Nucleic Acids Research, 1996, Vol. 24, No. 22 **4373–4378**

Genomic and cDNA sequence tags of the hyperthermophilic Archaeon *Pyrobaculum aerophilum*

Paul Völkl[1,2], Peter Markiewicz[1], Claudia Baikalov[1], Sorel Fitz-Gibbon[1], Karl O. Stetter[2] and Jeffrey H. Miller[1,*]

[1]Department of Microbiology and Molecular Genetics and the Molecular Biology Institute, University of California, 405 Hilgard Avenue, Los Angeles, CA 90024, USA and [2]Archaeenzentrum, Universität Regensburg, 93053 Regensburg, Germany

Received October 7, 1996; Accepted October 8, 1996

ABSTRACT

The hyperthermophilic archaeum, *Pyrobaculum aerophilum*, grows optimally at 100°C with a doubling time of 180 min. It is a member of the phylogenetically ancient *Thermoproteales* order, but differs significantly from all other members by its facultatively aerobic metabolism. Due to its simple cultivation requirements and its nearly 100% plating efficiency, it was chosen as a model organism for studying the genome organization of hyperthermophilic ancient archaea. By a G+C content of the DNA of 52 mol%, sequence analysis was easily possible. At least some of the mRNA of *P.aerophilum* carried poly-A tails facilitating the construction of a cDNA library. 245 sequence tags of a poly-A primed cDNA library and 55 sequence tags from a 1–2 kb *Sau*3AI-fragment containing genomic library were analyzed and the corresponding amino acid sequences compared with protein sequences from databases. Fourteen percent of the cDNA and >9% of genomic DNA sequence tags revealed significant similarities to proteins in the databases. Matches were obtained to proteins from archaeal, bacterial and eukaryal sources. Some sequences showed greatest similarity to eukaryal rather than to bacterial versions of proteins, other matches were found to proteins which had previously only been found in eukaryotes.

INTRODUCTION

Hyperthermophilic archaea represent an interesting source for studying phylogeny and organization of ancestral life. Based on 16S rRNA analysis, the most ancient living organisms known are hyperthermophiles (1,2). Therefore, their study may provide some clues as to the genome organization of the common ancestor of *Bacteria*, *Archaea* and *Eukarya* as well as to the basic requirements for thermophily. However, due to the difficult cultivation, especially of the deep branching and strictly anaerobic hyperthermophiles, these organisms are only poorly investigated. Most genetic studies with archaea were therefore done with the mesophilic extreme Halophiles (3,4) and with aerobic representatives of *Sulfolobus* (5). Both represent the longest evolutionary lineages of the *Euryarchaeota* and of the exclusively hyperthermophilic organisms containing *Crenarchaeota* kingdom, respectively. But, for a better understanding of the basic characteristics of phylogenetically ancient organisms it would be helpful to look at more slowly evolving archaea too. Recently, within the so far strictly anaerobic *Thermoproteales* order, which is comprised of slowly evolving organisms within the *Crenarchaeota*, having short evolutionary lineages based on 16S rRNA phylogenetic analysis (6), a novel isolate was described which could serve as a model organism for the molecular investigation of hyperthermophilic archaea. This new isolate, *Pyrobaculum aerophilum* (type strain: IM2), grows optimally at 100°C and pH 7.0 and represents the only facultatively aerobic member of this order (7). In contrast to other hyperthermophiles, *P.aerophilum* is therefore easy to grow either aerobically or anaerobically and can be plated to form colonies within four days with up to 100% efficiency (7). With these features *P.aerophilum* is an ideal candidate for molecular investigation and genetic studies. As a first step, sequence data from a large number of this organism's genes will provide invaluable comparative information and elucidate the early evolution of many biological systems. In contrast to members of *Pyrodictium* and some other hyperthermophilic archaea (1) where due to their extremely high G+C content DNA sequencing is hampered, *P.aerophilum* has a G+C content of 52 mol% which is suitable for sequencing. The genome sizes of hyperthermophiles are universally small, ~2 Mb, which is less than half the size of an *E.coli* genome. Thus random sequencing rapidly identifies a large proportion of the genes in the organism. The presence of poly-A mRNA in an organism makes it possible to reverse transcribe those genes back into DNA, thus making them available for cloning and sequencing. Polyadenylation of mRNA was believed to be a eukaryotic feature until short poly-A tails were also found at the 3′-terminal ends of bacterial mRNAs (8). In archaea, polyadenylation of mRNA has so far only been described in *Methanococcus vannielii*, a mesophilic archaeum within the *Euryarchaeota* kingdom (9). Based on this background, we tried to isolate poly-A mRNA from

