JOURNAL OF BACTERIOLOGY, Feb. 1996, p. 591–599
0021-9193/96/$04.00+0

Vol. 178, No. 3

Characterization of Uncultivated Prokaryotes: Isolation and Analysis of a 40-Kilobase-Pair Genome Fragment from a Planktonic Marine Archaeon

JEFFEREY L. STEIN,[1]* TERENCE L. MARSH,[2] KE YING WU,[3] HIROAKI SHIZUYA,[4] AND EDWARD F. DELONG[3]*

Recombinant BioCatalysis, Inc., La Jolla, California 92037[1]; Microbiology Department, University of Illinois, Urbana, Illinois 61801[2]; Department of Ecology, Evolution and Marine Biology, University of California, Santa Barbara, California 93106[3]; and Division of Biology, California Institute of Technology, Pasadena, California 91125[4]

Received 14 July 1995/Accepted 14 November 1995

One potential approach for characterizing uncultivated prokaryotes from natural assemblages involves genomic analysis of DNA fragments retrieved directly from naturally occurring microbial biomass. In this study, we sought to isolate large genomic fragments from a widely distributed and relatively abundant but as yet uncultivated group of prokaryotes, the planktonic marine *Archaea*. A fosmid DNA library was prepared from a marine picoplankton assemblage collected at a depth of 200 m in the eastern North Pacific. We identified a 38.5-kbp recombinant fosmid clone which contained an archaeal small subunit ribosomal DNA gene. Phylogenetic analyses of the small subunit rRNA sequence demonstrated its close relationship to that of previously described planktonic archaea, which form a coherent group rooted deeply within the *Crenarchaeota* branch of the domain *Archaea*. Random shotgun sequencing of subcloned fragments of the archaeal fosmid clone revealed several genes which bore highest similarity to archaeal homologs, including large subunit ribosomal DNA and translation elongation factor 2 (EF2). Analyses of the inferred amino acid sequence of archaeoplankton EF2 supported its affiliation with the *Crenarchaeote* subdivision of *Archaea*. Two gene fragments encoding proteins not previously found in *Archaea* were also identified: RNA helicase, responsible for the ATP-dependent alteration of RNA secondary structure, and glutamate semialdehyde aminotransferase, an enzyme involved in initial steps of heme biosynthesis. In total, our results indicate that genomic analysis of large DNA fragments retrieved from mixed microbial assemblages can provide useful perspective on the physiological potential of abundant but as yet uncultivated prokaryotes.

Characterization of complex microbial communities by isolation and analysis of phylogenetically informative gene sequences has been an exciting development in microbiology (27, 32). Studies using molecular phylogenetic approaches based on small subunit (ssu) rRNA sequence analyses have resulted in new estimates of the phylogenetic diversity contained within naturally occurring microbial assemblages. Entirely new phylogenetic lineages, which may sometimes represent major constituents of natural microbial communities, have been revealed by using such approaches (3, 4, 7, 13, 16, 22, 29). Although results of these studies have contributed substantially to assessments of naturally occurring microbial diversity, their utility as predictors of the physiological attributes of newly described phylotypes has been more limited. This is partly because many phylogenetically coherent prokaryote lineages, for example, the proteobacteria, often encompass a bewildering array of physiological and metabolic diversity. The problem is further complicated by the fact that many of the newly discovered phylotypes revealed by molecular phylogenetic analysis are proving difficult to enrich and isolate in pure culture. These limitations suggest the need for alternative approaches to characterize the physiological and metabolic potential of as yet uncultivated microorganisms, which to date have been characterized solely by analyses of PCR-amplified rRNA gene fragments, clonally recovered from mixed assemblage nucleic acids.

Recent work has demonstrated that prokaryotes of the domain *Archaea* (35, 37) are more phylogenetically diverse and ecologically widespread than has been previously suspected (3, 7, 13, 24). Marine planktonic archaea have now been shown to occur in relatively high abundance in a variety of marine planktonic environments. For instance, in surface waters of Antarctica during late winter, as well as in subsurface waters of coastal temperate regions, planktonic marine archaeal rRNA can account for a significant fraction (5 to 35%) of the total picoplankton rRNA (7, 8). These and other data indicate that archaeoplankton are widespread, abundant components of marine picoplankton and presumably are important participants in microbially mediated cycling of energy and matter in the sea. Specific relatives of the crenarchaeote-affiliated marine archaea have also been detected in soil microbial assemblages indicating their further diversification (33a).

Although the genotypic and phenotypic properties of marine pelagic archaea are unknown at present, their abundance and ecological distribution suggest that they may represent entirely new phenotypic groups within the domain *Archaea*. Available data on these microorganisms consist solely of quantitative estimates of their abundance and distribution derived from ssu rRNA hybridization experiments (7, 8) or of the recovery and sequence analysis of PCR-amplified rRNA gene fragments (7,

* Corresponding author. Mailing address for Jefferey L. Stein: Recombinant BioCatalysis, Inc., 505 Coast Blvd. South, La Jolla, CA 92037. Electronic mail address: jstein@vaxkiller.agi.org. Mailing address for Edward F. DeLong: Department of Ecology, Evolution and Marine Biology, University of California, Santa Barbara, CA 93106. Electronic mail address: delong@marbtech.lscf.ucsb.edu.

8, 13, 14, 24). To date, no successful attempts at cultivation and characterization of these microorganisms has been reported.

In an effort to obtain additional information on marine planktonic archaea, and to avoid uncertainties and limitations imposed by PCR amplification or cultivation approaches, we attempted to archive and stably propagate large fragments of genomic DNA from these as yet uncultivated prokaryotes. Using a recently constructed cloning vector derived from the *Escherichia coli* F factor, the fosmid (for F-factor-based cosmid [20]), we accessed genomic information from naturally occurring picoplankton by archiving their genome fragments in the form of a stable environmental DNA library. This report describes the construction of an environmental library from coastal marine picoplankton (planktonic microorganisms approximately 0.2 to 2 μm in size) and preliminary genomic characterization of as yet uncultivated marine archaea.

MATERIALS AND METHODS

Cell collection and preparation of DNA. Agarose plugs containing concentrated picoplankton cells were prepared from samples collected on an oceanographic cruise from Newport, Oregon, to Honolulu, Hawaii, in the fall of 1992. Seawater (30 liters) was collected in Niskin bottles, screened through 10-μm-pore-size Nitex filters, and concentrated by hollow fiber filtration (Amicon DC10) through 30,000-molecular weight-cutoff polysulfone filters. The concentrated bacterioplankton cells were collected on a 0.22 μm-pore-size, 47-mm-diameter Durapore filter and resuspended in 1 ml of 2× STE buffer (1 M NaCl, 0.1 M EDTA [pH 8.0], 10 mM Tris [pH 8.0]) to a final density of approximately 10^{10} cells per ml. The cell suspension was mixed with 1 volume of 1% molten SeaPlaque LMP agarose (FMC), cooled to 40°C, and then immediately drawn into a 1-ml syringe. The syringe was sealed with Parafilm and placed on ice for 10 min. The cell-containing agarose plug was extruded into 10 ml of lysis buffer (10 mM Tris [pH 8.0], 50 mM NaCl, 0.1 M EDTA, 1% Sarkosyl, 0.2% sodium deoxycholate, 1 mg of lysozyme per ml) and incubated at 37°C for 1 h. The agarose plug was then transferred to 40 ml of ESP buffer (1% Sarkosyl–1 mg of proteinase K per ml in 0.5 M EDTA), and incubated at 55°C for 16 h. The solution was decanted, replaced with fresh ESP buffer, and incubated at 55°C for an additional hour. The agarose plugs were then placed in 50 mM EDTA and stored at 4°C shipboard for the duration of the oceanographic cruise. After disembarkation, samples were transported in insulated containers containing ice packs to the University of California, Santa Barbara. At some collection sites, parallel samples were collected on Durapore filters and frozen for later analysis of relative archaeal rRNA abundance as previously described (7, 8, 28, 31).

