Applied and Environmental Microbiology, Jan. 2006, p. 135–143 Vol. 72, No. 1
0099-2240/06/$08.00+0 doi:10.1128/AEM.72.1.135–143.2006

Identification of Eukaryotic Open Reading Frames in Metagenomic cDNA Libraries Made from Environmental Samples†

Susan Grant,[1] William D. Grant,[1] Don A. Cowan,[2] Brian E. Jones,[3] Yanhe Ma,[4] Antonio Ventosa,[5] and Shaun Heaphy[1]*

Department of Infection Immunity and Inflammation, University of Leicester, University Road, Leicester LE1 9HN, United Kingdom[1]; Department of Biotechnology, University of the Western Cape, Bellville 7535, Cape Town, South Africa[2]; Genencor International B.V., Archimedesweg 30, 2333 CN Leiden, The Netherlands[3]; State Key Laboratory of Microbial Resource, Institute of Microbiology, Chinese Academy of Sciences, 100080 Beijing, China[4]; and Department of Microbiology and Parasitology, Faculty of Pharmacy, University of Sevilla, 41012 Sevilla, Spain[5]

Received 11 August 2005/Accepted 30 September 2005

Here we describe the application of metagenomic technologies to construct cDNA libraries from RNA isolated from environmental samples. RNAlater (Ambion) was shown to stabilize RNA in environmental samples for periods of at least 3 months at −20°C. Protocols for library construction were established on total RNA extracted from *Acanthamoeba polyphaga* trophozoites. The methodology was then used on algal mats from geothermal hot springs in Tengchong county, Yunnan Province, People's Republic of China, and activated sludge from a sewage treatment plant in Leicestershire, United Kingdom. The Tenchong libraries were dominated by RNA from prokaryotes, reflecting the mainly prokaryote microbial composition. The majority of these clones resulted from rRNA; only a few appeared to be derived from mRNA. In contrast, many clones from the activated sludge library had significant similarity to eukaryote mRNA-encoded protein sequences. A library was also made using polyadenylated RNA isolated from total RNA from activated sludge; many more clones in this library were related to eukaryotic mRNA sequences and proteins. Open reading frames (ORFs) up to 378 amino acids in size could be identified. Some resembled known proteins over their full length, e.g., 36% match to cystatin, 49% match to ribosomal protein L32, 63% match to ribosomal protein S16, 70% to CPC2 protein. The methodology described here permits the polyadenylated transcriptome to be isolated from environmental samples with no knowledge of the identity of the microorganisms in the sample or the necessity to culture them. It has many uses, including the identification of novel eukaryotic ORFs encoding proteins and enzymes.

Information about prokaryotic diversity has expanded considerably as molecular methods have become available, particularly the routine direct amplification and cloning of 16S rRNA genes from environmental samples (3, 28). Subsequent sequencing of these environmental rRNA gene libraries has revealed new, often uncultivated, groups of the *Bacteria* and *Archaea*, some present in great abundance (5, 9, 10, 11). More-recent studies involving molecular analysis of small subunit 18S rRNA genes are revealing a similar wide diversity among unicellular eukaryotes (4, 17, 21).

In the context of gene discovery, access to the entire metagenome of these as yet uncultured organisms would source a completely new gene pool, which could provide novel enzymes and proteins of potential industrial or medical use. It is now common practice to isolate DNA directly from environmental samples and construct DNA libraries to access the metagenome (12). Cloning into expression vectors allows the isolation of novel enzymes and other biologically active proteins and peptides (the proteome) without prior cultivation of the organisms from which they are derived (8, 14, 18, 25). This approach is very suitable for samples containing prokaryotic DNA where, in general, genes contain few introns, allowing direct transcription and translation of the metagenome in a host organism such as *Escherichia coli*.

However, in the case of eukaryotic organisms, such a direct route from genome to proteome is generally not possible because most genes contain a number of large introns. Consequently, libraries for the screening and expression of proteins from eukaryotes are generally derived from the transcriptome, particularly mRNA, by reverse transcription to form double-stranded DNA and cloning into a suitable vector (26). Such libraries are now available commercially for a wide diversity of organisms, particularly for species that are important in agriculture, commerce, medicine, or the laboratory.