*To whom correspondence should be addressed. Tel: +1 310 825 8460; Fax: +1 310 206 3088; Email: jhmiller@ewald.mbi.ucla.edu

the crenarchaeal member *P.aerophilum* and constructed genomic and cDNA libraries. Out of these libraries we generated sequence tags from randomly chosen clones and categorized them as potential homologs to known protein sequences. This sequence tag method has been successfully applied for studying human brain specific transcripts (10) and for the investigation of the genome of the eubacterial extreme thermophile *Thermotoga maritima* (11). The sequence tags may further be used for creating a physical map of the genome, phylogenetic analysis, and 'reverse genetics' (10). In the following study, DNA sequence tags were translated into protein sequences and compared with those available in the GenBank, EMBL and PIR databases. With this method we were able to identify numerous genes, and at the same time we generated sequence data usable for further studies such as full-length gene sequencing. Additionally, the results of this study allow us to draw conclusions about the relationship of single proteins of *P.aerophilum* with those of phylogenetically diverse organisms.

MATERIALS AND METHODS

Strains and culture conditions

P.aerophilum was grown aerobically in BS medium at 97°C as described previously (7). Cells were harvested in the late exponential growth phase by centrifugation and the cell masses were stored at –80°C until use. *Escherichia coli* strains used were XL1-Blue (*recA1, endA1, gyrA96, thi-1, hsdR17, supE44, relA1, lac*, [F′*proAB lacIq* ZΔ*M15*, Tn*10(tetr)*]) and SURE (*mcrA*, Δ(*mcrBC-hsdRMS-mrr)171*, *supE44*, *thi-1*, l-, *gyrA96*, *relA*, lac, *recB*, *recJ*, *sbcC*, *umuC*::Tn*5* (*kanr*), *uvrC*, [F′*proAB lacIqZDM15*, Tn*10(tetr)*]) (Stratagene, La Jolla, CA).

Escherichia coli strains were grown on Luria Bertani medium (LB) prepared as described by Miller (12), LB was supplemented with 0.2% maltose and 10 mM magnesium sulfate when strains were grown before and during Lambda phage infection. NCY broth contained 5 g/l NaCl, 2 g/l $MgSO_4.7H_2O$, 5 g/l yeast extract (Difco) and 10 g/l NZ amine (casein hydrolysate; Sigma). The pH was adjusted to 7.5 with NaOH. Agar plates contained 1.5% agar (Difco). Top agarose contained 0.7% agarose (BioRad) instead of agar. SM buffer contained 5.8 g NaCl, 2.0 g $MgSO_4.7H_2O$, 50 ml 1 M Tris–HCl, pH 7.5, and 5 ml 2% gelatin per liter. Ampicillin was added to 100 μg/ml and tetracycline to 20 μg/ml medium as described by Miller (12). Tetracycline was not added to media supplemented with maltose and magnesium sulfate.