Screening of agarose plugs for archaeal DNA. Lysates were prepared from the agarose plugs, and the DNA was screened for the presence of archaeal ssu ribosomal DNA (rDNA), using eubacterium- and archaeon-biased oligonucleotide primers and PCR amplification conditions previously described (7). Approximately 400 μl of each agarose plug was melted at 65°C for 10 min, cooled to 40°C, and digested for 1 h with 2 U of Gelase (Epicentre) as recommended by the manufacturer. After digestion, the lysate was extracted once with phenol-chloroform-isoamyl alcohol (24:24:1), washed twice with approximately 10 volumes of sterile water (Millipore) in a microconcentrater (Centricon 100; Amicon), and resuspended in a final volume of 100 μl of TE (10 mM Tris [pH 8.0], 1 mM EDTA). One microliter of each extract was used in subsequent eubacterium- or archaeon-specific ssu rDNA PCR amplifications (7). Amplification products were examined by agarose gel electrophoresis.

Preparation and cloning of picoplankton DNA. A picoplankton sample collected at a depth of 200 m off the Oregon coast (44°02.72′N, 124°57.30′W) on August 29, 1992, was found to contain relatively high levels of archaeal DNA, using the archaeal PCR assay. Quantitative hybridization experiments using archaeon-specific oligonucleotide probes indicated that ≈4.7% of the total rRNA in this sample was archaeal in origin (8). One slice of an agarose plug (72 μl) prepared from this sample was dialyzed overnight at 4°C against 1 ml of buffer A (100 mM NaCl, 10 mM bis Tris propane-HCl, 100 μg of acetylated bovine serum albumin per ml [pH 7.0 at 25°C]) in a 2-ml microcentrifuge tube. The solution was replaced with 250 μl of fresh buffer A containing 10 mM $MgCl_2$ and 1 mM dithiothreitol and incubated on a rocking platform for 1 h at room temperature. The solution was then changed to 250 μl of the same buffer containing 4 U of *Sau*3AI (New England Biolabs), equilibrated to 37°C in a water bath, and then incubated on a rocking platform in a 37°C incubator for 45 min. The plug was transferred to a 1.5-ml microcentrifuge tube and incubated at 68°C for 30 min to inactivate the enzyme and melt the agarose. The agarose was digested and the DNA was dephosphorylated by using Gelase and HK phosphatase (Epicentre), respectively, as recommended by the manufacturer. Protein was removed by gentle phenol-chloroform extraction, and the DNA was ethanol precipitated, pelleted, and then washed with 70% ethanol. This partially digested DNA was resuspended in sterile H_2O to a concentration of 25 ng/μl for ligation to the pFOS1 vector.

Vector arms were prepared from pFOS1 as described previously (20). Briefly, the plasmid was completely digested with *Aat*II, dephosphorylated with HK phosphatase, and then digested with *Bam*HI to generate two arms, each of which contained a *cos* site in the proper orientation for cloning and packaging ligated DNA between 35 and 45 kbp in size. The partially digested picoplankton DNA was ligated overnight to the pFOS1 arms in a 15-μl ligation reaction mixture containing 25 ng each of vector and insert and 1 U of T4 DNA ligase (Boehringer Mannheim). The ligated DNA in 4 μl of this reaction mixture was in vitro packaged by using the Gigapack XL packaging system (Stratagene), the fosmid particles were transfected to *Escherichia coli* DH10B (Bethesda Research Laboratories), and the cells were spread onto LB_{cm15} plates. The resultant fosmid clones were picked into 96-well microtiter dishes containing LB_{cm15} supplemented with 7% glycerol. This library was stored frozen at −80°C for later analysis.

Screening for archaeal clones. Multiplex PCR screening using archaeon-biased ssu rRNA-targeted oligonucleotide primers was used to identify fosmid clones containing archaeal rRNA genes (7). Pools of all clones contained in a single microtiter dish were inoculated into 5 ml of SOB medium containing 12.5 μg of chloramphenicol per ml and grown overnight at 37°C. Fosmids of each pool were extracted by using a standard alkaline lysis procedure (20). One microliter of each fosmid DNA pool, representing one microtiter plate, was used in 50-μl PCRs performed with archaeon-biased ssu rDNA primers. Microtiter dishes which tested positive with the archaeal rDNA primers were further screened by PCRs performed on row and column fosmid pools from that microtiter dish.

For further screening, high-density filter replicas of the fosmid library were generated by stamping the contents of the microtiter dishes in a five-by-five array onto nylon filters, using a Biomek 1000 workstation with attached high-density replicating tool (Beckman). The filters were stamped on a bed of LB_{cm15} agar in a microtiter dish lid and allowed to incubate overnight at 37°C to allow colony growth. Colonies were lysed, and DNA was fixed on the filters by using a turbo prep protocol in which the filters were laid on Whatman paper saturated with 2× SSC (1× SSC is 0.15 M NaCl plus 0.015 M sodium citrate)–5% sodium dodecyl sulfate (SDS) for 2 min and then microwaved at the highest power setting for 2.5 min on a turntable. The filters were then submerged in 20 ml of a solution containing 50 mM Tris-Cl (pH 8.0), 50 mM EDTA, 100 mM NaCl, 1% (wt/vol) sodium lauryl sarcosine (Sarkosyl), and 250 μg of proteinase K per ml and incubated at 45°C for 30 min. After removal from the solution, the filters were air dried at 55°C and the DNA was UV cross-linked with a Stratalinker (Stratagene).

An archaeal rDNA-containing fosmid clone (clone 4B7) was identified, purified, and subsequently digested with *Not*I to excise the cloned insert and separated on a 1% SeaKem agarose gel (FMC), using a CHEF DRII pulsed-field gel electrophoresis apparatus (Bio-Rad). The electrophoresis conditions were as follows: 198 V, 5-s switch interval, 20-h run time, in 1× Tris-acetate-EDTA buffer at 14°C. The cloned inserts were visualized with ethidium bromide staining, excised from the gel, and purified by using a Geneclean kit (Bio 101). Approximately 25 ng of the fragment was labeled with ^{32}P by random priming (Prime-It II; Stratagene) and hybridized to the high-density filters to identify potential contigs of the 16S containing clones. The filters were first prehybridized 2 h at 65°C in a blocking solution containing 5× SSC, 1.0% blocking reagent (Boehringer Mannheim), 0.1% Sarkosyl, and 0.02% SDS. The labeled insert was added to this solution and allowed to hybridize overnight at the same temperature. After hybridization, the filters were washed twice for 15 min each time at room temperature in 2× SSC–0.1% SDS and then twice for 15 min each time at 65°C in 0.5× SSC–0.1% SDS. The filters were then exposed to autoradiographic film (Kodak X-Omat) for 3 to 16 h at −80°C.

Subcloning and sequencing. A subclone library of fosmid 4B7 was prepared by digesting fosmid DNA with either *Eco*RI, *Eco*RI and *Bam*HI, or *Spe*I. The restriction fragments were subcloned into Bluescript vector SK+, which had been digested with the corresponding restriction enzymes and subsequently treated with HK phosphatase. Plasmid minipreparations of the fosmid subclones were prepared by alkaline lysis, and 200 to 300 bases of nucleotide sequence was determined for each end of the insert, using M13 forward and reverse primers and Sequenase 2.0 (United States Biochemical). Partial fosmid subclone sequences were compared with entries in the nonredundant protein and nucleic acid databases of the National Center for Biotechnology Information server, using BLASTN or BLASTX (1, 17).

The entire ssu rRNA gene contained on subclone 2i was sequenced in the forward and reverse directions, using rDNA-targeted oligonucleotide primers (21) and Sequenase 2.0 (United States Biochemical) as recommended by the manufacturer for double-stranded plasmid sequencing. The nucleotide sequence of the translation elongation factor 2 (EF2) gene was contained on two overlapping subclones of fosmid 4B7, subclone 3B (amino acid residues 410 to 725), and subclone B2 (amino acid residues 1 to 705). Sequencing was performed in the forward and reverse directions by primer walking, using the following oligodeoxynucleotide primers: EF3B-1r (CTC AGG AAC CTC ACC), EF3B-1f (CAA ATA TTA CGT GAT), EF2-2f (GTT ATT GCA CAC GTA), EF2-2r (TTC TTA CCT GTT CTT TTG), EF-3f (CGA GTA AGA CCT GTA), EF-3r (CAG AAA AGT CAA CGT), EF-4f (GGT GTT ATT GCA CAC), EF-4r (TAA TAC ACC CAT TCC), EF-5f (TTT GGT TCT GCA AAA), EF-6f (GCA CAA AAG TAT AGA), EF-5r (ACC AGT CTC TTC ATC), EF-6r (GTG ATT GTT TCT CTG), EFend-f (CAA GGT AAA CGT GGT AAA G), and EFend-r (GGT TTT GCC ATG ATC).