Libraries derived from environmental RNA have rarely been made. This is much more challenging, not least because of RNA instability. Methods have been described for isolating RNA from soils and sediments, most of which appear to have been derived from prokaryotes (16). Small ~150-bp PCR products corresponding to bacterial mRNA species could be identified in this material. Very recently, cDNA libraries have been made from environmental prokaryotic RNA (23). Clones up to 1 kb were identified, but most were in the 200- to 500-bp range. In another study, cDNA libraries were made from environmental RNA fractionated to contain prokaryotic 16S-sized rRNA; seven of these clones appeared to be mRNA related, including two fungal sequences (7).

* Corresponding author. Mailing address: Department of Infection Immunity and Inflammation, University of Leicester, University Road, Leicester LE1 9HN, United Kingdom. Phone: 44 116 252 2973. Fax: 44 116 252 5030. E-mail: sh1@le.ac.uk.

† Supplemental material for this article may be found at http://aem.asm.org/.

Our specific objective in this work was to identify novel protein-encoding open reading frames (ORFs) from eukaryotic microorganisms without their prior cultivation or identification. This requires the development of metagenomic cDNA technology: the ability to isolate full-length mRNAs, reverse transcribe them, and clone the cDNA to make libraries for sequence and expression studies. As far as we are aware, no attempts to specifically target eukaryotic mRNA from environmental samples have been previously described. We report here procedures suitable for stabilizing RNA in environmental samples in the field such that they can be transported back to the laboratory, the RNA isolated, and cDNA libraries made for subsequent sequencing and expression studies. The ability to stabilize RNA in environmental samples for subsequent purification and analysis in the laboratory may also have other uses, such as measuring changes in the environmental transcriptome of both prokaryotes and eukaryotes with time and in response to external change.

MATERIALS AND METHODS

Samples. *Escherichia coli* strain XLOLR (Stratagene) was grown overnight at 37°C in Luria Bertani (LB) broth. *Saccharomyces cerevisiae*, commercial baker's yeast, was grown overnight in malt extract broth (Oxoid) at 37°C. *Acanthamoeba polyphaga* strain ROS (kindly supplied by Simon Kilvington, Department of Infection Immunity and Inflammation, Leicester University) was grown at room temperature for 5 days in a semidefined axenic culture medium to produce trophozoites (15). Cells were pelleted by centrifugation at 1,000 × *g* for 10 min, and RNA was extracted immediately or the pellets were resuspended in RNAlater, an RNA stabilization solution (Ambion), prior to storage at −20°C. RNAlater is a proprietary compound mixture, fully miscible with water, with no dangerous components or hazards. Its precise chemical composition is not published. Clearly, any RNA stabilization buffer for tissues and cells must be rapidly permeable and denature or otherwise inactivate RNases and prevent degradation of RNA. A chaotropic, metal chelating, moderately acidic solution of 25 mM sodium citrate, pH 5.2, 10 mM EDTA, and 10 M ammonium sulfate has been reported to have properties similar to those of RNAlater (http://pen2.igc.gulbenkian.pt/cftr/vr/a/clarke_rnase_retarding_solution_storage_transport_to_rna_extraction.pdf). Two algal mat samples were collected from hot spring sites in Yunnan province, People's Republic of China, in April 2003. These were immediately mixed with a 10-fold excess of RNAlater and stored on ice for up to 8 days prior to return to the laboratory where they were stored at −20°C. Ten milliliters of sample TC2 was obtained in Rehai, a geothermal location and national park in Tengchong County, with coordinates 98° 26′ E, 24° 57′ N, elevation 1,520 m. The material was a dark green microbial mat collected from the Drumbeat Spring secondary source, which had a temperature of 52°C and a pH of 8.5. Thirty milliliters of sample LP4 came from Long Pu, site of an incomplete thermal spa development located at 98° 23′ E, 24° 54′ N, elevation 1,119 m. It was a green gelatinous microbial mat, 3 cm thick, stratified with a green upper layer with a brown fragmented and degraded layer beneath collected from the walls of a deep basin, with a temperature of 60 to 65°C and a pH of 8.5. Samples were also extracted on site using the GenomicPrep cells and tissue DNA isolation kit (Amersham Pharmacia). Activated sludge was obtained from a local sewage treatment plant run by Severn Trent Water Ltd. at Wanlip, Leicestershire, United Kingdom. This was pelleted by centrifugation (1,000 × *g* for 5 min), and RNA was extracted immediately on return to the laboratory, or the pellet was resuspended in RNAlater prior to storage at −20°C within 1 h of sampling. Samples were also extracted on return to the laboratory using the GenomicPrep cells and tissue DNA isolation kit (Amersham Pharmacia).