DNA isolation

DNA from *P.aerophilum* was purified by a method described by Sharp and Williams (13) and modified by Kim *et al.* (11). Approximately 0.7 g of frozen cell paste was suspended in 10 ml dH_2O. Protease K was added to a final concentration of 1.5 mg/ml and incubated at 37°C. After 1 h, 1.2 ml of a 4% SDS solution was added and the sample was incubated at 55°C for 1 h. The lysate was chilled on ice, and extracted with 12 ml phenol–chloroform–isoamyl alcohol (24:24:1 by volume). The phases were separated by spinning at 17 000 *g* for 10 min at room temperature. The upper aqueous phase was extracted three additional times then the DNA was precipitated by adding 0.1 vol 3 M sodium acetate and 2.5 vol absolute ethanol. After incubation for 10 min at room temperature, the precipitated DNA was spooled out with a closed end Pasteur pipette, rinsed in 70% ethanol, transferred into an Eppendorf tube, and dried under vacuum for 5 min. The dry pellet was resuspended in 5 ml TE buffer (pH 8.0). RNA was digested with RNaseA at a final concentration of 50 μg/ml at 37°C. After 1 h, 1.5 mg/ml Protease K was added. After incubation at 37°C for 1 h, 0.1 vol 2% deoxycholate was added and incubation was continued for another hour. The solution was extracted four times with 1 vol phenol–chloroform–isoamyl alcohol, ethanol precipitated, dried and resuspended in 1 ml TE buffer pH 8.0. The final yield was 0.67 mg DNA, as calculated by OD_{260} measurement.

Plasmid–DNA isolation

Plasmid DNA was isolated by the alkaline lysis procedure as described by Kraft *et al.* (14). Prior to sequencing the plasmid DNA was analyzed by restriction mapping and agarose gel analysis as described by Sambrook *et al.* (15).

RNA isolation

RNA was isolated from a frozen *P.aerophilum* cell paste from a culture grown aerobically in a 300 l fermentor. The Stratagene (La Jolla, CA) single step RNA isolation Kit using a guanidinium thiocyanate protocol (16) was used. Crude RNA was further purified by CsCl gradient ultra centrifugation. The RNA, in a volume of 3.7 ml denaturation solution D (4 M guanidinium thiocyanate, 25 mM sodium citrate, pH 7.0, 0.5% sarcosyl, 0.1 M 2-mercaptoethanol) was carefully layered over 8.3 ml 5.7 M CsCl, 10 mM EDTA in a 15 mm × 102 mm polyallomer ultracentrifuge tube and pelleted by centrifugation in a SW28 rotor at 25 000 r.p.m. for 24 h at 20°C. The supernatant was removed and the RNA pellet washed with 1 ml 75% ethanol, air dried and resuspended in 1 ml 10 mM Tris–HCl, pH 7.5, 0.5 M KCl (oligo-dT binding buffer). The recovery was ~25% of the crude RNA.

Purification of poly-A RNA

Polyadenylated RNA was separated from CsCl-purified RNA by binding on oligo-dT cellulose (New England Biolabs). Oligo-dT cellulose (100 mg) was washed in an Eppendorf tube with 1 ml 0.1 N NaOH, quickly spun down, and the supernatant removed. The oligo-dT cellulose was neutralized by washing in 1 ml elution buffer (10 mM Tris–HCl, pH 7.5) until the pH was 7.5 (5 times). Finally the cellulose was equilibrated with 1 ml binding buffer (10 mM Tris–HCl, pH 7.5, 0.5 M KCl) for 1 h at 4°C. The CsCl purified RNA in 1 ml binding buffer was mixed with the oligo-dT cellulose and slowly swirled at 4°C for 2 h. After brief centrifugation at 4°C, the supernatant was removed and the oligo-dT cellulose washed 5 times with 1 ml binding buffer. Poly-A RNA was eluted by adding 1 ml elution buffer and incubating at 45°C for 15 min. The elution was repeated 2 times. The concentration of RNA in all wash and elution steps was followed by ethidium bromide spot tests. The three elution samples were combined and ethanol precipitated by adding 0.1 vol 3 M sodium acetate and 2.8 vol absolute ethanol. The samples were incubated at –70°C for 12 h and centrifuged at 25 000 r.p.m. in a SW28 rotor at 4°C for 2 h. The poly-A RNA pellet was washed with 75% ethanol, dried and dissolved in 50 μl dH_2O.