Restriction mapping. Large genomic fragments isolated from fosmid clones were mapped by partial and double digestion with various restriction endonucleases. When the subclone sizes exceeded 10 kb, the F-factor-based vector pBAC108L (30) was used to accommodate the fosmid subfragments. Partial digestions were performed by adding 2.5 U of restriction enzyme to 1 μg of *Not*I-digested clone DNA in a 30-μl reaction mixture. The reaction mixture was incubated at 37°C, and 10-μl aliquots were removed at 10, 40, and 60 min. Restriction digestions were terminated by adding 1 μl of 0.5 M EDTA to the reaction mixtures and placing the tubes on ice. The partially digested DNA was separated by pulsed-field gel electrophoresis as described above except using a 1- to 3-s ramped switch time at 100 V for 16 h. The sizes of the separated fragments were determined relative to those of known standards. The distances of the restriction sites relative to the terminal T7 and SP6 promoter sites on the excised cassette were determined by end labeling 10 pmol of T7- or SP6-specific oligonucleotides with [γ-^{32}P]ATP (7,000 Ci/mmol) and hybridizing with Southern blots of the gels.

Southern blots of agarose gels containing fosmid and pBAC clones digested with two or more restriction enzymes were probed with labeled T7 and SP6 oligonucleotides as well as random-prime-labeled subclones and PCR fragments carrying gene sequences identified from the shotgun sequencing described above. This information was correlated with the size estimates from the partial digestions to generate physical and genetic maps of the fosmids and their subclones.

Phylogenetic analysis. Sequence alignment and DeSoete distance (9) analyses were performed on a Sun Sparc 10 workstation using GDE 2.2 and Treetool 1.0, obtained from the ribosomal database project (RDP) (23). DeSoete least squares distance analyses (9) were performed by using pairwise evolutionary distances, calculated by using the correction of Olsen to account for empirical base frequencies (34). Reference sequences were obtained from the RDP, version 4.0 (23). Maximum likelihood analyses (10) of ssu rRNA sequences were performed by using fastDNAml 1.0 (25), obtained from the RDP. For distance analyses of the inferred amino acid sequence of EF2, evolutionary distances were estimated by using the Phylip program (12) Protdist, and tree topology was inferred by the Fitch-Margoliash method, using random taxon addition and global branch swapping. For maximum parsimony analyses of protein sequences, the Phylip program Protpars was used with random taxon addition and ordinary parsimony options.

Nucleotide sequence accession numbers. Partial sequences reported in Table 1 have been submitted to GenBank under the following accession numbers: U40238, U40239, U40240, U40241, U40242, U40243, U40244, and U40245. The nucleotide sequences encoding ssu rRNA and EF2 have been submitted to GenBank under accession numbers U39635 and U41261.

RESULTS

Figure 1 shows an overview of the procedures used to construct an environmental library from the mixed picoplankton sample. Our goal was to construct a stable, large insert DNA library representing picoplankton genomic DNA, in order to gain information about the genetic and physiological potential of one constituent group in this community, the planktonic marine *Crenarchaeota*. Agarose plugs containing high-molecular-weight picoplankton DNA were prepared by concentrating cells from 30 liters of seawater, using hollow fiber filtration. These agarose plugs, representing picoplankton collected from a variety of sites and depths in the eastern North Pacific, were screened for the presence of archaebacteria by using both eubacterium-biased (to test for positive amplification) and archaeon-biased rDNA primers. PCR amplification results from several of the agarose plugs (data not shown) indicated the presence of significant amounts of archaeal DNA. Quantitative hybridization experiments using rRNA extracted from one sample, collected at a depth of 200 m off the Oregon coast, indicated that planktonic archaea in this assemblage comprised approximately 4.7% of the total picoplankton biomass (this sample corresponds to "PAC1"-200 m in Table 1 of reference 8). Results from archaeon-biased rDNA PCR amplification performed on agarose plug lysates confirmed the presence of relatively large amounts of archaeal DNA in this sample. Agarose plugs prepared from this picoplankton sample were chosen for subsequent fosmid library preparation. Each 1-ml agarose plug from this site contained approximately 7.5×10^9 cells; therefore, approximately 5.4×10^8 cells were present in the 72-μl slice used in the preparation of the partially digested DNA.

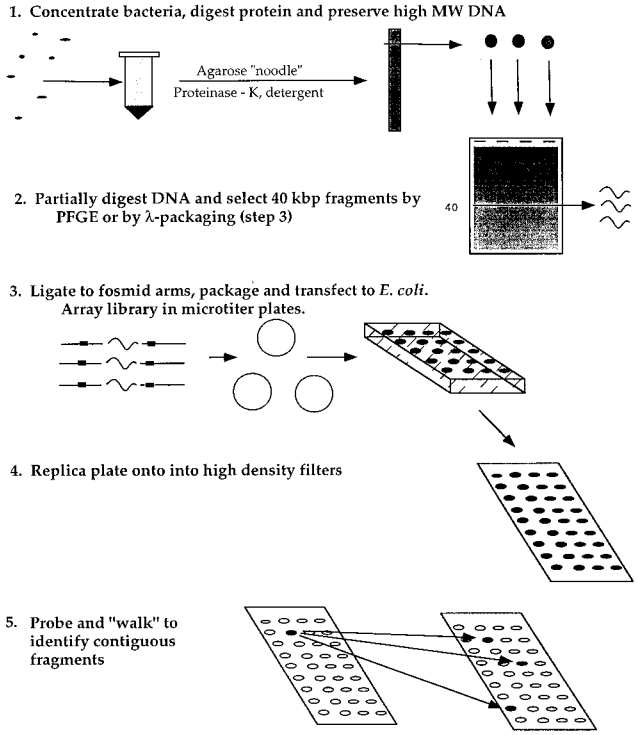

FIG. 1. Flowchart depicting the construction and screening of an environmental library from a mixed picoplankton sample. MW, molecular weight; PFGE, pulsed-field gel electrophoresis.

Recombinant fosmids, each containing ca. 40 kb of picoplankton DNA insert, yielded a library of 3,552 fosmid clones, containing approximately 1.4×10^8 bp of cloned DNA. All of the clones examined contained inserts ranging in size from 38 to 42 kbp (Fig. 2). Both the multiplex PCR (Fig. 3) and the hybridization experiments suggested that well B7 on microtiter

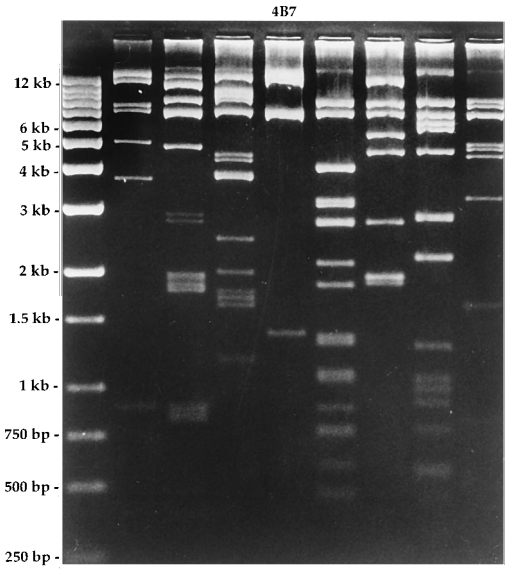

FIG. 2. Pulsed-field gel showing the separation of selected fosmid clones digested with *Not*I and *Bam*HI. The pFOS1 vector band is at 7.2 kbp. The top two bands of clone 4B7 are doublets.

