Syn LVAGRGGQMVRSAGAFAQVVAKEG IKLPSKEVRMVRKECVATLGRVSNAEFRNLKLGKAGRK PHVRGSVMNPCDHPHGGGEGR DLIIKPIVTEKATL
Hma SSPGDGGKFARASGVNAQLLTHDR VKLPSGEMKRLDPQCRATIGVVGGGGRTDKPFVKAGNK PNVRGVVMNAVDHPFGGGGRQ DVIKHPHVTEKAMN
Mja TVPGDGGKLVRAGGCYAHILTHDG VKLPSGHIKALHSMCRATIGVVAGGGRKEKPFVKAGKK PRVRGVVMNAVDHPFGGGRHQ DVIKAPVVTEKTVR

          II (28-61)                       III(82-89) L4 I(32-51)          II  (53-109)
Apy LME KYTFEVALDASKPEIKEAVEKLFNVKVKKVNTMI WKKAIVTL KRRQGTHSTKTRGEVAYSGR KILPQKGTGNARHGERGVNIFVGGGVAHGPK
Tma LRE KYVFEVNPLANKNLVKEAVEKLFNVKVEKVNILN WKKAVVTL NRRAGTASTKTRGEVSGGGR KPWPQKHTGRASHGSIRSPIWRHGGVVHGPK
Bsu LMT KYTFEVDVRANKTEAKDAVESIFGVKVDKVNIMN RRKAIVKL SLRQGTHKVKNRSEVRGGGR KPWRQKGTGRARQGSIRSPQWRGGGVVFGPT
Mtu LLD VYTFLVRPDSNKTQIKIAVEKIFAVKVASVNTAN TKRAIVTL AARQGTHSTKTRGEVSGGGR KPYRQKGTGRARQGSTRAPQFTGGGVVHGPK
Eco AME TIVLKVAKDATKAEIKAAVQKLFEVEVEVVNTLV WKKAYVTL GARQGTRAQKTRAEVTGSGK KPWRQKGTGRARSGSIKSPIWRSGGVTFAAR
Syn QLE KYVFDVRPEATKPEIKAAIELLFDVKVTGVNTAR VKRAVVTL NNRQGNASAKTRAEVRGGGR KPWKQKGTGRARAGSIRSPLWRGGGVIFGPK
Hma DMD KLQFAVDDRASKGEVADAVEEQYDVTVEQVNTQN EKKAVVRL NRKQGSDEYATPAESFGSGR QAHVPKQDGRAR----RVPQAVKGRSAHPPK
Mja MIE KLVFYVDRRATKQDIKRAMKELFDVEVEKVNTLI EKKAYVKL ARLQGSDPLATSAKNIGKGH RARVDRVPGWAA----RVPQAVGGRRAHPPK

                               III (125-139)   IV(147-194)                                      L3 I
Apy PRDYEYPLPKKVRKLGLKMALSDKAQ KTKKAVEFLKNLGVD VIPEKNEVLYKSFRNLQNVRVLLPEGLKLYDVLWANKLVIHKECLDRI GLIGEKV
Tma PRDWSKKLNKKMKKLALRSALSVKYR KTKSLKEILQNLQLS WKEEGYMNVKLSGRNLPDVKVIIADGLNVFDMLKYDYLVLTRDMVSKI MIIGRKV
Bsu PRSYSYKLPKKVRRLAIKSVLSSKVI KTKEMAAILKGLSVE VTADANEAVALSARNIPGVTVVEANGINVLDVVNHEKLLITKAAVEKV ------I
Mtu PRDYSQRTPKKMIAAALRGALSDRAR STKSARAFLASLTER VIGRSDEAGAKSVRNLPGVHILAPDQLNTYDVLRADDVVFSVEALNAY GILGTK-
Eco PQDHSQKVNKKMYRGALKSILSELVR KTKLLAQKLKDMALE ITGELDENLFLAARNLHKVDVRDATGIDPVSLIAFDKVVMTADAVKQV GLVGKKL
Syn PRDYSQKMNRKERRLALRTAIASRAD KTKELATALTRWGAK ILDEIPENVFLSGRNIPYLKILRADNLNIYDVLVADTIVATATALEKI GILGTKV
Hma TEKRSLDLNDKERQLAVRSALAATAD KTQEVVSLLEALDVH IKAGQGSARGRKYRRPASILFVTSDNLAGADVATASEVNTEDLAPGGA GFAGYKL
Mja VEKLWERVNKKERIKAIKSAIAATAN KTKDVFAVFEKLGIS DRAGKGKMRGRKYKKPRSILVVVGDNLPGVDVITAKDLGIIHLAPGGV AFPVYKA

    (4-19)     II*   III (112-143)                    IV**     V (205-222)        S10     I (5-56)
Apy GMTRVLLKD QELRV DVFKPGDLVDVWGISKGRGFAGVMKRWDFAGF ENAILVKG ALPGHNKGIVVLFPAVER KIRIKLKAFDHRVLDQSVKQII
Tma GMTRVFVGN QVIKV DVFEKGELVDVTGWTKGRGFAGAMKRWGFSGG NDLLVVKG GVPGARGGLVLIRSAKAP KIRIKLKAYDHELLDESAKKIV
Bsu --------- QEVKV EIFSAGEIVDVTGVSKGKGFQGAIKRHGQSRG RNLLLIKG NVPGAKKSLITVKSAVKS KIRIRLKAYDHRILDQSAEKIV
Mtu GMTQVFDES QELTA EIFADGSYVDVTGTSKGKGFAGTMKRHGFRGQ NGVLLIKG AVPGRTGGLVMVRSAIKR KIRIRLKAYDHEAIDASARKIV
Eco GMTRIFTED QSISV ELFADVKKVDVTGTSKGKGFAGTVKRWNFRTQ RNLLLVKG AVPGATGSDLIVKPAVKA RIRIRLKAFDHRLIDQATAEIV
Syn GMTQIFDQE DAVTA DIFQAGDLVDVAGQSMGRGFAGYQKRHNFRRG RNLLIIKG ALPGKPGTLLNITPAKTV KIRIRLKAFDRRLLDTSCDKIV
Hma GMTHVVLVN LDIVE DIFRAGEYADVAGVTKGKGTQGPVKRWGVQKR GPYTLVKG SVPGPDKRLVPFRPAVRP QARVRLAGTSPEDLDDICADVR
Mja GMSHAFIKE QLNIT DVFQEGELVDTIGVTKGKGFQGQVKRWGVKIQ NNYVVLKG SVQGPAKRLIVLRRAIRP RARIKLSSTDHKVLDEICRQIK

                                   II(63-76)      III(86-96)
Apy ETVKRTGGVVRGPIPLPTRKRKWCVLRSPH EHFEIRAFSRIIDI ALMEINLPAGV
Tma EVAKSTNSKVSGPIPLPTERTLYCVLRSPM EHFEKRVHKRLIDI ALMRINLPAGV
Bsu ETAKRSGASVSGPIPLPTEKSVYTILLAVH EQFEMRTHKRLIDI ALMRLDLPSGV
Mtu ETVVRTGASVVGPVPLPTEKNVYCVIRSPH REHFEMRTHKRLID LMRIDLPASVD
Eco ETAKRTGAQVRGPIPLPTRKERFTVLISPH DQYEIRTHHRLVDI ALMRLDLAAGV
Syn DTANRTNAAAVGPIPLPTKRKIYCVLRSPH EHFETRTHRRIIDI ALMKLDLPAGV
Hma EIANKTGVELSGPVPLPTKTLEVPSRKSPD EHWEMRVHKRLIDI QLMRIQVPNDV
Mja EIAEKTGVDISGPIPLPTKVLRVVTRKSPD DRWTMKIHKRLIDI HIMKIRIPDNV
```



```
EF-G(2)
    7                            35 39    45 46                        70 78
Apy EKLRNIGIVAHIDAGKTTTTERIP-TTGKD GEVTEGA ATMDWMPQEKERGITITAATTACYW QINIIDTPGHVDFSVEVVRSMKVLDGIVFIFSAVEGVQPQSEAN
Tma DKLRNIGIMAHIDAGKTTTTERILYYTGRK GDVDEGN TTTDWMPQEKERGITIQSAATTCFW RINIIDTPGHVDFTAEVERALRVLDGAIRVFDATAGVEPQSETV
Eco ARYRNIGISAHIDAGKTTTTERILFYTGVN GEVHDGA ATMDWMEQEQERGITITSAATTAFW RINIIDTPGHVDFTIEVERSMRVLDGAVMVYCAVGGVQPQSETV
Mva DQIRNMGICAHIAHGKTTLSDNLLAGAGMI SKDLAGD LALDFDEEEAARGITIYAANVSMVH LINLIDTPGHVDFGGDVTRAMRAIDGAVVVCCAVEGVMPQTETV
                   140 259   267 271        284 310          323 329             348 350 354 374                396
Apy WRWADRFKVPRIAFINKMD VLCGSAFKN QPLLDAVIVTYPLP PFCAYAFKVMADPY TYIRVFSGTLKAGSYVYNAT DEKQR AGEICAVVGL-DAATGDTLCDE
Tma WRQADKYNVPRIAFMNKMD VLCGAAKAN QPLLDAVIDYLPSP PFTALVFKVQVDPY VYFRVYSGRLEKGSYVYNST GQRER PGDIAAGVGLKVSQTGDTLWHE
Eco WRQANKYKVPRIAFVNKMD VTCGSAFKN QAMLDAVIDYLPSP PFSALAFKIATDPF TFFRVYSGVVNSGDTVLNSV AARER AGDIAAAIGLKDVTTGDTLCDP
Mva LRQALKEKVKPVLFINKVD VAFGSAYNN EVILDMAIKHLPNP PLAGVVTKIIVDKH SACRLFSGRIKQGDELYLVG KQKAR AGNICALTGLREATAGETVCSP
      411                                                                            495 497 528
Apy K ISMAIEPKTKKDQEKLSQVLNLSSKEDPTFRATTDPETGQILIHGMGELHLEIMVDRMRREYGIEVNVGKPQVAYKETIRKKAI EGK FIDDIHGGVIPKEFIPSV
Tma K ISLAVEPVTKADEEKLVKALLALSEEDPTLQVRVDKETGETIISGMGELHLEIVVDRLKREFGVNVRVGQPQVAYRETIKKSAE EGK F-----------EFMPAI
Eco D ISIAVEPKTKADQEKMGLALGRLAKEDPSFRVWTDEESNQTIIAGMGELHLDIIVDRMKREFNVEANVGKPQVAYRETIRQKVT EGK FINDIKGGVIPGEYIPAV
Mva S ITVAIEAKNTKDLPKLIEILRQIGREDNTVRIEINEETGEHLISGMGELHIEVTDTKIGRDGGIEVDVGEPIIVYRETITGTSP EGK IVNMTKGIVQLDEARELI
                            577  602                        638 649                           687
Apy EKGVKEAMQNGILAGYPVVDVRVRLFDGSYHE DPVLLEPIMEVEVETPEDYVGDVIGDLNSRRGTIMGM AHVPLAEMFGYATTLRSLTQGRGTFIMRFSHYDEVPQH
Tma EAGIKEAMMAGPLAGYPVVRVRAIVLDGSYHE QPVLLEPIMKLEITTPEEYMGNIISDLNSRRAKVESL AKVPLSETFGYATVLRSLSQGRASYIMQFSHYQEVPEK