Genomic library construction

Partial digestions of *P.aerophilum* DNA were made with different dilutions of *Sau*3AI endonuclease incubated exactly 1 h at 37°C

(15). The restriction fragments were analyzed on a 1% TAE agarose gel. A digestion of 10 µg DNA with 0.2 U *Sau*3AI for 1 h at 37°C yielded DNA fragments ranging in length from 500–2500 base pairs. A digestion with 0.04–0.08 U *Sau*3AI for 1 h at 37°C gave fragments 3–15 kb in size. From these two sets of restriction digests, fragments of 1–2 kb and 3–10 kb in size were excised from a preparative 1% TAE agarose gel, and the DNA recovered using the GeneClean DNA Isolation kit (BIO101, La Jolla, CA). The fragments were ligated into pBluescript II (SK–) vector (Stratagene) digested with *Bam*HI and dephosphorylated with calf intestine alkaline phosphatase (0.25 U/10 µg vector DNA, 30 min, 37°C). Ligation products were introduced into *E.coli* strain 'SURE' by electroporation using a BioRad electroporator (BioRad, Richmond, CA). Transformants were selected by plating on LB agar plates supplemented with ampicillin (100 µg/ml), X-gal (40 µg/ml) and IPTG (0.1 mM). Single colonies were grown in 3 ml liquid LB/amp medium, and the cell masses used for plasmid isolation.

cDNA library construction

cDNA libraries were constructed using the Uni-ZAP XR cDNA Cloning kit (Stratagene, La Jolla, CA). The protocol was followed as described by the manufacturer. The construction of a poly-T primed library was performed with 5 µg of purified poly-A RNA using a 50 base oligonucleotide containing an 18 base oligo-dT 3′-end, a 'GAGA' 5′-end and an internal *Xho*I site as primer for reverse transcription. For creating a randomly primed library, 20 µg of crude RNA was used with a 46 base random primer containing an internal *Xho*I site (Stratagene).

Sequencing and data analysis

Clones from the oligo-dT primed library and genomic libraries were sequenced by the Sanger chain termination method (17) using a Sequenase Version 2.0 kit (US Biochemical) and [α-^{32}P]dATP (NEN). Sequencing products were separated on 6% polyacrylamide–urea gels at two intervals to obtain overlapping sequencing runs. Genomic clones were sequenced from both directions with the SK and KS primers (Stratagene). For sequencing oligo-dT primed clones SK and M13 –20 primers (Stratagene) were used.

Sequences were analyzed for similarity to known proteins with the NCBI BLAST program (18), using the NCBI non-redundant database containing GenBank, PIR, SwissProt and EMBL. Database sequences were retrieved using BLAST retrieve or Internet Gopher (19).

RESULTS

cDNA libraries

To obtain sequence information of expressed genes in *P.aerophilum* two cDNA libraries were constructed. For the first library, polyadenylated RNA was isolated from crude RNA by binding to oligo-dT cellulose. About 1% of the crude RNA had bound and could be recovered from the oligo-dT cellulose indicating the presence of abundant poly-A RNA in the cell. For the second library, crude RNA was used as template for randomly primed reverse transcription. Both cDNA libraries were cloned unidirectionally into Lambda ZAP II (Stratagene). The number of clones obtained from the poly-T primed library was 3.2×10^6, while the randomly primed library yielded only 4.5×10^4 clones. The analysis of the first 18 poly-A cDNA clones by *Not*I/*Apa*I restriction enzyme digestion revealed an average insert size of 1 kb, ranging from 0.6–2.0 kb. In contrast, the randomly primed cDNA clones had very short or no inserts and were therefore not further used for random sequencing analysis.