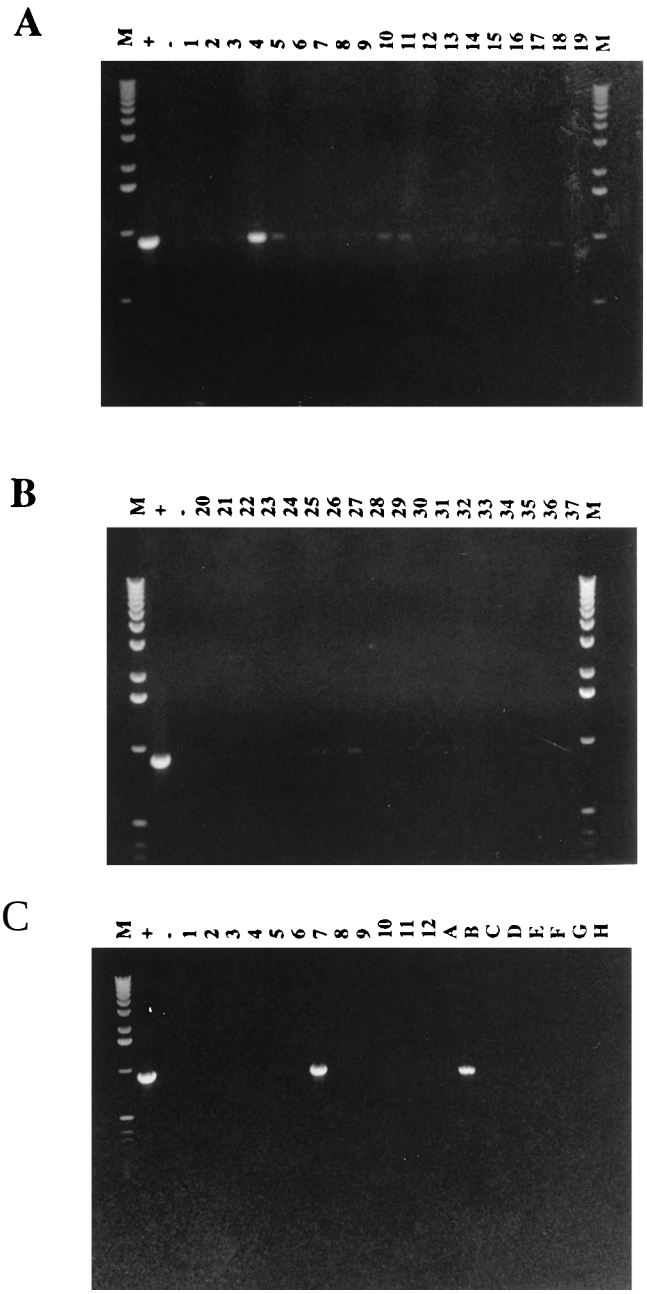

FIG. 3. Multiplex PCR analysis of the fosmid bacterioplankton DNA library. (A and B) Agarose gel electrophoreses of fosmid minipreparations pooled from each microtiter dish in the library (dishes 1 to 37) and amplified with archaeon-biased ssu rRNA-specific primers. Lane M, 1-kb molecular size marker (Bethesda Research Laboratories); lane +, positive control containing archaeal genomic DNA (*Haloferax volcanii*); lane −, negative control containing eubacterial genomic DNA (*Shewanella putrefaciens*). (C) Agarose gel electrophoresis of archaeon-biased ssu rRNA PCR amplifications of fosmid clones pooled from columns (1 to 12) and rows (A to H) of microtiter dish 4 (positive reaction in panel A). Positive reactions were detected in microtiter dish 4, row B, column 7 (clone 4B7).

dish 4 contained a clone encoding a 16S rDNA gene specific to the archaea. This clone, designated 4B7, was selected for more detailed examination. Using the cloned 4B7 fragment as a probe, we did not detect other cloned fragments in the fosmid library that contained overlapping regions (Fig. 4). No other archaeal ssu rRNA genes were detected in the library.

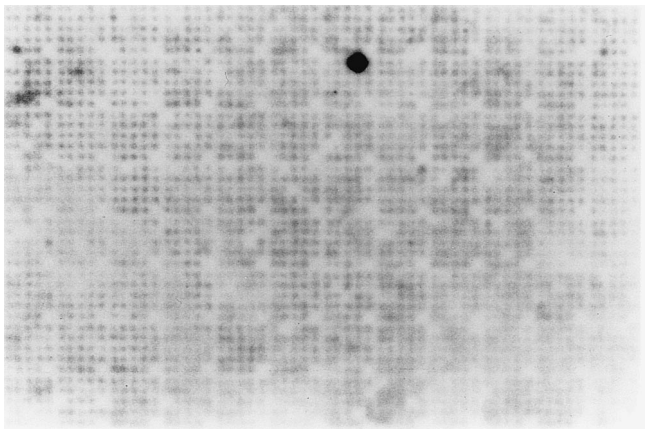

FIG. 4. High-density filter replica of 2,304 fosmid clones containing approximately 92 million bp of DNA cloned from the mixed picoplankton community. The filter was probed with the labeled insert from clone 4B7 (dark spot). The lack of other hybridizing clones suggests that contigs of 4B7 are absent from this portion of the library. Similar experiments with the remainder of the library yielded similar results.

Sequence analysis and identification of protein- and rRNA-encoding genes. Fragments of archaeal DNA contained in the 4B7 insert were subcloned by digesting the fosmid with either *Eco*RI, *Eco*RI plus *Bam*HI, or *Spe*I. The resulting restriction fragments were recovered in Bluescript vector (Stratagene), and the distal 200 to 300 nucleotides of each subclone were sequenced by using M13 forward and reverse sequencing primers. Of a total of 18 subclones analyzed in this fashion, 6 (33%) contained nucleotide sequences with significant identity to previously characterized genes archived in the National Center for Biotechnology Information nucleic acid or protein nonredundant database. Table 1 shows the rank order similarity of subclone sequences which share significant similarity to known sequences, based on Poisson probabilities of random homology [$P(n)$] (1, 17). Putative genes contained on fosmid 4B7 identified in this fashion include EF2, glutamate 1-semialdehyde aminotransferase (GSAT), RNA helicase, DnaJ, ssu rRNA, and large subunit (lsu) rRNA. Three of these nucleotide sequences (EF2, lsu rRNA, and ssu rRNA) are most similar to known archaeal homologs. Relatively high $P(n)$ values obtained with the DnaJ (subclone 22K-R) and lsu rRNA (Spe14-R) sequences are due to the fact that only a small portion of each of these subcloned fragments encodes the indicated homolog.

Gene organization. The results of the sequence analyses of subclones containing the 16S-23S rRNA operon, GSAT, EF2, and RNA helicase were used to confirm the locations of restriction sites determined by partial and double digestions. The sequenced genes mapped to the distal *Bam*HI-*Not*I fragments of the 38.5-kbp insert in clone 4B7 (Fig. 5). The GSAT gene and the 16S-23S operon reside on one of these fragments and are transcribed in the same direction. We have not found evidence for a 5S rRNA gene encoded on the 4B7 clone. The genes encoding RNA helicase and EF2 were on the distal *Bam*HI-*Not*I fragment. The distal location of the archaeoplankton EF2 gene relative to the rRNA operon on clone 4B7, as well as the striking similarities of fosmid-encoded EF2 and ssu and lsu rRNA genes to archaeal homologs (Table 1), provides evidence that the 4B7 fosmid insert represents a contiguous genomic fragment from a planktonic marine archaeon.