Eco DKGIQEQLKAGPLAGYPVVDMGVRLHFGSYHD KPVLLEPIMKVEVETPEENTGDVIGDLSRRRGMLKGQ AEVPLSEMFGYATQLRSLTKGRASYTMEFLKYDEAPSN
Mva IEGFKEGVKGGPLASERAQGVKIKLIDATFHE KPILLEPMQKIYINTPQDYMGDAIREINNRRGQIVDM GSVPVAEMFGFAGAIRGATQGRCLWSVEFSGFERVPNE
  EF-Tu(1α)
    10                         39 54                                                      114 116
Apy I KEHVNVGTIGHVDHGKSTLTSAITCVLAAG IDKAPEEKERGITINITHVEYETAKRHYAHVDCPGHADYIKNMITGAAQMDGAILVVSAAD PMPQTREHVLLARQ
Tma I KPHVNVGTIGHIDHGKSTLTAAITKYLSLK IDKAPEEKARGITINITHVEYETEKRHYAHIDCPGHADYIKNMITGAAQMDGAILVVAATD PMPQTREHVLLARQ
Eco V KPHVNVGTIGHVDHGKTTLTAAITTVLAKT IDNAPEEKARGITINTSHVEYDTPTRHYAHVDCPGHADYVKNMITGAAQMDGAILVVAATD PMPQTREHILLGRQ
Mva I KPILNVAFIGHVDAGKSTTVGRLLLDGGAI MDGLKEERERGVTIDVAHKKFPTAKYEVTIVDCPGHRDFIKNMITGASQADAAVLVVNVDD IQPQTREHVFLIRT
              143 150              167 172        184 202      209                                 252 262
Apy VNVPYIVVFMNKCD LLELVELEVRALLSKYEY VPVIRGSALGALQ LNAMDE IPTPEREVDKPFLMPIEDVFSISGRGTVVTGRVERGVLRPGDEV LKTVAT I
Tma VEVPYMIVFINKTD LIDLVEMEVRDLLSQYGY VPVIRGSALKAVE LDAMDN IPDPQRDVDKPFLMPIEDVFSITGRGTVVTGRIERGRIRPGDEV KKTVVT V
Eco VGVPYIIVFLNKCD LLELVEMEVRELLSQYDF TPIVRGSALKALE AGFLDS IPEPERAIDKPFLLPIEDVFSISGRGTVVTGRVERGIIKVGEEV QKSTCT V
Mva LGVRQLAVAVNKMD YNELKKMIGDQLLKMIGF INFVPVASLHGDN AEVIDG FQPPEKPTNLPLRLPIQDVYTITGVGTVPVGRVETGIIKPGDKV IGEKIT V
    269                                     315 336          349 363                  384 385                405
Apy EMFRKVLDEALPGDNIGVLLRGVGKDDVERGQVLAQPGSVKAHRKF NYRPQFYFRTADVT MPGDNVELEVELIAPVALEEGL RFAIREGGRTVGAGVVTKILD
Tma EMFRKELDEGIAGDNVGCLLRGIDKDEVERGQVLAAPGSIKPHKRF GYKPQFYIRTADVT MPGDHVEMEIELIYPVAIEKGQ RFAVREGGRTVGAGVVTEVIE
Eco EMFRKLLDEGRAGENVGVLLRGIKREEIERGQVLAKPGTIKPHTKF GYRPQFYFRTTDVT MPGDNIKMVVTLIHPIAMDDGL RFAIREGGRTVGAGVVAKVLS
Mva EMHHEQLPSAEPGDNIGFNVRGVGKKDIKRGDVLGHTTNPTVATDF GYTPVFHTHTAQIA KAGDAAIVKLIPTKPMVIESVK RFAIRDMGMTVAAGMAIQVTA
```

Fig. 2. Abridged multiple alignment of EF-G(2) and EF-Tu(1α) sequences. *Numbers above blocks* indicate amino acid positions (*A. pyrophilus* numbering). The structure-based parts of the alignment (Ævarsson 1995) spanned *A. pyrophilus* EF-G and EF-Tu residues 1–400 and 1–384, respectively. Positions that are C terminal to the *vertical arrows* (*A. pyrophilus* residues 401–700 and 321–450 for EF-G and EF-Tu, respectively) were selected by BLAST-guided scrutiny of the alignment schemes inferred by CLUSTAL and MULTALIN (Cammarano et al. 1999) and by visually matching conserved motifs constraining the alignment topology. *Highlighting* delimits positions occupied by identical or similar amino acids in no fewer than 19 of 21 sequences.

1996) and BLAST-guided exclusion of spuriously matched positions (Fig. 2 legend). Deselecting regions comprising the artifactual matches left 423 positions for EF-G(2) and 310 positions for EF-Tu(1α). The taxonomic sampling of the joint data set was limited to 21 taxa by the availability of the EF-G(2) sequences. In addition to the groupings present in the *r*-protein data set, the EF-Tu(1α) + EF-G(2) alignment (shown in abridged form in Fig. 2) included *Thermus thermophilus* (Tt; Deinococci) and Proteobacteria of the α subdivision (Pα).

A data set of RNAP β+β′-type sequences (1133 positions) was assembled from global alignments of the individual subunits after BLAST-guided exclusion of spuriously matched segments. This left 613 and 520 positions for the β- and β′-type subunits, respectively. The taxonomic spectrum of the β+β′ data set (22 taxa) was constrained by the number of available β′ sequences. In addition to Archaea, *Aquificales* (*A. aeolicus*), and *T. maritima,* the final alignment (not shown) included low- and high-G+C gram-positives, Cyanobacteria, Proteobacteria (α, β, and γ divisions), and Chlamydiae–Spirochaetes (S).

The “S10 operon,” EF, and RNAP-subunit data sets were estimated to contain 77, 123, and 172 constant positions, which included 53, 115 and 155 invariant sites, respectively.

The three data sets and the global alignments of the individual proteins are obtainable upon request to P. Cammarano.

Phylogenetic Results

Outgroup-rooted bacterial phylogenies were reconstructed using ML and MP methods. In the ML analyses different models for the amino acid substitution process were used to evaluate the robustness of the ML tree against the violation of assumed Markov model.

r-Protein-Based Phylogenies

The 16 taxa comprising the *r*-protein data set were organized into seven topological units and the 945 possible topologies were analyzed for the best tree by the exhaustive search method with ProtML. The ML and AIC values for the different models of amino acid substitution are compared in Table 1. The JTT-F model was slightly better than the JTT and Dayhoff models, while the proportional and Poisson models were less fitting, most probably reflecting overly simplistic assumptions concerning the dynamics of the amino acid substitution process. However, all models recovered the tree topol-

Table 1. Exhaustive-search ML analysis of concatenated *r*-protein sequences with different amino acid substitution models

Model	Tree topology	$-\ln L^{a} \pm$ SE	$-\Delta\ln L$	AIC[b]	ΔAIC
JTT	(O,A,(T,((IG+,Pγ),(hG+,C))))[c]	−19,503.0 ± 318	−28.3	39,064.0	+17
JTT-F	(O,A,(T,((IG+,Pγ),(hG+,C))))	−19,475.7 ± 322	0	39,047.2	0
Dayhoff	(O,A,(T,((IG+,Pγ),(hG+,C))))	−19,617.1 ± 320	−141.4	39,292.1	+45
Proportional	(O,A,(T,((IG+,Pγ),(hG+,C))))	−20,527.9 ± 324	−1052.2	41,151.8	+2,104.8
Poisson	(O,A,(T,((IG+,Pγ),(hG+,C))))	−21,169.7 ± 331	−1694.0	42,397.5	+3,350.5

[a] Log-likelihood.