On average, from a single clone a sequence tag of ~310 nucleotides in length could be determined by one sequence reaction. Preliminary analysis of 245 oligo dT-primed cDNA sequence tags (abbreviated as ESPAT: expressed sequence of *P.aerophilum* from oligo-dT primed library), by comparison of the translated DNA sequence with known protein sequences from PIR, SwissProt and GenBank databases resulted in strong matches to 34 different proteins from other organisms, which is a match rate of ~14%. Matches were found to proteins of various functions including several intermediary metabolism proteins, DNA/RNA related proteins, and surface/transport proteins (Table 1). Other non-metabolic matches included the eukaryotic elongation factor ef-2, human DNA polymerase replication factor C and one clone (ESPAT-71) that weakly matched a wide variety of DNA-binding proteins. In this case the closest matches were with eukaryotic proteins (not shown). Sequence tag ESPAT-136 had strong similarity to archaeal and eukaryotic translation elongation factor ef-2 showing the histidine which, if post-translationally modified, would be the target of diphtheria toxin. The remaining clones had open reading frames whose amino acid sequence did not correspond to any known protein sequence in the databases. However, these sequences were not ribosomal DNA, and frequently showed short matches to protein 'motifs', such as nucleotide binding folds or cysteine patterns similar to those found in iron–sulfur cluster containing proteins.

Genomic libraries

In order to study DNA sequences from transcribed and non-transcribed regions of the *P.aerophilum* genome a 1–2 kb *Sau*3AI genomic library was constructed. The average insert size of the genomic clones was 1.5 kb. With an estimated *P.aerophilum* genome size of 2×10^6 base pairs, a given gene would be represented in a non-biased library of this size (>4000 clones) with >95% probability. From a total of 30 different clones 55 sequence tags, with an average length of 320 nucleotides, were analyzed by comparing translated sequences with amino acid sequences in databases using the program BLAST (18). Five sequence tags exhibited notable similarity to known protein sequences in the databases (Table 1). This is a match rate of ~9%, in contrast to the 14% match rate of the cDNA sequence tags. As in the oligo-dT primed library, some clones exhibited greater similarity to eukaryotic rather than to bacterial proteins.

In order to verify some of the putative matches, sequences up- and downstream of the matching region of the nitrate reductase and the ribokinase were generated and analyzed. For all sequences the similarities continued in these elongated sequence stretches (not shown). Similarly, starting with the similarity found by sequence tag GSPA-35, the entire gene encoding a subtilisin like serine protease was cloned and the DNA sequence determined together with flanking regions (20). The genomic DNA sequences were not interrupted by introns as shown by comparison with the corresponding cDNA sequences.

Table 1. Summary of cDNA and genomic sequence tags of *P.aerophilum* exhibiting similarities to known protein sequences