Phylogenetic analysis of fosmid-encoded ssu rRNA and EF2. Southern blot hybridization using probes produced from sub-

TABLE 1. Protein- and rRNA-encoding genes revealed shotgun sequencing[a]

Subclone	Putative gene	Total positions	Rank	Species	*P*(*n*)
3B-R	EF2	189	1	*Sulfolobus solfataricus*	5.0e-10
			2	*Desulfurococcus mobilis*	6.5e-9
			3	*Sulfolobus acidocaldarius*	3.0e-7
			4	*Chlorella kessleri*	4.0e-5
			5	*Methanococcus vannielii*	5.0e-5
3B-F	EF2	298	1	*Sulfolobus solfataricus*	7.3e-23
			2	*Sulfolobus acidocaldarius*	6.5e-22
			3	*Thermoplasma acidophilum*	1.4e-19
			4	*Desulfurococcus mobilis*	1.8e-18
			5	*Pyrococcus woessi*	2.2e-17
2I-R	GSAT	145	1	*Arabidopsis thaliana*	2.5e-12
			2	Soybean	3.3e-12
			3	*Synechococcus* sp. strain 6301	4.0e-12
			4	Tobacco	8.9e-12
			5	Barley	3.0e-11
2E-F	RNA helicase	238	1	*Escherichia coli*	1.5e-23
			2	*Klebsiella pneumoniae*	1.5e-23
			3	*Streptococcus pneumoniae*	2.1e-22
			4	*Drosophila melanogaster*	2.9e-18
			5	*Saccharomyces cerevisiae*	5.4e-18
22K-R	DnaJ heat shock protein	77	1	*Clostridium acetobutylicum*	4.7e-4
			2	*Methanosarcina mazeii*	2.1e-3
			3	*Mycoplasma genitalium*	3.5e-2
			4	*Caulobacter crescentus*	3.8e-2
			5	*Synechococcus* sp.	4.5e-2
Spe14-F	lsu rRNA	131	1	*Archaeoglobus fulgidus*	1.0e-11
			2	*Solfolobus solfataricus*	1.3e-9
			3	*Solfolobus acidocaldarius*	6.9e-9
			4	*Methanococcus vannielii*	2.8e-5
			5	*Rickettsia prowazekii*	9.1e-5
Spe14-R	lsu rRNA	76	1	*Methanosarcina frisius*	9.1e-6
			2	*Methanothermus fervidus*	9.3e-6
			3	*Rhodothermus marinus*	9.7e-5
			4	*Halococcus morrhuae*	1.2e-4
			5	*Desulfurococcus mobilis*	3.2e-4
Spe18-F	ssu rRNA	177	1	*Methanobacterium thermoautotrophicum*	4.5e-24
			2	*Methanothermus fervidus*	6.6e-23
			3	*Pyrococcus* sp.	1.7e-22
			4	*Methanospirillum hungatei*	1.3e-20
			5	*Archaeoglobus fulgidus*	3.1e-20

clones containing either 23S or 16S rRNA (Table 1) indicated that the ssu and lsu rRNA genes are adjacent to one another on fosmid 4B7 (Fig. 5). Phylogenetic analysis indicated the close relationship of the cloned rRNA gene sequence to sequences of a group of marine planktonic *Crenarchaeota* previously identified in PCR-based, molecular phylogenetic environmental surveys (7, 8, 13) (Fig. 6). The GC content of the rDNA gene found on fosmid 4B7 is relatively low (51.7% GC), similar to that found in PCR-amplified, planktonic crenarchaeotal rDNA fragments. Sequence similarities between the fosmid-encoded ssu rRNA and PCR-amplified rDNA fragments previously retrieved from other planktonic *Crenarchaeota* (7, 8, 13) ranged from 92 to 95% unrestricted sequence similarity. Maximum likelihood analyses of the full ssu rRNA sequence placed the fosmid-encoded ssu rRNA sequence within the *Crenarchaeota* branch of the domain *Archaea* in 217 (87%) of 250 bootstrap replicate analyses (Fig. 6B).

The EF2 amino acid sequence bore highest overall similarity with the EF2 genes of *Sulfolobus solfataricus* and *Sulfolobus acidocaldarius*, as indicated by sequence similarities (49.1 and 47.8% similarity, counting conservative replacements, respectively) and P(*n*) values (1). The DNA sequence of fosmid-encoded EF2 was also most similar to that of *S. solfataricus* and *S. acidocaldarius*. Conserved sequence motifs characteristic of EF2s and other GTP-binding proteins are also present and conserved in the inferred planktonic archaeal EF2 protein sequence (Fig. 7). In addition, a conserved amino acid sequence motif surrounding the diphtheria toxin-sensitive, ADP-ribosylatable histidine residue (2, 5) is also present in the amino acid sequence of the cloned EF2 (Fig. 7). Both distance and parsimony analyses of the amino acid sequence of EF2 also indicated that the planktonic archaeal EF2 sequence is most closely related to other crenarchaeotal EF2 sequences, at high bootstrap confidence levels (90 and 74%, respectively) (Fig. 8).

DISCUSSION

We initiated this study to further understand the genetic properties, physiological potential, and ecological significance of planktonic marine archaea, which previously had been characterized solely on the basis of ssu rRNA gene sequences. Our approach consisted of cloning random fragments of picoplankton DNA into an F-factor-based fosmid cloning vector, which provided an environmental library composed of the genomic DNA from the mixed picoplankton assemblage. Because the pFOS1 vector was designed to accommodate and propagate large genomic fragments with a high degree of fidelity (20, 30), we reasoned that the composition of such a library would likely reflect the composition of the sample and that the individual clones would accurately preserve the arrangement of cloned genes. Since an average-size fosmid clone represents approximately 2% of a typical archaeal genome, our approach has

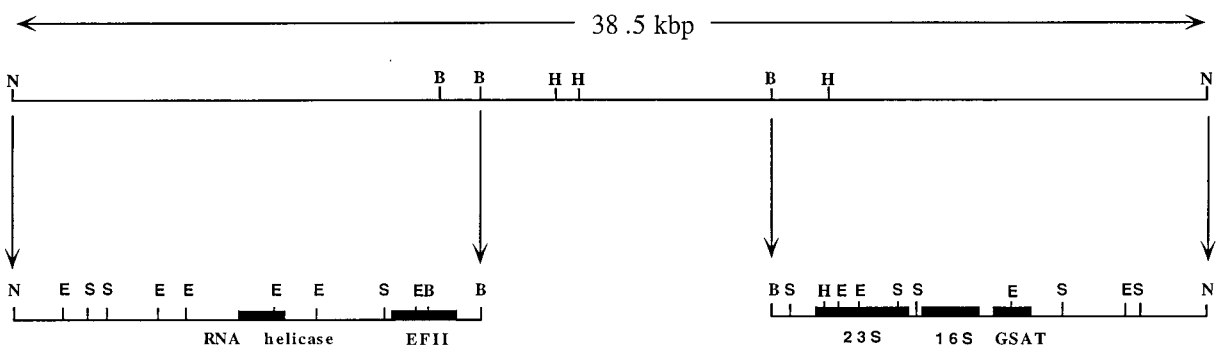

FIG. 5. Genetic map of clone 4B7 showing the positions of genes encoding RNA helicase, EF2 (EFII), 23S and 16S rRNAs, and GSAT relative to sites for *Eco*RI (E), *Spe*I (S), *Not*I (N), *Bam*HI (B), and *Hin*dIII (H).