[b] Akaike information criterion.

[c] O, outgroup archaeal sequences; A, *A. pyrophilus;* T, *T. maritima;* IG+, low-G+C gram-positive Bacteria; hG+, high-G+C gram-positive Bacteria; C, Cyanobacteria-chloroplasts; Pγ, Proteobacteria, γ subdivision. The topologies shown are the ML topologies among 945 alternatives generated from a constrained starting tree in which the 16 taxa were organized into seven topological elements: [constrained tree {((Mva, Mja), Hma), (((Mpn,Mge), Mca), (Bst,Bsu)), ((Cpa,Chl), Syn), (Eco,Hin), Mtu, Tma, Apy}]. Species abbreviations are listed under Methods.

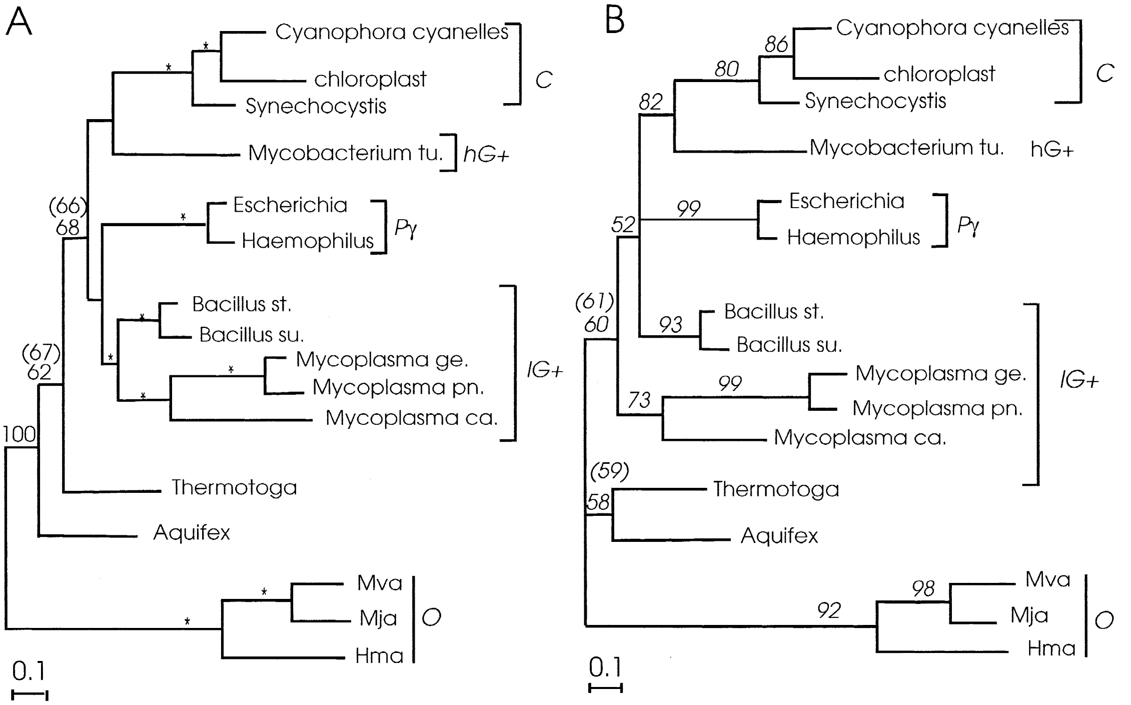


Fig. 3. Maximum-likelihood analysis of the "S10 operon" data set. **A** Tree inferred by the exhaustive-search method (ProtML program) with the JTT-F amino acid substitution model ($\ln L = -19{,}475.7$ in Table 2). *Numbers above internal branches* are BP values calculated as the sum of the BPs of all the topologies showing the node being analyzed in the RELL bootstrap analysis. BP values of trees inferred after exclusion of invariable sites are given in *parentheses.* The 68% BP value attached to the node dividing the hyperthermophiles from the successive lineages is the sum of the BPs of the trees showing the topologies (O,(A,(T,(P,hG+,IG+C))) (15 trees, $\Sigma \mathrm{BP}_i = 0.618$), (O,(T,(A,(P,hG+,IG+,C))) (11 trees, $\Sigma \mathrm{BP}_i = 0.048$), and (O,((A,T),(P,hG+,IG+,C))) (6 trees, $\Sigma \mathrm{BP}_i = 0.0150$). Constrained nodes are indicated by *asterisks.* **B** Tree reconstructed by the QP algorithm with the best (Blosum-F) amino acid substitution model using a mixed rate-heterogeneity model with eight gamma rate categories. *Numbers above nodes* (italics) are QP reliability values. QP reliability values of trees inferred from the data set lacking invariable sites are shown in *parentheses.* ln*L* values of trees inferred with different models of amino acid substitution were −18,971.1 (Blosum-F), −19,053.5 (JTT-F), and −19,062.3 (Dayhoff-F); the corresponding α parameters of the gamma rate distribution were 1.85 ± 0.14, 1.42 ± 0.1, and 1.49 ± 0.1. The ln*L* values of the QP tree inferred assuming uniform amino acid substitution rates were −19,351.5 (Blosum-F), −19,538.9 (JTT-F),and −19,501.8 (Dayhoff-F). *Scale bars* are in units of amino acid substitutions per sequence position.

ogy shown in Fig. 3A having *A. pyrophilus* and *T. maritima* as the deepest and second deepest bacterial offshoot, respectively. In that tree the node dividing the *A. pyrophilus–T. maritima* pair from the successive mesophilic lineages was given at a bootstrap probability (BP) of 68% (see Fig. 3 legend). Examination of the 945 possible trees identified eight topologies (Table 2) that were not significantly worse than the ML tree by the criterion of 1 SE of log-likelihood difference. Of these trees, two (trees 4 and 5 in Table 2) showed the mesophiles as being the deepest offshoots, between 0.6 and 0.8 SE of log-likelihood difference.

The basal placement of *A. pyrophilus* and *T. maritima* was also moderately supported by the QP method (60%

Table 2. Analysis of concatenated *r*-protein sequences: Phylogenetic placement of hyperthermophilic Bacteria inferred by ML with the best (JTT-F) model of amino acid substitution[a]

Tree topology	ΔL_i[b]	(ΔL/SE)	BP$_i$[c]
1. (O,A,(T,((IG+,Pγ),(hG+,C))))[d]	(−19475.7)	Best tree	0.226
2. (O,A,(T,(IG+,(Pγ,(hG+,C)))))	−3.8 ± 9.6	(0.40)	0.111
3. (O,A,(T,(Pγ,(IG+,(hG+,C)))))	−4.8 ± 9.9	(0.49)	0.083
4. (O,Pγ,(IG+,((A,T),(hG+,C))))	−9.1 ± 15.2	(0.60)	0.101
5. (O,Pγ,((A,T),(IG+,(hG+,C))))	−13.1 ± 17.0	(0.77)	0.025
6. (O,A,(T,(hG+,(C,(IG+,Pγ)))))	−9.3 ± 10.4	(0.89)	0.064
7. (O,A,(T,(IG+,(C,(hG+,Pγ)))))	−13.1 ± 14.4	(0.91)	0.026
8. (O,A,(T,(IG+,(hG+,(Pγ,C)))))	−14.0 ± 14.4	(0.97)	0.050

[a] Only 8 of the possible 945 topologies are shown. These could not be significantly discriminated from the ML tree by the criterion of 1 SE of the log-likelihood difference.

[b] ΔL_i is the difference of the log-likelihood of tree *i* from that of the ML tree (tree 1) and ± is 1 SE of the log-likelihood difference.

[c] Relative bootstrap probability for tree *i* being the ML tree among the 945 alternatives during bootstrap resampling, estimated by RELL.

[d] See Table 1, footnote c.

QP reliability), which accounts for site-by-site variation of evolutionary rates and does not require a prior clustering of the taxa (Fig. 3B). Unlike ProtML, however, *A. pyrophilus* and *T. maritima* formed a weakly supported monophyletic grouping (58% QP reliability) near the base of the bacterial clade.

In both the ProtML and the QP analyses exclusion of the invariant sites did not affect the tree topology and the statistical support for the tree branches (see BP and QP reliability values given in parentheses in Figs. 3A and B).

A MP analysis of the *r*-protein data set recovered a single most parsimonious tree (4529 steps) whose topology—(O,(A,(T,(IG+,(Pγ,(hG+,C))))))—roughly mirrored that of the ML tree. However, the nodes linking the six bacterial groupings were given at bootstrap confidence levels between 25 and 50%.

EF-Based Phylogenies

The 21 taxa comprising the joint EF data set were organized into nine topological units and the possible 135,135 topologies were examined for the ML tree by the exhaustive search method. The JTT amino acid substitution model was the best among the alternatives (Table 3). However, all five models gave *A. pyrophilus* and *T. maritima* as the deepest bacterial branches—deeper than *T. thermophilus* (Deinococcus lineage). The ML tree under the best model (JTT) is shown in Fig. 4A along with abridged versions of the ML trees inferred from the separate EF-Tu(1α) and EF-G(2) sequence data sets. Importantly, the divide between the hyperthermophiles and the successive lineages was robustly supported by a BP of 82% (see Fig. 4 legend), in contrast to trees inferred from the separate EF data sets (BP = 55 and 58%), i.e., the statistical confidence for the *Aquifex–Thermotoga* pair being basal to the other bacterial lineages was significantly enhanced when the length of the data set was increased by combining the EF-G(2) and EF-Tu(1α) sequences. Examination of the 1000 trees selected by the approximate likelihood method identified 22 topologies (trees 1–22 in Table 4) that were not significantly worse than the ML tree by the criterion of 1 SE of log-likelihood difference. Of these trees, three (trees 15, 17, and 21) showed the mesophiles as the deepest offshoots, between 0.87 and 0.98 SE of log-likelihood difference.