Clone[a]	Sequence[b]	Matching protein (organism)[c]	Identity[d]	Accession[e]
1. Homologies related to DNA/RNA-metabolism				
ESPAT-99 SK	410	*DNA pol. replication factor C (Human)	37/59 (48%)	SP:P35249
ESPAT-136 SK	345	(*)Elongation factor ef-2 (*Sulfolobus*)	69/117 (59%)	SP:P23112
ESPAT-145 SK	650	*Rec*F protein (*S.thyphimurium*)	25/40 (62%)	SP:P24900
ESPAT-217 SK	400	DNA-ligase (*Desulfurolobus*)	21/53 (40%)	SP:Q02093
ESPAT-217 M13	400	Ribokinase (*E.coli*)	38/92 (41%)	SP:P05054
ESPAT-258 SK	337	50S ribosomal protein (*Mc. vannielii*)	16/30 (53%)	SP:P15824
2. Homologies to proteins with metabolic function				
ESPAT-33 SK	226	Aconitate hydratase (*Bacillus*)	19/33 (57%)	SP:P09339
ESPAT-80 SK	349	*Carbamoyl phosphatase (Yeast)	13/18 (47%)	SP:P07259
ESPAT-91 SK	320	Salty Thiosulfate Reductase [*Salmonella*]	14/27 (25%)	SP:P37600
ESPAT-224 M13	400	Polysulfide reductase B(*Wollinella*)	31/50 (49%)	SP:P31076
ESPAT-207 SK	324	Molybdopterin biosynth. (*Moe*A) (*E.coli*)	15/25 (60%)	SP:P12281
ESPAT-215 SK	290	Formate dehydrogenase (*E.coli*)	21/60 (35%)	SP:P24183
ESPAT-224 SK	322	Dimethy sulfoxide reductase (*E.coli*)	38/75 (50%)	SP:P18776
ESPAT-238 SK	246	Disulphide oxidoreductase [*Entamoeba*]	12/22 (43%)	gi|895831
ESPAT-249 SK	273	Dihydroxyacid dehydratase [*Clostridium*]	16/28 (43%)	SP:P31959
ESPAT-257 SK	302	Resp. nitrate reductase A (*E.coli*)	47/72 (65%)	SP:P09152
ESPAT-310 SK	318	*NADH plastochin. oxidored. (*Marchantia*)	17/51 (33%)	SP:P12131
ESPAT-317 M13	306	Glutamate dehydrogenase (*Clostridium*)	42/82 (51%)	SP:P27346
ESPAT-330 SK	379	Glucose-1 dehydrogenase A (*Bacillus*)	25/70 (35%)	SP:P10528
		7 alpha-hydroxysteroid hydrogenase (Eubac.)	24/54 (44%)	PIR:A42468
GSPA-35 SK	463	Subtilisin (*Bacillus*)	33/66 (50%)	SP:P00780
GSPA-70 SK	320	*Homoaconitase (*Yeast*)	32/72 (37%)	SP:P49367
GSPA-81 KS	301	GMP synthase (*Bacillus subtilis*)	11/28 (39%)	GP:S88687_1
3. Homologies to surface/transport proteins				
ESPAT-53 SK	298	*Erythrocyte membrane prot. 7 (Human)	45/99 (46%)	SP:P27105
ESPAT-111 SK	357	*drr*A: Daunorubicin res. prot. (*S.peucetius*)	45/94 (40%)	SP:P32010
ESPAT-260 M13	300	ABC Transporter (*Haemophilus*)	16/33 (22%)	SP:P44531
ESPAT-299 M13	291	*Cu 2+-transporting ATPase, P-type (Human)	30/66 (15%)	PIR:JC2465
ESPAT-314 SK	292	Daunorubicin res. ATP bind. prot. (*Strep*)	16/32 (50%)	SP:P32010
ESPAT-322 M13	300	Inner membran protein lack (*Agrobacterium*)	29/49 (59%)	SP:Q01937
GSPA-15 KS	293	ATP-dependent transporter [*Cyanophora*]	25/41 (23%)	SP:P48255
4. Homologies to proteins of miscellaneous function				
ESPAT-235 M13	274	Cation-transporting ATPase [*Synechococcus*]	17/28 (54%)	SP:P37279
ESPAT-246 SK	264	*Adenosylhomocysteinase (*C.elegans*)	32/61 (52%)	SP:P27604
ESPAT-309 M13	263	HisF protein (*Azospirillum*)	25/41 (60%)	SP:P26721
ESPAT-326 M13	308	*Activator 1 37 KD subunit (Human)	22/41 (53%)	PIR:A45253
GSPA-15 MD1	387	Ribonuclease E (*E.coli*)	14/35 (40%)	PIR:JG0009

[a]ESPAT, expressed sequence from *Pyrobaculum aerophilum* oligo-dT primed cDNA library followed by number of clone and primer used for sequencing. GSPA, genomic sequence of *P.aerophilum*.
[b]Number of bases sequenced.
[c]Most homologous protein. Eukaryotic matches are indicated by an asterisk (*).
[d]Numbers are: identical amino acids/amino acids in the most similar area. In parenthesis: percentages of identical amino acids within the total sequence tag.
[e]Accession ID of the matching protein sequence: SP, swissprot; GP, genpept; PIR, PIR databases.

Sequence similarities of *Thermotoga maritima*

In a similar sequencing project using the hyperthermophilic bacterium *Thermotoga maritima*, 32% of 244 cDNA and 14% of 21 genomic DNA sequence tags revealed similarities to known protein sequences (11). With two exceptions (AVP-3 vacuolar H^+-phosphatase from *Arabidopsis* and aspartate aminoacyl tRNA synthase from *Saccharomyces*), strong matches were exclusively to eubacterial proteins. The basic results of the two sequence tag approaches from *T.maritima* and *P.aerophilum* are summarized in Table 2.