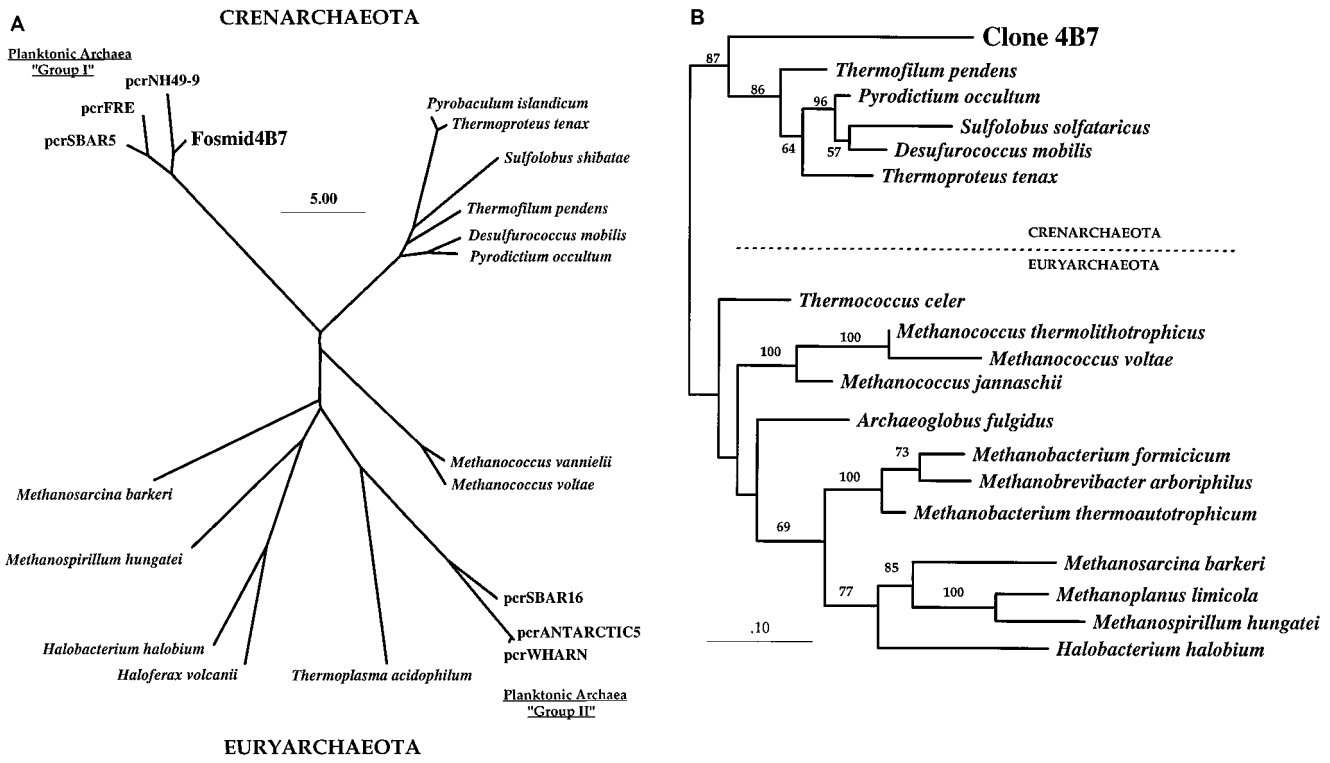

FIG. 6. Phylogenetic analysis of the fosmid-encoded ssu rRNA sequence. (A) DeSoete least squares distance analyses (9) were performed by using pairwise evolutionary distances, calculated by using the correction of Olsen to account for empirical base frequencies (25). Analyses were performed with the programs GDE 2.2 and Treetool 1.0, obtained from the RDP (23). Reference sequences, obtained from the RDP (version 4.0), include rDNA sequences from cultivated archaea as well as PCR-amplified rDNA fragments retrieved from mixed plankton assemblages (indicated as pcr-). The scale bar represents five nucleotide substitutions per 100 sequence positions. (B) Bootstrap maximum likelihood analysis (10) of the ssu rRNA gene contained on fosmid 4B7, performed by using fastDNAml 1.0 (25). Analyses were performed by using empirical base frequencies on a total of 1,295 unambiguously alignable sequence positions. Numbers represent percentage of 250 bootstrap replications (11) that support the branching pattern appearing to the right of the value. *Thermotoga maritima* was used as an outgroup in the analysis. The scale bar corresponds to the expected number of changes per sequence position for those positions changing at the median rate.

allowed us to identify intact genes, operons, and their control regions on a single cloned fragment.

Using the archaeal ssu rDNA as a phylogenetic marker, we identified a 38.5-kb fragment in clone 4B7 that contains an archaeal rDNA operon. Several different screening techniques suggested that 4B7 was the only archaeal rDNA-containing clone in the entire library. Assuming that the average genome size of planktonic archaea is about half of that of eubacterial picoplankton and that archaeal cells comprised a maximum of 5% of the total population sampled, a total of 89 archaeal clones would be expected in our fosmid library composed of 3,552 clones. If planktonic marine *Crenarchaeota* contain only one rDNA operon per genome, as do all other characterized *Crenarchaeota* (33), then approximately 1/50 of the total archaeal clones [(40,000 bp per clone)/(2×10^6 bp/1 rDNA operon)] should contain an rDNA operon. If the above-mentioned assumptions concerning rDNA copy number and planktonic archaeal genome size are correct, the recovery of only a single archaeal rDNA-containing clone in a library containing a total of 3,552 40-kbp clones is not unexpected.

Phylogenetic analyses of the rRNA gene encoded on fosmid 4B7 indicated its affiliation with a crenarchaeotal group of *Archaea*, members of which were previously identified solely by analyses of PCR-amplified rDNA fragments (7, 8, 13, 24) (Fig. 6A). Maximum likelihood analysis of the full ssu rRNA sequence encoded on fosmid 4B7 support the affiliation of this group with the *Crenarchaeota* at reasonably high bootstrap confidence levels (87%; Fig. 6B and references 11 and 38). This result is consistent with transversion signature analyses of rDNA sequences of members of this archaeal group (35a) as well as analyses of the higher-order structure of the rDNA operon contained on fosmid 4B7 (23a, 33). However, large differences in base composition between marine planktonic *Crenarchaeota* rDNA (≈51% G+C) and that of their putative thermophilic crenarchaeotal relatives (≈62 to 67% G+C [36]) complicates interpretation of rDNA-based phylogenies. In this context, phylogenetic analyses of inferred amino acid sequences of protein-encoding genes may be useful for resolving phylogenetic affiliation of this divergent and novel marine archaeal group. EF2 sequences are found in all cellular organisms, are highly conserved, and have proven particularly useful for constructing global phylogenies (5, 6, 19). Analyses of the EF2-encoding sequence on fosmid 4B7 (Fig. 7) allowed an independent assessment of the phylogenetic placement of the planktonic *Crenarchaeota*. Both distance and parsimony analyses of the EF2 amino acid sequence indicated its specific affiliation with EF2 sequences of the *Crenarchaeota* (Fig. 7). The congruence between the rRNA gene and EF2 protein phylogenies strongly supports the close relationship of this planktonic marine archaeal group with the hyperthermophilic *Crenarchaeota.*

Neither RNA helicase nor GSAT has been previously found in archaea. RNA helicase belongs to the DEAD-box proteins that catalyze the ATP-dependent alteration of RNA secondary structure. This family of proteins has been implicated in splicing, translation, cell growth, and development in eubacteria