RNA extraction. Total RNA was extracted from about 3 g (wet weight) of the samples using the QIAGEN RNeasy mini kit. For stored material, RNAlater was removed after centrifugation, and the sample was resuspended in the lysis buffer provided in the kit. In the case of the *Acanthamoeba* material, the amoebae were disrupted in the lysis buffer by repeated pipetting, and homogenization was ensured by passing the lysate down a QIAshredder column (QIAGEN) by following the manufacturer's protocol. The rest of the method followed the RNeasy protocol. The extraction from *E. coli* was preceded by lysozyme treatment (TE buffer, pH 8.0, with 400 μg/ml lysozyme) as detailed in the QIAGEN RNeasy protocol. For the algal mat, yeast, and activated sludge materials, a similar protocol was followed, but instead of using a QIAshredder column, disruption and homogenization were carried out by bead beating with the FastPrep apparatus (BIO 101, Qbiogene) using tubes containing lysing matrix E and a setting of 5.5 for 30 s. After centrifugation to pellet the debris, the rest of the method followed the RNeasy protocol. The eluted RNA was stored at −20°C after addition of the RNase inhibitor SUPERase-In (Ambion).

Purification of mRNA from total RNA. For the activated sludge material, poly(A)-tailed mRNA was extracted from about 300 μg of total RNA using the poly(A) Purist MAG kit (Ambion). The method involves binding of the poly(A) tails to oligo(dT) magnetic beads, capture of the beads magnetically, and elution of the poly(A) RNA. The RNA was then ethanol precipitated and resuspended in a small volume (15 μl) of RNA storage solution supplied with the kit.

Library construction. The Smart cDNA library construction kit (BD Biosciences, Clontech) was used to create Lambda cDNA libraries in the phagemid Lambda TriplEx2 vector. The method uses primers for reverse transcription (RT) that should optimize the production of full-length transcripts. First-strand DNA synthesis is based on a dT-rich oligonucleotide hybridizing to poly(A) RNA sequences such as at the 3′ end of mRNAs. Second-strand synthesis only occurs if the reverse transcriptase reaches the 5′ end of the RNA and adds additional non-template-encoded C residues. It also results in directional cloning of the cDNA and can generate polypeptides from all three reading frames in a single recombinant Lambda TriplEx2 clone. Plasmid can be excised from the lambda phagemid. The starting material for cloning can be total RNA or purified mRNA.

The cDNA library construction was initially carried out from total RNA extracts rather than from mRNA, since the amount of total RNA from most environmental samples is likely to be small. However, since abundant quantities of activated sludge were available for comparison, a library was also constructed from the mRNA isolated from this sample. The RT step was carried out using PowerScript reverse transcriptase with the SMART IV primer (5′-AAGCAGTGGTATCAACGCAGAGTGGCCATTACGGCCGGG-3′) and the lock-docking oligo(dT) primer CDS III/3′ [5′-ATTCTAGAGGCCGAG GCGGCCGACATG-$d(T)_{30}$ (AGC)N-3′] provided in the kit. This should result in full-length single-stranded cDNA containing a sequence complementary to the SMART IV Oligo, which then serves as a universal priming site for subsequent amplification by long-distance PCR (LD-PCR). This was carried out with the Advantage 2 PCR kit and the 5′ PCR primer (5′-AAGCAGTGGTATCAA CGCAGAGT-3′) with the CDS III/3′ PCR primer used for the RT step to produce double-stranded cDNA. After digestion with SfiI to produce suitable restriction sites for cloning, the construction of the library essentially followed the manufacturer's protocol, except that size fractionation of the double-stranded cDNA (pooled from 4 to 6 PCRs) was carried out after agarose gel electrophoresis. Material of approximately 500 to 5,000 bp in size was extracted using the QIAGEN QIAEX II gel extraction kit. The cDNA was ethanol precipitated, ligated into the Lambda TriplEx2 vector, and packaged with the Stratagene Gigapack III Gold packaging extract, and titers were determined using *E. coli* XL1-Blue as the host. Blue/white screening of plaques was carried out to assess the percentage of clones with inserts. After amplification, the completed cDNA libraries were stored in 7% dimethyl sulfoxide at −80°C.