Table 2. Comparison of the results of the random sequence tag approaches from the bacterium *T.maritima* and the archaeum *P.aerophilum*

Subject	*Thermotoga maritima*	*Pyrobaculum aerophilum*
Total of sequence tags analyzed	265	300
expressed sequences	244	245
genomic sequences	21	55
Match rate of expressed sequence tags	32%	14%
Match rate of genomic sequence tags	14%	9%
Matches specific to eukaryotic proteins	2/52	9/34

DISCUSSION

A critical requirement for genetic analysis is that single cells of an organism form visible colonies on defined media with high efficiency and in a reasonable time. Unfortunately, the deep branching and therefore phylogenetically most interesting archaea have extremely low plating efficiencies. Additionally, they are strictly anaerobic and due to their high growth temperatures and other features difficult to grow and to handle. With the new isolate *P.aerophilum* a suitable candidate for genetic studies of hyperthermophilic primitive archaea is available for the first time. By constructing an oligo-dT primed cDNA library *P.aerophilum* has been shown to possess polyadenylated mRNA. Each of the 245 clones sequenced to date was unique and also did not contain ribosomal DNA, indicating that polyadenylation is widespread among different mRNA species of *P.aerophilum* and is therefore useful for the construction of a highly representative cDNA library. With the sequence tag strategy a large amount of sequence information was accumulated, creating a broad basis for further studies. Using cloned sequences, it may be possible to introduce defined mutations into the genome of this organism, and thereby analyze the function of the gene by 'reverse genetics.' Furthermore, the identified genes could be used for creating a map of the *P.aerophilum* genome.

In the poly-A cDNA library, the match rate to known proteins was 14%, while the rate in the genomic libraries was lower (9%), which is expected due to the additional regulatory and non-coding DNA. Despite their lower match rate, the genomic clones are useful when analyzing non-coding regions of the genome. Introns in archaeal protein genes are unknown, although they have been shown to occur within some ribosomal genes like the 16S rRNA gene of *P.aerophilum* (21). If they occur at all they can't be very frequent because of the small genome sizes. The genomic sequence encoding a serine type protease in *P.aerophilum* did not contain an intervening sequence when compared with the corresponding cDNA sequence (20). Furthermore, the genome seems to be organized in polycistronic transcription units as in the intron-scarce eubacteria, as was previously seen for Archaeal Methanogen genes (22).

The cDNA match rate for *P.aerophilum* was lower than for the hyperthermophilic eubacterium *Thermotoga maritima*, which also employed a sequence tag approach (11). In this study, >30% of cDNA sequence tags could be assigned as likely homologs of genes known from other organisms. Similar projects, such as sequencing human brain cDNAs (10) or cDNAs of the nematode *Caenorhabditis elegans* (23), yielded around 20% and up to 40% known sequences, respectively. It may be that *P.aerophilum* contains many genes which are unique to its lineage, which do not exist in eukaryotes or eubacteria. But, like in the other cases the real number of homologous sequences is probably higher because with a 100 amino acid long sequence tag an unconserved region of an otherwise homologous protein will not be detected. The clearest difference to the results from *T.maritima* was that several matches from *P.aerophilum* were to eukaryotic rather than to prokaryotic proteins, where with few exceptions, the sequence tags from *T.maritima* matched exclusively eubacterial proteins. The elongation factor ef-2 from *P.aerophilum* was clearly of the eukaryotic type. It exhibited highest similarity to *Sulfolobus* but, it still had 40% identity to the human elongation factor ef-2. This finding is absolutely in line with the phylogenetic positions of *P.aerophilum* and supports the three domain concept of life (6,24). In other cases amino acid sequence similarities were highest to eukaryotic proteins for which no prokaryotic equivalent is known, such as the erythrocyte membrane protein. The finding of both similarities to typically eukaryotic and similarities to bacterial proteins reflect the phylogenetic position of archaea sharing both eubacterial and eukaryotic features. The study of archaea will in many cases provide the missing link in our understanding of the relationship between homologous systems in diverse organisms (25).

ACKNOWLEDGEMENTS

We would like to thank Jean J. Lee for excellent technical assistance. This work was supported in part by grants of the Office of Naval Research (ONR) to J.H.M. (N00014-95-1-0938) and the Deutsche Forschungsgemeinschaft to K.O.S.