The branching pattern of the bacterial phyla repeated itself in the QP analysis shown in Fig. 4B. Once again, the composite EF data set outperformed the separate EF-G(2) and EF-Tu(1α) data sets regarding both resolution and statistical robustness of the inferred phylogenies. In the separate EF trees the relative branching order of the three thermophilic taxa (A, T, Tt) was indeterminate, and the divide between thermophilic and mesophilic Bacteria was modestly supported.

The topology and statistical robustness of both the ProtML and the QP trees were basically unaffected by the exclusion of inferred invariant sites as indicated by bootstrap and QP reliability values given in parentheses in Figs. 4A and B.

Except for an inverted branching order, the placement of the *A. pyrophilus–T. maritima* pair at the base of the bacterial clade was recovered, at 85% bootstrap confirmation, in the MP analysis (Fig. 5). Also, similarly to ML, the basal branching of the two hyperthermophiles was weakly supported (<60% bootstrap confirmation) by trees inferred from the separate EF-Tu(1α) and EF-G(2) data sets.

In order to assess whether the MP tree was statistically distinguishable from competing alternatives, the tree in Fig. 5 was challenged with the spectrum of topologies that fell within 1 SE of the log-likelihood difference in the ML analysis (trees 1–22 in Table 4). Any alternative was declared significantly worse than the MP tree when the mean of Δ_{sbst} (the difference in number of inferred substitutions) was $\geq 1.96 \times$ SD. As Table 4 (Section II) shows, the MP tree (tree 23) could be discriminated, at the 0.05 level, from all the alternatives having the mesophilic lineages, or *T. thermophilus*, as the deepest radiations (trees 15, 17, 21), and also from alternatives showing a monophyletic *Aquifex–Thermotoga* grouping as the deepest bacterial offshoot (tree 10). However, the analysis did not discriminate whether *A. pyrophilus* or *T. maritima* was the deepest branch.

RNAP-Based Phylogeny

A preliminary analysis performed with PUZZLE demonstrated that (i) the amino acid compositions of the *Mycoplasma genitalium* and *Mycoplasma pneumoniae* β+β′ sequences (M) are highly and significantly different from the frequency distributions assumed in the ML models ($p < 0.01$; 5% χ^2 test) and (ii) the two sequences show faster-than-average evolutionary rates (see below). As both biased composition (Lake and Rivera 1994; Lento et al. 1995; Foster and Hickey 1999) and acceler-

Table 3. Exhaustive-search ML analysis of the EFG(2)–EF-Tu(1α) data set with different amino acid substitution models

Model	Tree topology	$-\ln L^{a}$ ± SE	$-\Delta\ln L$	AIC[b]	ΔAIC
JTT	(O,A,(T,(Tt,(IG+,(C,(Pα,(Pγ,hG+)))))))[c]	−16,804.8 ± 396	0	33,687.6	0
JTT-F	(O,A,(T,(Tt,(IG+,(C,(Pα,(Pγ,hG+)))))))	−16,827.7 ± 395	−22.9	33,771.5	+83.9
Dayhoff	(O,A(T,(Tt,(C,(IG+,(Pα,(Pγ,hG+)))))))	−16,914.3 ± 389	−109.5	33,906.7	+219.1
Proportional	(O,A,(T,(Tt,(C,(IG+,(hG+,(Pα,Pγ)))))))	−17,998.9 ± 409	−1,194.1	36,113.8	+2462.2
Poisson	(O,A,(T,(Tt,(IG+,(C,(hG+,(Pα,Pγ)))))))	−18,492.8 ± 418	−1,688.0	37,063.7	+3376.1

[a] Log-likelihood.

[b] Akaike information criterion.

[c] Tt, *T. thermophilus;* Pα, Proteobacteria, α subdivision; all other abbreviations are as listed in Table 1, footnote c. The topologies shown are the ML topologies among 135,135 alternatives generated from a constrained starting tree in which the 21 taxa were organized into nine topological elements [constrained tree {(((((Mja,Mva), Pwo), Tac), Hha), Sso), Apy, Tma, Tth, (Atu,Rpr), ((Mpn,Mge), Bsu), ((Ani,Spl), Syn), (Mle,Mlu), (Eco,Hin)}]; species abbreviations are as listed under Methods.

ated evolution (Felsenstein 1978; Budin and Philippe 1998) might distort the tree topology, the β+β′-type subunit data set was analyzed with and without the two mycoplasma sequences.

ML trees inferred by the exhaustive search method from the two data subsets are listed in Table 5. The ML trees reconstructed under the best (JTT-F) model are compared in Figs. 6 A and A′. Overall, the results show that the positioning of the two hyperthermophiles is critically affected by inclusion of the *M. genitalium* and *M. pneumoniae* sequences, there being two conflicting positions for *A. aeolicus.* Notably, in the absence of mycoplasmas, *T. maritima* and *A. aeolicus* fell basal to all other lineages, while inclusion of mycoplasmas led to (i) the repositioning of *A. aeolicus* between Chamydiae–Spirochaetes (S) and Proteobacteria and (ii) the rooting of the bacterial tree at the mycoplasma level in all amino acid substitution models to the exception of the best-fit model (JTT-F), which had mycoplasmas affiliated, albeit weakly, with low-G+C gram-positives, similarly to trees of 16S rRNA (Woese 1987) and EF sequences (Figs. 4B and 5). The displacement of *Aquifex* among the mesophiles (Figs. 6 A and A′) was accounted for by idiosyncrasies of the mycoplasma sequences (rather than by the addition of new taxa), since it persisted when the inclusion of *M. genitalium* and *M. pneumoniae* was counterbalanced by deselecting the low-G+C gram-positives *Bacillus subtilis* and *Staphylococcus aureus* (results not shown).

However, with both subsets of RNAP-subunit sequences the branching order of the bacterial phyla was not significantly supported in the exhaustive search analysis (Figs. 6A and A′) and was unresolved in the QP ML analysis (Fig. 6B), which uses multifurcations for ambiguous groupings. Furthermore, a glance at the branch lengths in the QP tree renders it visually obvious that the branches representing the compositionally biased *M. genitalium* and *M. pneumoniae* sequences are up to three times longer than those of the other free-living Bacteria, suggesting accelerated evolutionary rates.

The MP analysis of the RNAP data set lacking mycoplasmas recovered the basal placement of the two hyperthermophiles [topology, (O,T,(A,(P,(S,((C,(IG+, hG+))))))))]in less than 40% bootstrap samples. On the other hand, a parallel analysis of the all-inclusive data set recovered the rooting of the bacterial tree at the mycoplasma level seen in all of the ML trees reconstructed with suboptimal amino acid substitution models [topology, (O,M,(T,(((A,S),P),(C,(IG+,hG+))))))]. In this case, however, the division of *Mycoplasmatales* from the successive lineages was robustly inferred (96% bootstrap confirmation). The latter result is not surprising if the early branching of mycoplasmas reflects a long branch attraction artifact, to which MP is known to be particularly sensitive (Felsenstein 1978).

Discussion

Congruence between different markers is the most reliable criterion to assess evolutionary history. Hence, we have reconstructed bacterial phylogenies inclusive of the *Aquifex* and *Thermotoga* lineages using the concatenated *r*-proteins encoded in the S10 operon, the combined sequences of translational elongation factors Tu(1α) and G(2) encoded in the *str* operon, and the joint RNAP β- and β′-type subunit sequences encoded in the *rpoBC* gene cluster.