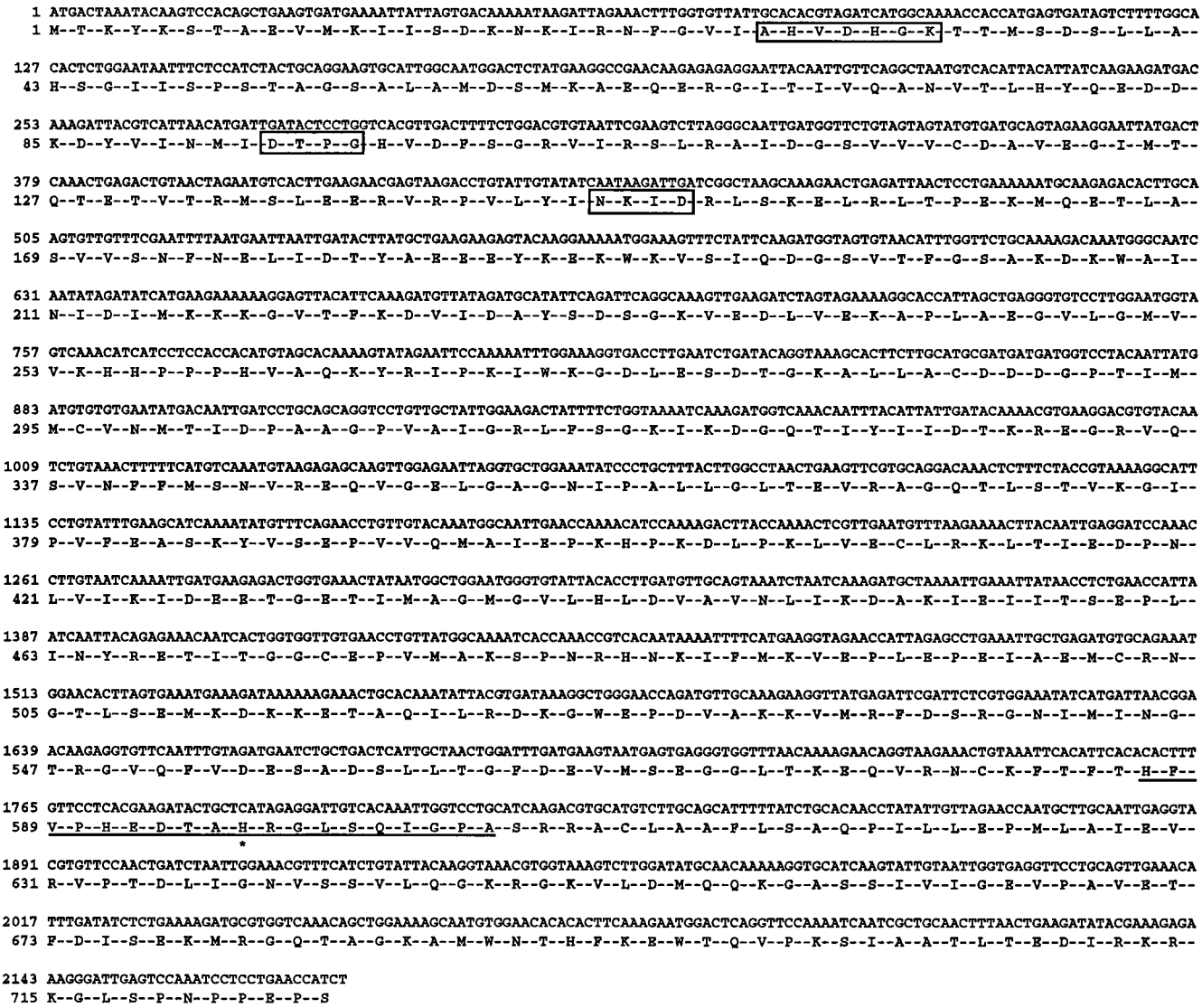

FIG. 7. Nucleotide and inferred amino acid sequences of fosmid 4B7-encoded EF2. The conserved GTP-binding motifs found in other EF2 and EFG homologs (AHVDHGK, DTPG, and NKID [2, 5]) are boxed. The histidine site susceptible to ADP ribosylation by diphtheria toxin in other EF2s is marked by an asterisk, and the surrounding, conserved amino acid residues are underlined.

and eukaryotes (26), and all members described to date share a highly conserved type B motif, LDEADxxL. The archaeoplankton RNA helicase encoded on clone 4B7 contains this DEAD-box motif as well as several other conserved regions diagnostic of the family. It has recently been shown that another member of this family, DbpA from *E. coli*, will hydrolyze ATP in the presence of *E. coli* 23S rRNA but not in the presence of eukaryotic rRNA (15). It will be of interest to determine whether the archaeal RNA helicase can distinguish rRNAs from archaea, bacteria, and eucarya as well. GSAT is involved in the synthesis of 5-aminolevulinic acid via one of two pathways: either a C_4 pathway found in animals, yeasts, and some bacteria or a C_5 pathway found in higher plants (18). The C_5 pathway leads to the biosynthesis of chlorophyll, and it is noteworthy that the archaeoplankton GSAT gene appears strongly affiliated with homologs from photoautotrophs, including higher plants (Table 1).

Our results have interesting general implications with regard to the physiological diversity and ecological distribution of the *Crenarchaeota*. In total, the available phylogenetic and ecological data suggest that a hyperthermophilic ancestor of planktonic *Crenarchaeota*, perhaps originating from a hydrothermal vent habitat, adapted to growth at low seawater temperatures, giving rise to the planktonic *Crenarchaeota*. However, another alternative interpretation of the deep placement of planktonic archaeal 16S rRNA and EF2 genes within the crenarchaeotal branch is that the *Crenarchaeota* as a group have a nonthermophilic origin. This latter alternative seems more unlikely, however, in light of several newly discovered hyperthermophilic archaeal rDNA sequences (3) specifically related to the planktonic archaeal rDNAs, which root more deeply within the *Crenarchaeota* than do the planktonic marine crenarchaeotes (2a).

Analysis of large genomic fragments from as yet uncultivated microorganisms can provide information on their genetic composition, organization, and physiological potential as well

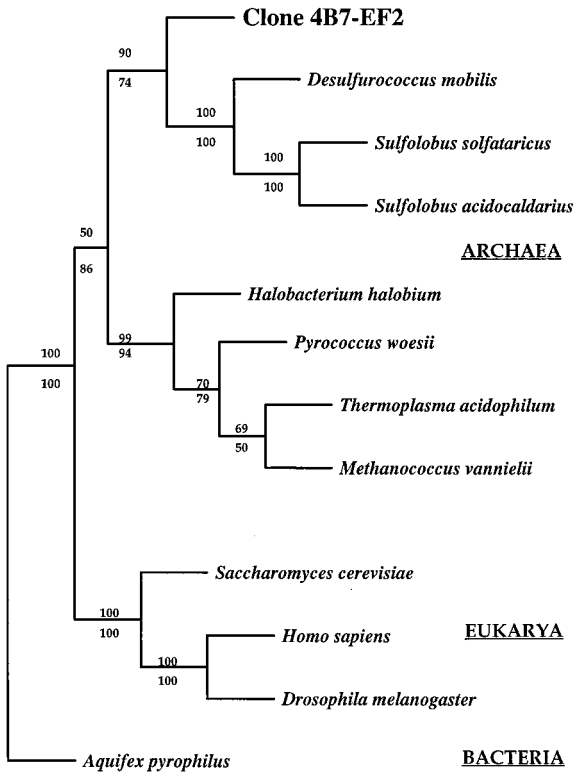

FIG. 8. Phylogenetic analyses of the inferred amino acid sequence of the EF2 gene contained on fosmid 4B7. The distance and parsimony analyses used identical alignments of 557 amino acid positions, and a total of 500 bootstrap replicates were performed. Values appearing above the branch correspond to bootstrap confidence levels of distance analyses (Protdist and Fitch programs, Phylip, version 3.2), and values appearing below branches correspond to bootstrap confidence levels of parsimony analyses (Protpars).

as their phylogenetic affiliation. The accelerated accumulation of prokaryotic genome sequence, structure, and organizational data provides a large comparative database for interrelating homologous genomic properties of as yet uncultivated microorganisms. Isolation, cultivation, and physiological characterization of prokaryotes by classical pure culture techniques obviously has distinct advantages and should be aggressively pursued. However, it may well be the case that some microorganisms will continue to elude even the best attempts to grow them at unnaturally high cell densities in artificial media in the monocultural state that we refer to as pure culture. As one alternative to the problem of cultivability, genomic walking from a phylogenetic anchor such as 16S rRNA may help provide a better understanding of the physiological and genetic potential of autochthonous, uncultivated microorganisms which can often represent major components of naturally occurring prokaryotic assemblages (7, 8, 16). In addition, the data derived from such genomic studies may suggest more useful strategies for optimizing future attempts at cultivation and physiological characterization of uncultivated microbes, using more traditional microbiological approaches.

ACKNOWLEDGMENTS

We are indebted to Tracy Villareal and the officers and crew of the RV Wecoma for sampling opportunities at sea and to Melvin Simon and Carl Woese for helpful advice and encouragement. We acknowledge Paul Fowler for helpful technical assistance.

This work was supported by NSF grants OCE-9317734 (J.L.S.) and OCE-9218523 (E.F.D.), an Alexander Hollaender Distinguished Postdoctoral Fellowship to J.L.S., a DuPont Young Professor Award to E.F.D., and a UCSB Faculty Career Development Award to E.F.D. T.L.M. was supported by a grant to Carl Woese, NASA NAGW2554.