REFERENCES

1 Stetter,K.O. (1992) In Tran Than Van,J.K., Mounolou,J.C., Schneider,J. and McKay,C. (eds), *Colloque Interdisciplinaire du Comite National de la Recherche Scientifique, Frontiers of Life*, pp. 195–219.
2 Woese,C.R. (1987) *Microbiol. Rev.*, **51**, 221–271.
3 Pfeifer,F., Offner,S., Kruger,K., Ghahraman,P. *et. al.* (1994) *Systematic Appl. Microbiol.*, **16**, 569–577.
4 Tchelet,R. and Mevarech,M. (1994) *Systematic Appl. Microbiol.*, **16**, 578–581.
5 Schleper,C., Roder,R., Singer,T. and Zillig,W. (1994) *Mol. Gen. Genet.*, **243**, 91–96.
6 Woese,C.R., Kandler,O. and Wheelis,M.L. (1990) *Proc. Natl Acad. Sci. USA*, **87**, 4576–4579.
7 Völkl,P., Huber,R., Drobner,E., Rachel,R., Burggraf,S., Trincone,A. and Stetter,K.O. (1993) *Appl. Environ. Microbiol.*, **59**, 2918–2926.
8 Gopalakrishna,Y., Langley,D. and Sarkar,N. (1981) *Nucleic Acids Res.*, **9**, 3545–3554.
9 Brown,J.W. and Reeve,J.N. (1985) *J. Bacteriol.*, **162**, 909–917.

10 Adams,M.D., Kelley,J.M., Gocayne,J.D., Dubnick,M., Polymeropoulos,M.H., Xiao,H., Merril,C.R., Wu,A., Olde,B., Moreno,R.F. *et al.* (1991) *Science*, **252**, 1651–1656.
11 Kim,C.W., Markiewicz,P., Lee,J.J., Schierle,C.F. and Miller,J.H. (1993) *J. Mol. Biol.*, **231**, 960–981.
12 Miller,J.H. (1972) *Experiments in Molecular Genetics*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
13 Sharp,R.J. and Williams,R.A.D. (1988) *Appl. Environ. Microbiol.*, **54**, 2049–2053.
14 Kraft,R., Tardiff,J., Krauter,K.S. and Leinwand,L.A. (1988) *Biotechniques*, **6**, 544–546.
15 Sambrook,J., Fritsch,E.F. and Maniatis,T. (1989) *Molecular Cloning: A Laboratory Manual*, 2nd Edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
16 Chomczynski,P. and Sacchi,N. (1987) *Anal. Biochem.*, **162**, 156–159.
17 Sanger,F., Nicklen,S. and Coulson,A.R. (1992) *Biotechnology*, **24**, 104–108.
18 Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) *J. Mol. Biol.*, **215**, 403–410.
19 Krol,E. (1992) In Krol,E. (ed.), *The Whole Internet User's Guide and Catalog*. O'Reilly & Associates, Sebastopol, CA, pp. 189–210.
20 Völkl,P., Markiewicz,P., Stetter,K.O. and Miller,J.H. (1994) *Protein Sci.*, **3**, 1329–1340.
21 Burggraf,S., Larsen,N., Woese,C.R. and Stetter,K.O. (1993) *Proc. Natl Acad. Sci. USA*, **90**, 2547–2550.
22 Reeve,J.N., Hamilton,P.T., Beckler,G.S., Morris,C. J. and Clarke,C.H. (1986) *System. Appl. Microbiol.*, **7**, 5–12.
23 Bargmann,C.I. (1992) *Nature Genet.*, **1**, 79–80.
24 Cammarano,P., Palm,P., Creti,R., Ceccarelli,E., Sanangelantoni,A.M. and Tiboni,O. (1992) *J. Mol. Evol.*, **34**, 396–405.
25 Keeling,P.J. and Doolittle,W.F. (1995) *Proc. Natl Acad. Sci. USA*, **92**, 5761–5764.