The genes used here do not appear to be implicated in lateral gene transfers involving hyperthermophilic Bacteria and Archaea (Tiboni et al. 1993; Brown et al. 1994; Gribaldo et al. 1999). This possibility was in fact excluded by (i) the lack, in the bacterial *r*-proteins, of discrete insertions unique to the (generally longer) archaeal–eukaryal *r*-proteins; (ii) the presence, in EF-G, of a unique 100- to 120-aa insertion (the G′ subdomain) having no counterpart in the archaeal–eukaryal homologues (EF-2) (Ævarsson, 1995); (iii) the lack, in EF-Tu, of five conserved insertions that are unique to the archaeal–eukaryal EF-1α sequences; and (vi) several signature motifs distinguishing bacterial and archaeal β- and β′-type RNAP subunits. Also, because *r*-proteins and EFs are highly integrated with the cognate ribosomal

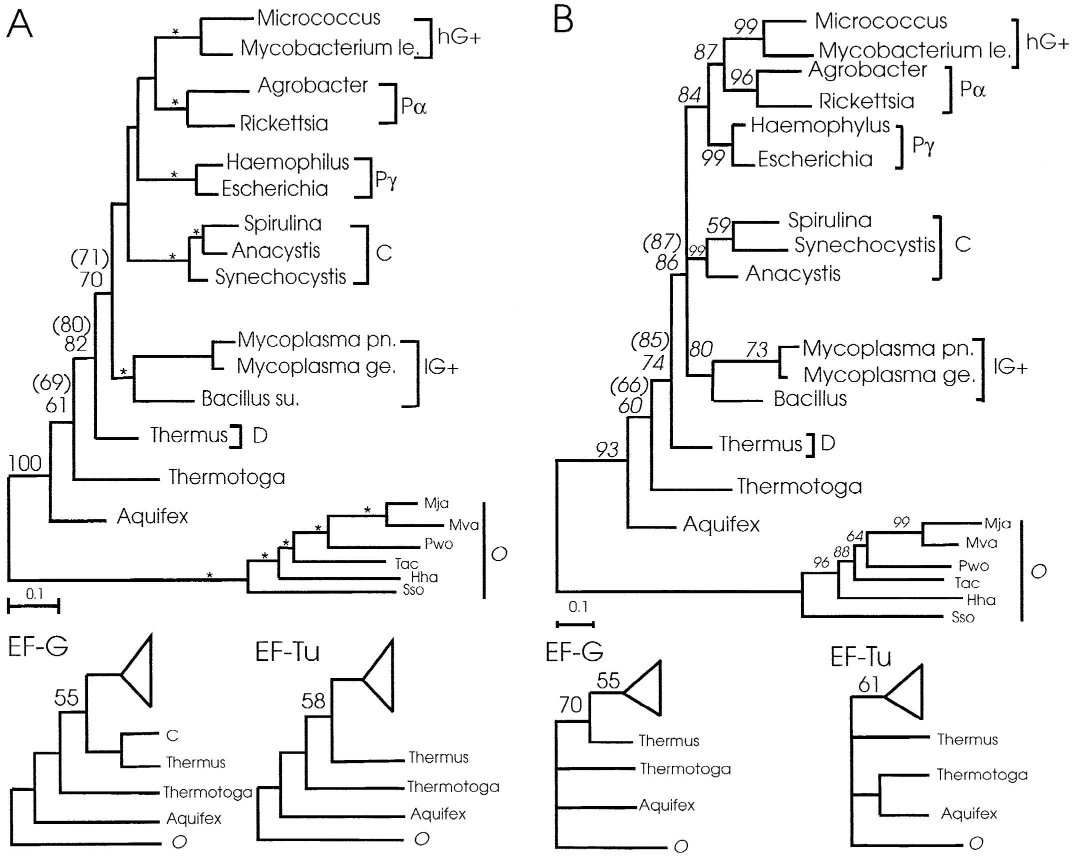


Fig. 4. Maximum-likelihood analysis of the EF data set. **A** Tree inferred from the composite data set by the exhaustive-search method (ProtML program) with the JTT amino acid substitution model (ln*L* = −16,804.8 in Table 4). Constrained nodes are indicated by *asterisks. Numbers* are the BPs of internal branches estimated as the sum of the BPs of all the trees showing the node in question in the RELL analysis. BPs for a data set lacking invariant sites are given in *parentheses.* The BP = 82% assigned to the node dividing the two hyperthermophiles from the successive lineages is the sum of the BPs of trees having the topologies (O,(A,(T,(P,Tt,hG+,lG+,C))) (37 trees, $\Sigma BP_i = 0.611$), (O,(T,(A,(Tt,P,hG+,lG+,C))) (27 trees, $\Sigma BP_i = 0.198$), and (O,((A,T),(Tt,P,hG+,lG+,C))) (9 trees, $\Sigma BP_i = 0.016$). The *bottom two trees* are abridged versions of the ML trees inferred from the separate EF-G(2) and EF-Tu(1α) data sets. **B** Tree reconstructed by the QP algorithm with the best (Blosum-F) model of amino acid substitution using a mixed rate-heterogeneity model with eight gamma rate categories. *Numbers* (italics) are QP reliabilities of internal branches. QP reliability values for trees inferred from a data set lacking inferred invariable sites are shown in *parentheses.* The ln*L* values were −16,215.5 (Blosum-F model) and −16,228.0 (JTT-F model); the corresponding α parameters of the gamma rate distribution were 0.96 ± 0.05 (Blosum-F) and 0.83 ± 0.04 (JTT-F). The ln*L* values of QP trees reconstructed assuming uniform amino acid substitution rates were −16,750.0 (Blosum-F) and −16,839.5 (JTT-F). The *bottom two trees* are abridged versions of the QP trees inferred with the Blosum-F model from the separate EF-G(2) and EF-Tu(1α) data sets with eight categories of site-by-site rate variation (ln*L* values were −9720.4 and −6465.2, respectively). D, Deinococci. *Scale bars* are in units of amino acid substitutions per sequence position.

components, successful transfer of the corresponding genes across the bacterial phyla appears unlikely.

r-*Protein- and EF-Based Phylogenies*

The “S10 operon” and EF data sets do not enable a (statistically significant) resolution of the order of emergence of the bacterial groupings above the *Aquifex* and *Thermotoga* level, possibly reflecting a massive evolutionary radiation at the base of the mesophilic phyla. However, both sets of sequences confirm with moderate (*r*-proteins) to high (EFs) statistical support the early emergence of the two hyperthermophilic lineages inferred from analysis of small-subunit rRNA sequences (Burggraf et al. 1992; Barns et al. 1996). The most compelling evidence for *Aquifex* and *Thermotoga* being basal to the other bacterial lineages is provided by the MP analysis of the concatenated EF-Tu(1α)–EF-G(2) sequences. In that analysis, trees showing mesophiles, or *T. thermophilus,* as the deepest bacterial branches, and trees showing *A. pyrophilus* and *T. maritima* as a deep monophyletic grouping, can be confidently rejected. The analysis, however, does not allow us to assess whether *Aquifex* or *Thermotoga* is the most basal lineage.

The possibility that the early emergence of *A. pyrophilus* and *T. maritima* is a reconstruction artifact resulting from the attraction between the two “hyperthermophilic” branches and the long branch of the outgroup (Felsenstein 1978; Budin and Philippe 1998) can be reasonably excluded, as in both the *r*-protein- and the EF-

Table 4. Phylogenetic placement of hyperthermophilic Bacteria by ML and MP analyses of the EF-G(2) + EF-Tu(1α) data set[a]

	I. ML			II. MP		
Tree topology	ΔL_i[b] ± SE	(ΔL/SE)	BP$_i$[c]	sbst	Δ_{sbst}[d] ± SD	N/W
1. (O,A,(T,(Tt,(IG+,(C,(Pγ,(Pα,hG+)))))))[e]	(−16,804.8) Best tree		0.1480	3,847	20 ± 13.96	N
2. (O,A,(T,(Tt,(C,(IG+,(Pγ,(Pα,hG+)))))))	−1.1 ± 9.6	(0.12)	0.089	3,844	17 ± 14.19	N
3. (O,A,(T,(Tt,((C,IG+),(Pγ,(Pα,hG+))))))	−4.1 ± 9.0	(0.46)	0.057	3,846	19 ± 13.90	N
4. (O,T,(A,(Tt,(C,(IG+,(Pγ,(Pα,hG+)))))))	−5.9 ± 12.5	(0.47)	0.028	3,832	5 ± 11.28	N
5. (O,A,(T,(Tt,C),(IG+,(Pγ,(Pα,hG+))))))	−7.8 ± 14.9	(0.52)	0.036	3,850	23 ± 15.21	N
6. (O,A,(T,(Tt,(C,(IG+,(Pα,(Pγ,hG+)))))))	−8.6 ± 16.1	(0.53)	0.052	3,840	13 ± 12.05	N
7. (O,A,(T,(Tt,(C,(IG+,(hG+,(Pγ,Pα)))))))	−10.5 ± 16.7	(0.63)	0.030	3,846	19 ± 13.90	N
8. (O,A,(T,(Tt,(IG+,(C,(Pγ,(Pα,hG+)))))))	−4.8 ± 7.5	(0.64)	0.033	3,832	5 ± 10.63	N
9. (O,T,(A,(Tt,((C,lG+),(Pγ,(Pα,hG+))))))	−8.3 ± 11.9	(0.70)	0.019	3,832	5 ± 11.10	N
10. (O,(A,T),(Tt,(C,(IG+,(Pγ,(Pα,hG+))))))	−9.9 ± 11.8	(0.84)	0.002	3,855	28 ± 14.00	W
11. (O,T,(A,(Tt,(C,(IG+,(Pα,(Pγ,hG+)))))))	−13.1 ± 18.1	(0.72)	0.012	3,828	1 ± 9.23	N
12. (O,T,(A,((Tt,C),(IG+,(Pγ,(Pα,hG+))))))	−13.3 ± 17.0	(0.78)	0.008	3,839	12 ± 13.04	N
13. (O,T,(A,(Tt,(C,(IG+,(hG+,(Pα,Pγ)))))))	−14.9 ± 18.8	(0.79)	0.011	3,835	8 ± 11.05	N
14. (O,A,(T,((Tt,C),(IG+,(hG+,(Pα,Pγ))))))	−16.5 ± 20.2	(0.82)	0.017	3,848	21 ± 14.87	N
15. (O,Tt,((A,T),(C,(IG+,(Pγ,(Pα,hG+))))))	−12.9 ± 14.8	(0.87)	0.016	3,863	36 ± 15.04	W
16. (O,A,(T,(Tt,((C,IG+),(Pγ,(Pα,hG+))))))	−14.7 ± 16.5	(0.89)	0.011	3,851	24 ± 13.57	N
17. (O,C,(Tt,((A,T),(IG+,(Pγ,(Pα,hG+))))))	−17.7 ± 18.8	(0.94)	0.014	3,863	36 ± 16.32	W
18. (O,T,(A,((Tt,C),(IG+,(hG+,(Pα,Pγ))))))	−21.3 ± 22.0	(0.97)	0.005	3,837	10 ± 12.82	N
19. (O,A,(T,((Tt,C),(IG+,(Pα,(Pγ,hG+))))))	−18.7 ± 19.3	(0.97)	0.010	3,849	22 ± 13.50	N
20. (O,A,(T,(Tt,(IG+,(C,(Pα,(Pγ,hG+)))))))	−13.7 ± 14.1	(0.97)	0.020	3,842	15 ± 11.28	N
21. (O,Tt,(C,((A,T),(IG+,(Pγ,(Pα,hG+))))))	−18.7 ± 19.0	(0.98)	0.010	3,874	47 ± 16.17	W
22. (O,A,(T,(Tt,(IG+,(C,(hG+,(Pα,Pγ)))))))	−13.9 ± 14.0	(0.99)	0.013	3,851	24 ± 13.50	N
23. (O,T,(A,(Tt,(IG+,(C,(Pα,(Pγ,hG+)))))))	−18.2 ± 16.0	(1.14)	0.001	(3,827) Best tree		

[a] N and W: topology not significantly worse (N) and significantly worse (W) than the best tree in the MP analysis. User-defined trees (topologies 1–22 in Section I) are declared significantly worse than the best tree when $\Delta_{sbst} \geq 1.96 \times$ SD. Of the 135,135 possible trees, trees 1–22 could not be significantly discriminated from the ML tree by the criterion of 1 SE of the log-likelihood difference.