Characterization of cDNA inserts. The inserts in a number of clones from each library were sequenced after PCR amplification from individual plaques excised into nanopure water using vector-encoded amplification primers (5′-CTCGGG AAGCGCGCCATTGTGTTGG and 3′-ATACGACTCACTATAGGGC). This was carried out using *Taq* polymerase (Abgene), with an initial denaturation for 2 min at 94°C, followed by 30 cycles with parameters of 94°C for 30 s and 68°C for 3 min, as indicated by the BDClontech protocol. The products were cleaned with the QIAquick PCR purification kit (QIAGEN), and sequenced using the 5′ PCR primer listed above by Lark Technologies, Takely, United Kingdom. Some clone sequences were completed by primer walking.

PCR and characterization of 18S rRNA genes. DNA was extracted from environmental samples which had been stored at −20°C using the GenomicPrep cells and tissue DNA isolation kit (Amersham Pharmacia). In the case of *Acanthamoeba polyphaga*, the kit was used to extract DNA from pelleted freshly grown cells to provide a positive control. PCR amplification of approximately the first 520 bp of the 18S rRNA gene was carried out with forward primer 5′-CCG AAT TCG TCG ACA ACC TGG TTG ATC CTG CCA GT-3′ and reverse primer 516R 5′-ACC AGA CTT GCC CTC C-3′. The program consisted of 2 min of denaturation at 95°C and 30 cycles of 95°C for 30 s, 55°C for 40 s, and 72°C for 2 min, with a final 10-min extension at 72°C. The cleaned PCR products were ligated into the pGEM-T Easy vector and transformed into JM109 high-efficiency competent cells (Promega), with blue/white screening. Colony PCR was performed using M13 primers to amplify the inserts in a number of clones. In the case of the activated sludge clones, the PCR products in individual clones were screened using restriction digestion with HaeIII and clones with different restric-

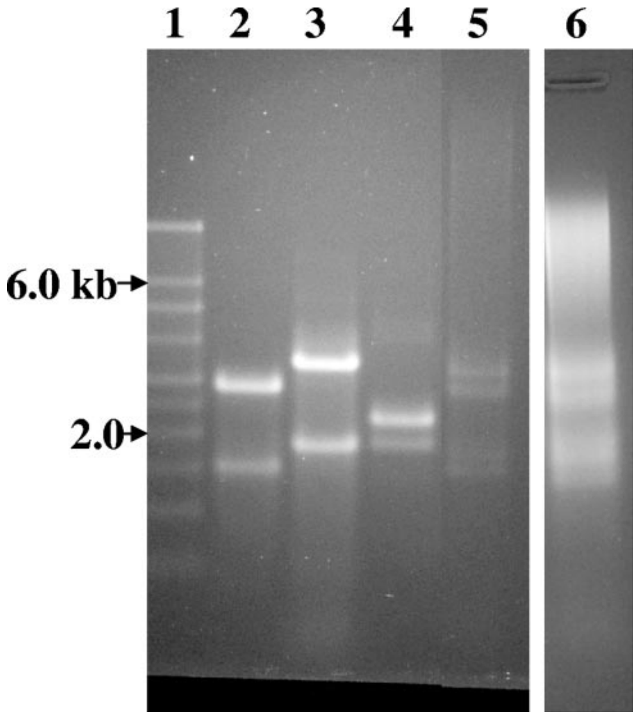

FIG. 1. Typical results of total RNA extraction from named species and environmental samples electrophoresed on a 1.2% denaturing formaldehyde agarose gel. Lane 1, RNA size markers (Ambion Millennium markers); lane 2, *Escherichia coli*; lane 3, *Saccharomyces cerevisiae*; lane 4, *Acanthamoeba polyphaga*; lanes 5 and 6, activated sludge.