REFERENCES

1. **Altschul, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman.** 1990. Basic local alignment search tool. J. Mol. Biol. **215:**403–410.
2. **Amils, R., P. Cammarano, and P. Londei.** 1993. Translation in the Archaea. New Comp. Biochem. **26:**393–438.
2a. **Barns, S., and N. Pace.** Personal communication.
3. **Barns, S. M., R. E. Fundyga, M. W. Jeffries, and N. R. Pace.** 1994. Remarkable archaeal diversity detected in a Yellowstone National Park hot spring environment. Proc. Natl. Acad. Sci. USA **91:**1609–1613.
4. **Britschgi, T. B., and S. J. Giovannoni.** 1991. Phylogenetic analysis of a natural marine bacterioplankton population by rRNA gene cloning and sequencing. Appl. Environ. Microbiol. **57:**1707–1713.
5. **Ceccarelli, E., M. Bochetta, R. Creti, A. M. Sanangelantoni, O. Tibinoni, and P. Cammarano.** 1995. Chromosomal organization and nucleotide sequence of the genes for elongation factors EF-α and EF-2 and ribosomal proteins S7 and S10 of the hyperthermophilic archaeum Desulfurococcus mobilis. Mol. Gen. Genet. **246:**687–696.
6. **Creti, R., E. Ceddarelli, M. Bocchetta, A. M. Sanangelantoni, O. Tiboni, P. Palm, and P. Cammarano.** 1994. Evolution of translational elongation factor (EF) sequences: reliability of global phylogenies inferred from EF-1α(TU) and EF-2(G) proteins. Proc. Natl. Acad. Sci. USA **91:**3255–3259.
7. **DeLong, E. F.** 1992. Archaea in coastal marine environments. Proc. Natl. Acad. Sci. USA **89:**5685–5689.
8. **DeLong, E. F., K. Y. Wu, B. B. Prezelin, and R. V. M. Jovine.** 1994. High abundance of Archaea in Antarctic marine picoplankton. Nature (London) **371:**695–697.
9. **DeSoete, G. A.** 1983. A least squares distance algorithm for fitting additive trees to proximity data. Psychometrika **48:**621–626.
10. **Felsenstein, J.** 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. **17:**368–376.
11. **Felsenstein, J.** 1985. Confidence limits on phylogenies: an approach using the bootstrap. Evolution **39:**783–791.
12. **Felsenstein, J.** 1989. Phylip-phylogeny inference package. Cladistics **5:**164–166.
13. **Fuhrman, J. A., K. McCallum, and A. A. Davis.** 1992. Novel major archaebacterial group from marine plankton. Nature (London) **356:**148–149.
14. **Fuhrman, J. A., K. McCallum, and A. A. Davis.** 1993. Phylogenetic diversity of subsurface marine microbial communitites from the Atlantic and Pacific Oceans. Appl. Environ. Microbiol. **59:**1294–1302.
15. **Fuller-Pace, F. V., S. M. Nicol, A. D. Reid, and D. P. Lane.** 1993. DbpA: a DEAD box protein specifically activated by 23S rRNA. EMBO J. **12:**3619–3626.
16. **Giovannoni, S. J., T. B. Britschgi, C. L. Moyer, and K. G. Field.** 1990. Genetic diversity of Sargasso Sea bacterioplankton. Nature (London) **345:** 60–63.
17. **Gish, W., and D. J. States.** 1993. Identification of protein regions by database similarity search. Nat. Genet. **3:**266–272.
18. **Hansson, M., L. Rutberg, I. Schroder, and L. Hederstedt.** 1991. The *Bacillus subtilis hemAXCDBL* gene cluster, which encodes enzymes of the biosynthetic pathway from glutamate to uroporphyrinogen III. J. Bacteriol. **173:** 2590–2599.
19. **Iwabe, N., K. Kuma, M. Hasegawa, S. Osawa, and T. Miyata.** 1989. Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. Proc. Natl. Acad. Sci. USA **86:**9355–9359.
20. **Kim, U.-J., H. Shizuya, P. J. de Jong, B. Birren, and M. Simon.** 1992. Stable propagation of cosmid sized human DNA inserts in an F factor based vector. Nucleic Acids Res. **20:**1083–1085.
21. **Lane, D. J.** 1991. 16S/23S sequencing, p. 115–176. *In* E. Stackebrandt and M. Goodfellow (ed.), Nucleic acid techniques in bacterial systematics. John Wiley & Sons, New York.
22. **Liesak, W., and E. Stackebrandt.** 1992. Occurrence of novel groups of the domain *Bacteria* as revealed by analysis of genetic material isolated from an Australian terrestrial environment. J. Bacteriol. **174:**5072–5078.
23. **Maidak, B. L., N. Larsen, M. J. McCaughey, R. Overbeek, G. J. Olsen, K. Fogel, J. Blandy, and C. R. Woese.** 1994. The ribosomal database project. Nucleic Acids Res. **22:**3485–3487.
23a. **Marsh, T. L., et al.** Unpublished data.
24. **McInerney, J. O., M. Wilkinson, J. W. Patching, T. M. Embley, and R. Powell.** 1995. Recovery and phylogenetic analysis of novel archaeal rRNA sequences from a deep-sea deposit feeder. Appl. Environ. Microbiol. **61:** 1646–1648.
25. **Olsen, G. J., H. Matsuda, R. Hagstrom, and R. Overbeek.** 1994. fastDNAml: a tool for construction of phylogenetic trees of DNA sequences using maximum likelihood. Comput. Appl. Biosci. **10:**41–48.
26. **Ouzounis, C. A., and B. J. Blencowe.** 1991. Bacterial DNA replication initiation factor *pri*A is related to proteins belonging to the 'Dead-box' family. Nucleic Acids Res. **19:**6953.
27. **Pace, N. R., D. A. Stahl, D. J. Lane, and G. J. Olsen.** 1986. The analysis of

natural microbial populations by ribosomal RNA sequences. Adv. Microb. Ecol. **9:**1–55.
28. **Raskin, L., J. M. Stomley, B. E. Rittman, and D. A. Stahl.** 1994. Group-specific 16S rRNA hybridization probes to describe natural communities of methanogens. Appl. Environ. Microbiol. **60:**1232–1240.
29. **Schmidt, T. M., E. F. DeLong, and N. R. Pace.** 1991. Analysis of a marine picoplankton community by 16S rRNA gene cloning and sequencing. J. Bacteriol. **173:**4371–4378.
30. **Shizuya, H., B. Birren, U.-J. Kim, V. Mancino, T. Slepak, Y. Tachiri, and M. Simon.** 1992. Cloning and stable maintenance of 300-kilobase pair fragments of human DNA in Escherichia coli using an F-factor-based vector. Proc. Natl. Acad. Sci. USA **89:**8794–8797.
31. **Stahl, D. A., and R. I. Amman.** 1991. Development and application of nucleic acid probes, p. 205–248. *In* E. Stackebrandt and M. Goodfellow (ed.), Nucleic acid techniques in bacterial systematics. John Wiley & Sons, New York.
32. **Stahl, D. A., D. J. Lane, G. J. Olsen, and N. R. Pace.** 1984. Analysis of hydrothermal vent associated symbionts by ribosomal RNA sequences. Science **224:**409–411.
33. **Thomm, M., and W. Hausner.** 1993. Genes for stable RNAs and their expression in *Archaea*, p. 36–53. *In* M. Sebald (ed.), Genetics and molecular biology of anaerobic bacteria. Springer-Verlag, New York.
33a. **Ueda, T., Y. Suga, and T. Matsuguchi.** 1995. Molecular phylogenetic analysis of a soil microbial community in a soybean field. Eur. J. Soil Sci. **46:**415–421.
34. **Weisburg, W. G., J. G. Tully, D. L. Rose, J. P. Petzel, H. Oyaizu, D. Yang, L. Mandelco, J. Sechrest, T. G. Lawrence, J. Van Etten, J. Maniloff, and C. R. Woese.** 1989. A phylogenetic analysis of the mycoplasmas: basis for their classification. J. Bacteriol. **171:**6455–6467.
35. **Woese, C. R.** 1987. Bacterial evolution. Microbiol. Rev. **51:**221–271.
35a. **Woese, C. R.** Personal communication.
36. **Woese, C. R., L. Achenbach, P. Rouviere, and L. Mandelco.** 1991. Archaeal phylogeny: reexamination of the phylogenetic position of Archaeoglobus fulgidus in light of certain composition induced artifacts. Syst. Appl. Microbiol. **14:**364–371.
37. **Woese, C. R., O. Kandler, and M. L. Wheelis.** 1990. Towards a system of organisms; proposal for the domains Archaea, Bacteria, and Eucarya. Proc. Natl. Acad. Sci. USA **87:**4576–4579.
38. **Zharkikh, A., and W. H. Li.** 1992. Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. Mol. Biol. Evol. **9:**1119–1147.