[b] ΔL_i is the difference of the log-likelihood of tree *i* from that of the ML tree (tree 1), and ± is 1 SE of the log-likelihood difference.

[c] Relative bootstrap probability for tree *i* being the ML tree among the 1000 best trees selected.

[d] Difference in number of inferred substitutions between the MP tree (tree 23) and alternative trees listed in Section I.

[e] See Table 3, footnote c.

based phylogenies the two lineages are represented by relatively short, i.e., slow-evolving branches (see Figs. 3 and 4). Interestingly, a basal radiation of *Aquifex* and *Thermotoga* (with *Aquifex* emerging below *Thermotoga*) is also robustly supported by phylogenies of available FtsZ protein sequences (S. Gribaldo, unpublished results).

A remarkable aspect of the present results is the evidence that the phylogenetic signal delivered by the joint EF data set is stronger than that delivered by the separate EF-Tu(1α) and EF-G(2) components. However, this situation cannot be generalized. In an attempt to improve the robustness of the *r*-protein phylogeny, we developed an *r*-protein data set (516 aa positions; 13 bacterial taxa and 3 Archaea) comprising proteins L14, L24, L5, S14, S8, L5, L6, and L18, encoded in that order in the spectinomycin (*spc*) operon, which is situated immediately downstream from (or fused to) the last gene (*rps*17) of the S10 operon (Sanangelantoni et al. 1994). Unlike the "S10 operon," however, both exhaustive search and quartet puzzling ML analyses have a multifurcating tree with *A. pyrophilus* and *T. maritima* forming a monophyletic cluster [topology, (O, (A, T), Pγ, C, hG+, IG+)]. When added to the S10 operon data set, the uninformative *spc* operon data set did not obscure the phylogenetic signal delivered by the former but affected the support for the *Aquifex–Thermotoga* pair being basal to the mesophilic lineages (45% bootstrap confidence instead of 68% in the exhaustive search ML analysis).

RNP-Based Phylogenies: Are Mycoplasmas Ancestral to Other Bacteria?

Unlike *r*-protein- and EF-based phylogenies, phylogenies of RNAP-subunit sequences are equivocal with respect to the *Aquifex* and *Thermotoga* placements. First, different placements of *Aquifex* are inferred by the exhaustive search method, depending on whether or not *M. genitalium* and *M. pneumoniae* are included in the analysis (Table 5). Second, the phylogenetic content of the data set appears generally unsuited to afford a statistically significant resolution of the deepest bacterial relationships.

As the mycoplasma β and β′ sequences are significantly biased in composition, and exhibit faster-than-average evolutionary rates, we implicate these factors as being responsible for both (i) the displacement of *Aquifex* from its basal position (Fig. 6) and (ii) the rooting of Bacteria at the mycoplasma level seen in all of the

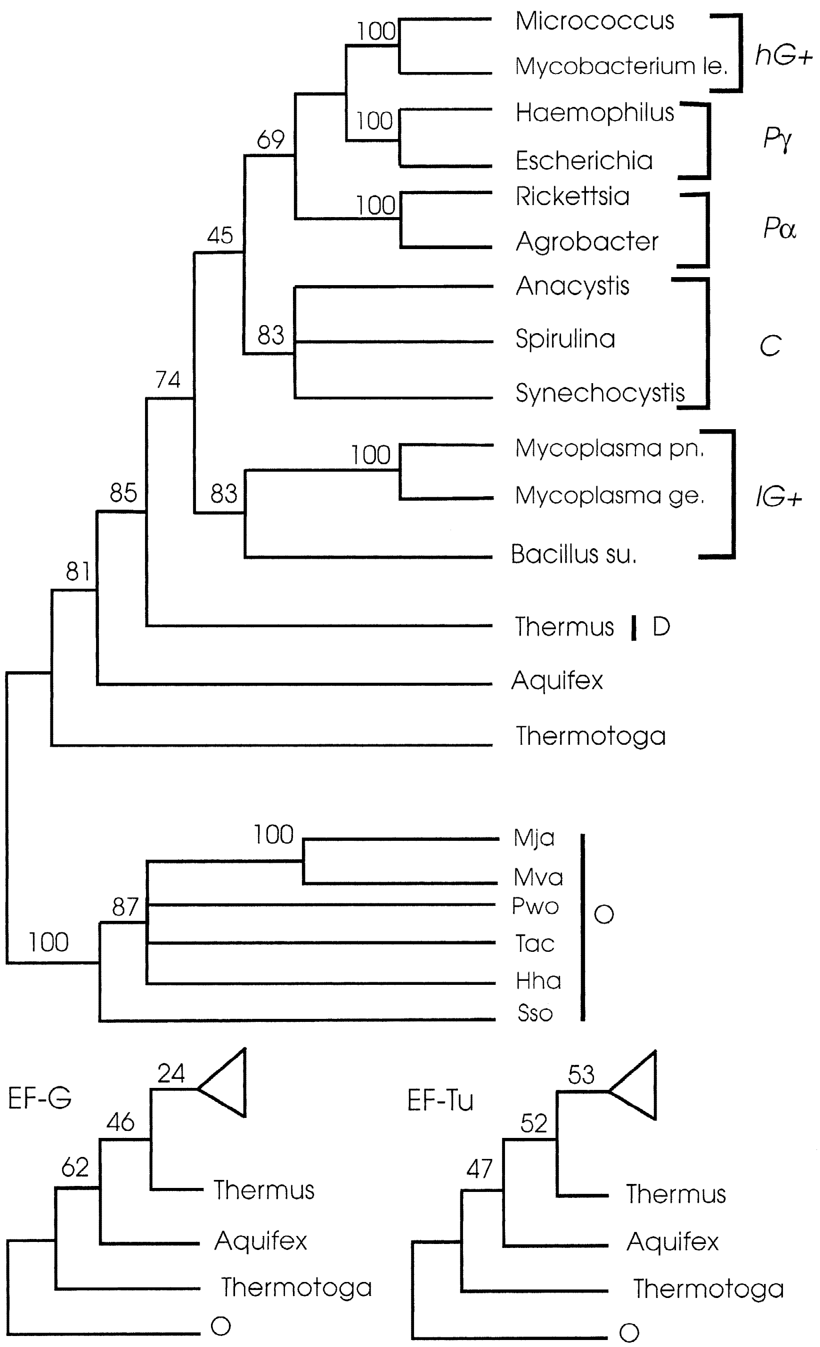


Fig. 5. Maximum-parsimony analysis (PROTPARS program) of the EF data set. *Numbers above nodes* represent the number of times the underlying branch was given in 100 bootstrap replications. The tree shown is topology 23 in Table 4. The *bottom two trees* are MP trees inferred from the separate EF-G(2) and EF-Tu(1α) data sets. Branch lengths do not represent numbers of inferred substitutions.

RNAP-based phylogenies, except the one reconstructed with the best-fit amino acid substitution model (Table 5). Most probably, a long branch attraction artifact is responsible for the positioning of the mycoplasma β+β′ sequences at the base of the bacterial clade, and additional perturbations of the tree topology may arise from the correlated compositional biases (Lake and Rivera 1994; Lockhart et al. 1996; Foster and Hickey 1999). It is of interest that chloroplast sequences, which are on a long branch of the RNAP-based tree but are not compositionally biased ($p > 0.15$, 5% χ^2 test), are firmly grouped with their cyanobacterial kins, away from the root of the bacterial tree (Fig. 6B).

Indeed, if mycoplasmas are excluded, the two hyperthermophiles fall basal to all other bacterial lineages, in agreement with the 16S rRNA-, *r*-protein-, and EF-based phylogenies. Unlike the latter, however, the order of emergence of the bacterial phyla is tenuously supported in the ML trees inferred by the exhaustive search method and is unresolved in trees inferred by the QP algorithm, which gives multifurcations for groupings that are reconstructed only occasionally during multiple puzzling steps (Strimmer and von Haeseler 1996).

All the more important, the rooting of Bacteria at the mycoplasma level conflicts with phylogenies of EF sequences, which robustly affiliate *M. genitalium* and *M.*

Table 5. Exhaustive-search ML analysis of two subsets of RNAP β- and β′-type subunit sequences including or lacking mycoplasma sequences

Model	Tree topology	$-\ln L$[a] ± SE	$-\Delta\ln L$	AIC[b]	ΔAIC
JTT	(O,T,(A,((S,P),(C,(IG+,hG+)))))[c]	−20,905.9 ± 379	−59.7	41,869.8	+81.3
JTT-F	(O,T,(A,(S,(P,(C,(IG+,hG+))))))	−20,846.2 ± 381	0	41,788.5	0
Dayhoff	(O,T,(A,(S,(P,(C,(IG+,hG+))))))	−21,234.6 ± 384	−388.4	42,527.2	+738.7
Proportional	(O,T,(A,((S,P),(C,(IG+,hG+)))))	−22,470.7 ± 397	−1624.5	45,037.4	+3248.9
JTT	(O,M,(T,((S,(A,P)),(C,(IG+,hG+)))))	−22,973.0 ± 421	−93.3	46,011.9	+99.9
JTT-F	(O,T,((S,(A,P)),(C,(hG+,(IG+,M)))))	−22,904.0 ± 414	0	45,912.0	0
Dayhoff	(O,M,(T,((S(A,P)),(C,(IG+,hG+)))))	−23,320.6 ± 421	−317	46,707.1	+795.1
Proportional	(O,M,(T,((S,(A,P)),(C,(IG+,hG+)))))	−24,706.9 ± 436	−1,823	49,557.4	+3,645.4

[a] Log-likelihood.

[b] Akaike information criterion.

[c] S collectively designates Chlamydiae (*Chlamydia trachomatis*) and Spirochaetes (*Borrelia burgdorferi*), which were clustered together in the ProtML analysis; M designates the mycoplasmas (*M. pneumoniae* and *M. genitalium*); P collectively designates the Pγ and Pα subdivision of Proteobacteria. All other abbreviations are as listed in Table 1, footnote c. The topologies shown are the ML topologies among the 10,385 alternatives generated by organizing the 16 taxa into eight topological elements [constrained tree {(Syn,chl), (Bsu,Sau), (Mtu, Mle), (Rpr, (Ppu, (Eco,Hin))), (Ctr,Bbu), (Tce,Sso), Tma, Aae}] and among the 2,027,025 alternatives generated by organizing the 18 taxa into 10 topological units [constrained tree {(Syn,chl), (Bsu,Sau), (Mtu, Mle), (Rpr, (Ppu, (Eco,Hin))), (Ctr,Bbu), (Tce,Sso), Tma, Aae, Mpn, Mge}]; species abbreviations are as listed under Methods.

pneumoniae with low-G+C gram-positive Bacteria, remote from the root of the bacterial tree (see Figs. 4B and 5). As the *M. genitalium* and *M. pneumoniae* EF-Tu and EF-G sequences show unbiased amino acid compositions ($P > 0.85$, χ^2 analysis) and do not display faster-than-average evolutionary rates (compare the ML branch lengths in Fig. 4), we believe that EF-based phylogenies provide a more trustworthy picture of the mycoplasma placement than phylogenies of the RNAP-subunit sequences, which probably suffered from a relaxation of evolutionary constraints in *Mycoplasmatales*. Furthermore, the affiliation of mycoplasmas with the low-G+C gram-positives predicted by 16S rRNA (Woese 1987) and EF-based phylogenies (this paper) is also strongly supported by phylogenetic trees of Hsp70(*dnaK*) sequences (Gribaldo et al. 1999).

The drawbacks implicated by the idiosyncratic *M. genitalium* and *M. pneumoniae* β and β′ sequences have been overlooked in a recent RNAP phylogeny showing the two hyperthermophiles nested among the mesophilic phyla (Klenk et al. 1999) and *Mycoplasmatales* as the deepest bacterial radiation. This picture roughly mirrors that of trees inferred from the full spectrum of β+β′ sequences under suboptimal amino acid substitution models (Table 5). Interestingly, a similar biased amino acid composition has suggested the exclusion of *M. genitalium* and *M. pneumoniae* from phylogenies of valyl-tRNA synthase sequences (Hashimoto et al. 1998).

Overall, different phylogenetic markers either do not deliver a significant phylogenetic signal with respect to the branching order of the bacterial phyla (RNAP subunits, *spc* operon) or, if a signal is delivered (16S rRNA, S10 operon, EFs, FtsZ), this is a coherent one placing the hyperthermophiles at the base of the bacterial clade.

Contradictory Facts

The deep positioning of *Thermotoga* as inferred from 16S rRNA-, *r*-protein-, and EF-based phylogenies conflicts with analyses of Hsp70 and GS1 sequences, which suggest *T. maritima* as being specifically and closely related to the gram-positive Bacteria.

The *T. maritima* Hsp70 sequence lacks a relatively conserved insertion that distinguishes gram-negative Bacteria and Eucarya (all of which possess the Hsp70 insertion) from gram-positive Bacteria and Archaea (which lack the Hsp70 insert). Therefore, by this criterion Thermotogales should be affiliated with the gram-positives (Gupta et al. 1997; Gupta 1998), even though *T. maritima* does not cluster with the gram-positive Bacteria in phylogenetic trees of Hsp70 sequences (Gribaldo et al. 1999). The incongruity between the placement of Thermotoga argued from gene trees and from the presence/absence of the insertion could be rationalized by assuming that (i) the *Hsp70/dnaK* gene of *T. maritima* was recruited from a gram-positive Bacterium via lateral gene transfer but the identity of the source gene has been obscured by subsequent mutations; (ii) the insert is an unstable character that was lost more than once during bacterial evolution; and (iii) different bacterial phyla retained either one or the other of two paralogous versions of the *dnaK* gene (one possessing and the other lacking the insert).

A comparable incongruity is offered by the anomalous clustering of *T. maritima* (T) with Euryarchaeotes (E) and low-G+C gram-positives (IG+) in phylogenetic trees of GS1 sequences (Tiboni et al. 1993). Here, the configuration of the GS1 cluster [topology, (O,(E,((T,hG+))))] could be most parsimoniously explained by the single transfer of a euryarchaeal GS1 gene provided *Thermotogales* were affiliated with (or shared a

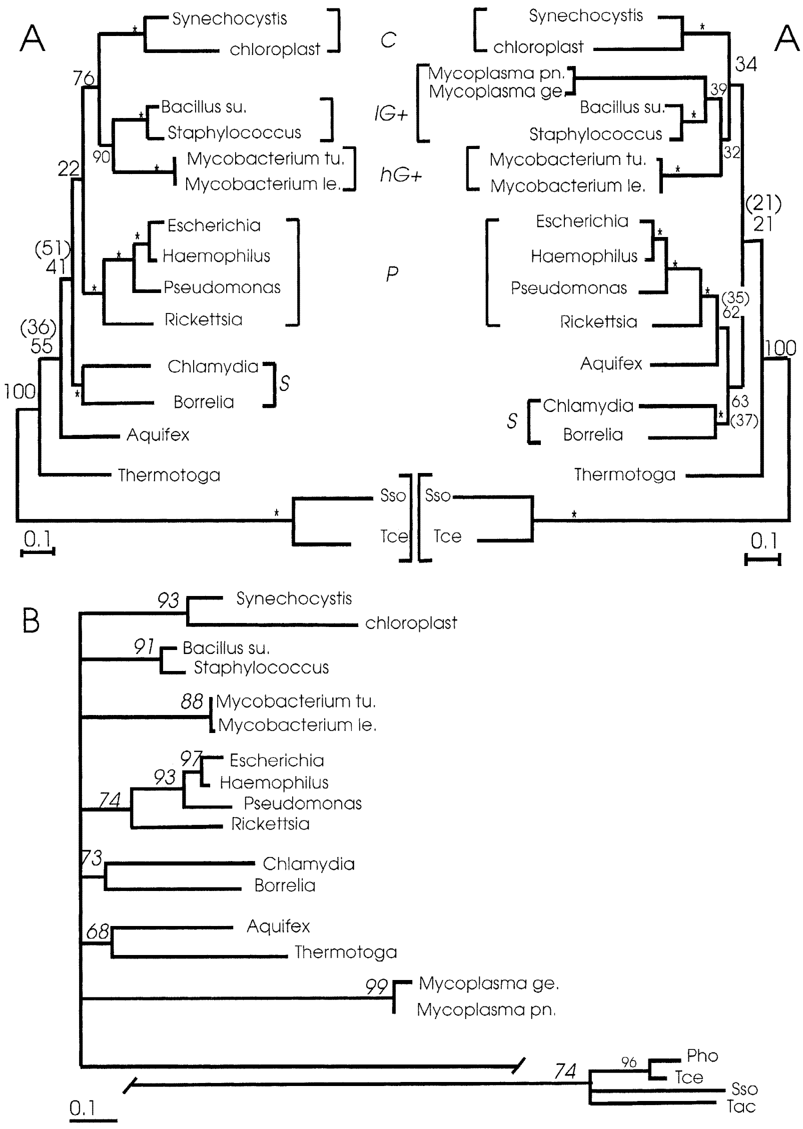


Fig. 6. Maximum-likelihood analysis of the RNAP β+β′-type subunits data sets. **A, A′** Trees inferred by ProtML (JTT-F amino acid substitution model) from data sets comprising (A) and lacking (A′) the *M. genitalium* and *M. pneumoniae* sequences. **B** Tree inferred from the all-inclusive alignment with the QP algorithm under the best (JTT-F) amino acid substitution model, using eight rate categories of site-by-site rate variation and an α parameter of the gamma rate distribution of 0.96 ± 0.03 estimated from the data set. Numbers are bootstrap (A, A′) and QP (B) reliability values. *Numbers in parentheses* represent bootstrap and QP reliability values inferred from data sets lacking invariant sites. *Scale bars* are in units of amino acid substitutions per sequence position.

last common ancestor with) the low-G+C gram-positives (Brown et al. 1994; Pesole et al. 1995).

However, in light of 16S RNA-, EF-, and *r*-protein-based phylogenies, the clustering of *T. maritima* with low-G+C gram-positives in trees of GS1 sequences can be interpreted only by implicating either (a) two independent transfers of the euryarchaeal GS1 gene—one to *T. maritima* and the second to an ancestor of the low-G+C gram-positives; or (b) the sequential propagation of the archaeal GS1 gene from Euryarchaeotes, to *Thermotogales,* to low-G+G gram-positives.

Evolutionary Implications

The basal placement of the *Aquifex* and *Thermotoga* lineages inferred from the 16S rRNA-based phylogenies, together with the fact that hyperthermophiles also occupy the deepest archaeal branches, has been interpreted as indicating that the last common ancestor of extant life on Earth was a hyperthermophile (Burggraf et al. 1992; Barns et al. 1996), possibly a survivor of early cataclismic events that annihilated attendant mesophilic or psycrophilic flora (Lazcano and Miller 1996). The phylogenetic antiquity of the *Aquifex–Thermotoga* pair inferred from *r*-protein and EF sequences in the present report is consistent with, and supports, this interpretation. Nevertheless, calculations of the base composition of primordial ribosomal RNAs indicate a relatively low guanine *plus* cytosine content, which is probably incompatible with a hyperthermophilic lifestyle of the last universal ancestor (Galtier et al. 1999). If this is so, hyperthermophily would be a derived rather than ancestral character, i.e., the descendants of the last universal ancestor, and possibly its predecessors, may have adapted several times to hot terrestrial environments during the history of life on this planet.

Acknowledgments. This work has been supported by a grant from MURST Progetto Nazionale Protein–Nucleic Acids interactions and from the CNR Progetto Finalizzato Biotecnologie. We are grateful to E. Conway de Macario for critically reading the manuscript and for helpful suggestions.

Note Added in Proof

After submission of this work, evidence for a hyperthermophilic last common ancestor has been presented by M. Di Giulio ("The universal ancestor lived in a thermophilic or hyperthermophilic environment." J. Theor. Biol. In press). A high G+C content of the primordial rRNAs has been inferred by parsimony analysis, in contrast to ML results by Gaultier et al. 1999.

References

Adachi J (1995) Modelling 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

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, Zheltonosova J, Chirgadze Yu, Al-Karadaghi S, Svensson LA, Liljas A (1994) Three-dimensional structure of the ribosomal translocase: Elongation factor G from *Thermus thermophilus.* EMBO J 13:3669–3677

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 eucaryotes based on the elongation factor phylogeny. Proc Natl Acad Sci USA 93:7749–7754

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

Bocchetta M, Ceccarelli E, Creti R, Sanangelantoni AM, Tiboni O, Cammarano P (1995) Arrangement and nucleotide sequence of the gene *fus* encoding elongation factor G (EF-G) from the hyperthermophilic bacterium *Aquifex pyrophilus:* Phylogenetic depth of hyperthermophilic bacteria inferred from analysis of EF-G/*fus* sequences. J Mol Evol 41:803–812

Brown JR, Masuchi FT, Robb F, Doolittle WF (1994) Evolutionary relationships of bacterial and archaeal glutamine synthetase genes. J Mol Evol 38:566–576

Budin K, Phylippe H (1998) New insights into the phylogeny of eukaryotes based on ciliate Hsp70 sequences. Mol Biol Evol 15:943–956

Burggraf S, Olsen GJ, Stetter DO, Woese CR (1992) A phylogenetic analysis of *Aquifex pyrophilus.* Syst Appl Microbiol 15:352–356

Cammarano P, Creti R, Sanangelantoni AM, Palm P (1999) The Archaea-monophily issue. A phylogeny of translational elongation factor G(2) sequences inferred from an optimized selection of alignment positions. J Mol Evol 49:524–537

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

Deckert GW, Warren PV, Gaasterland T, Young WG, Lenox AL, Grahm DE, Overbeck R, Snead MA, Keller M, Aujay M, Huber R, Feldman RA, Short JM, Olsen GJ, Sawnson RV (1998) The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus.* Nature 392:353–358

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 (1978) Cases in which parsimony or compatibility methods will be positively misleading. Syst Zool 27:403–410

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

Foster PG, Hickey DA (1999) Compositional bias may affect both DNA based and protein based phylogenetic reconstructions. J Mol Evol 48:284–290

Galtier N, Tourasse N, Gouy M (1999) A nonhyperthermophilic common ancestor to extant life forms. Science 283:220–221

Gribaldo S, Lumia V, Creti R, Conway de Macario E, Sanangelantoni A, Cammarano P (1999) Discontinuous occurrence of the *Hsp70*(*dnaK*) gene among Archaea and sequence features of HSP70 suggest a novel outlook on phylogenies inferred from this protein. J Bacteriol 181:434–443

Gupta RS (1998) Protein phylogenies and signature sequences: A reappraisal of evolutionary relationships among Archaebacteria, Eubacteria and Eukaryotes: Microbiol Mol Biol Rev 62:1435–1491

Gupta RS, Bustard K, Falah M, Singh D (1997) Sequencing of heat shock protein 70 (DnaK) homologs from *Deinococcus proteolyticus* and *Thermomicrobium roseum* and their integration in a protein-based phylogeny of prokaryotes. J Bacteriol 179:345–357

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

Hashimoto T, Sanchez LB, Shirakura T, Müller M, Hasegawa M (1998) Secondary absence of mitochondria in *Giardia lamblia* and *Trichomonas vaginalis* revealed by valyl-tRNA synthetase phylogeny. Proc Natl Acad Sci USA 95:6860–6865

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

Klenk H-P, Meier T-D, Durovic P, Schwass V, Lottspeich F, Dennis P, Zillig W (1999) RNA Polymerase of Aquifex pyrophilus: Implications for the evolution of the bacterial rpoBC operon and extremely thermophilic Bacteria. J Mol Evol 48:528–541

Lake JA, Rivera MC (1994) Was the nucleus the first endosymbiont? Proc Natl Acad Sci USA 91:1455–1459

Lazcano A, Miller S (1996) The origin and early evolution of life: Prebiotic chemistry, the pre-RNA world, and time. Cell 85:793–798

Lento GM, Hickson FE, Chambers GK, Penny D (1995) Use of spectral analysis to test hypotheses on the origin of pinnipeds. Mol Biol Evol 12:28–52

Lockhart PJ, Larkum AWD, Steel MA, Waddel PJ, Penny D (1996) Evolution of chlorophyll and bacteriochlorophyll: The problem of invariant sites in sequence analysis. Proc Natl Acad Sci USA 93: 1930–1934

Macario AJL, Dugan CB, Conway de Macario E (1991) A *dna*K homolog in the archaebacterium *Methanosarcina mazei* S-6. Gene 108:133–137

Pearson WR, Lipman DJ (1988) Improved tools for biological sequence comparisons. Proc Natl Acad Sci USA 85:2444–2448

Pesole G, Gissi C, Lanave C, Saccone C (1995) Glutamine synthetase gene evolution in Bacteria. Mol Biol Evol 12:189–197

Sanangelantoni AM, Bocchetta M, Cammarano P, Tiboni O (1994) Phylogenetic depth of S10 and *spc* operons: Cloning and sequencing of a ribosomal protein gene cluster from the extremely thermophilic bacterium *Thermotoga maritima.* J Bacteriol 176:7703–7710

Strimmer K, von Haeseler A (1996) Quartet puzzling: A quartet maximum-likelihood method for reconstructing tree topologies. Mol Biol Evol 13:964–969

Strimmer K, von Haeseler A (1997) Likelihood mapping. A simple method to visualize the phylogenetic content of a sequence alignment. Proc Natl Acad Sci USA 94:6815–6819

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

Tiboni O, Cammarano P, Sanangelantoni AM (1993) Cloning and sequencing of the gene encoding glutamine synthetase 1 from the archaeum *Pyrococcus woesei:* Anomalous phylogenies inferred from analysis of archaeal and bacterial glutamine synthetase 1 sequences. J Bacteriol 175:2961–2969

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

Woese CR (1987) Bacterial evolution. Microbiol Rev 51:221–271

Zillig W, Palm P, Klenk H-P, Langer D, Hüdepohl U, Hain J, Lanzendorfer M, Holz I (1993) Transcription in Archaea. In: Kates M, Kushner DJ, Matheson AT (eds) Biochemistry of Archaea (Archaebacteria). New comprehensive biochemistry, Vol 26. Elsevier, Amsterdam, pp 367–391