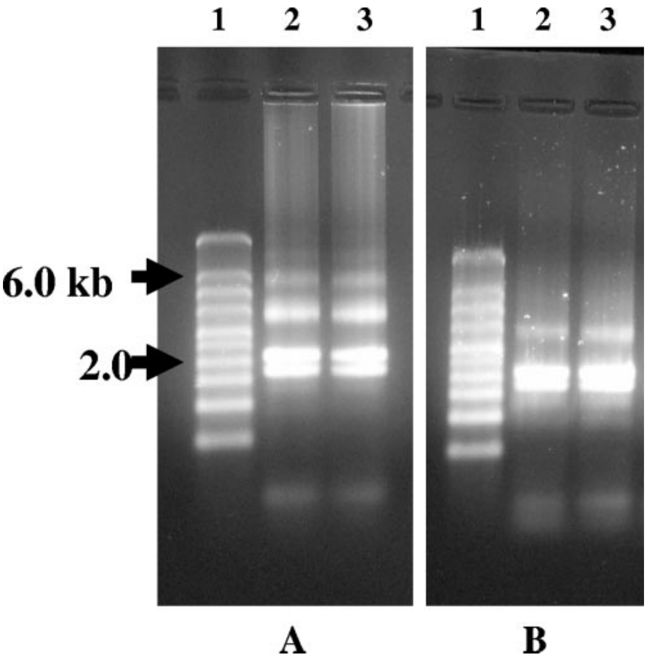

FIG. 2. Effect of storage in RNAlater on the stability of total RNA extracted from *Acanthamoeba polyphaga*. (A) TBE 1.2% agarose gel. Lane 1, RNA size markers (Millennium markers; Ambion); lanes 2 and 3, total RNA from a fresh culture of *Acanthamoeba polyphaga.* (B) Same as panel A, but lane 2, total RNA from *A. polyphaga* stored in RNAlater for 10 days at ambient temperature; lane 3, total RNA from *A. polyphaga* stored for 10 days at 4°C.

tion fragment length polymorphism patterns were selected for sequencing using the forward 18S primer.

Clone sequence analysis. Nucleotide sequences were analyzed using the ORF finder and BLAST (2) facilities at NCBI (http://www.ncbi.nlm.nih.gov/) during May 2005. Complete nucleotide sequences were compared using BLASTN and BLASTX. ORFs identified using ORF finder were compared using BLASTP.

Nucleotide sequence accession number. The sequences of the clones from the cDNA libraries described here have been deposited in the EMBL database. Those from the *A. polyphaga* library have accession numbers AJ876805 to AJ876809, those from the TC2 algal mat library have accession numbers AJ879837 to AJ879844, and those from the LP4 algal mat library have accession numbers AJ879808 to AJ879836. Sequences from the activated sludge library made with total RNA have accession numbers AJ879846 to AJ879868, and those from the activated sludge library made from poly(A)-enriched RNA have accession numbers AJ879869 to AJ879891.

RESULTS

RNA stabilization and extraction. Total RNA was extracted from *Escherichia coli*, *Saccharomyces cerevisiae*, *Acanthamoeba polyphaga*, and activated sludge to establish methods suitable for a range of organisms and samples. It was analyzed by electrophoresis on TBE (Tris-borate-EDTA) and formaldehyde denaturing agarose gels with similar results; Fig. 1 shows the formaldehyde denaturing gel. The RNA isolated from *E. coli* (lane 2) and *S. cerevisiae* (lane 3) both contain, as expected, predominantly rRNA with two major bands. These corresponded to the to 16S small subunit (SSU) species and the 23S large subunit (LSU) species in the case of *E. coli* and the typical eukaryotic 18S SSU (1,799 nucleotides) and the 28S LSU (3,394 nucleotides) in the case of *S. cerevisiae*. For *A. polyphaga* (lane 4), we observed again two major RNA bands, the presumptive SSU RNA had a similar mobility to the yeast 18S RNA. The presumptive LSU species appears to be considerably smaller than the yeast 28S LSU (no sequence information could be found in the databases for the *A. polyphaga* LSU gene). In the activated sludge sample, there were four bands, presumably corresponding to large and small rRNA subunits of both prokaryotes and eukaryotes (lane 5 and lane 6 [overloaded]).

Investigations were also carried out to assess whether RNA in environmental samples could be stabilized after collection under field conditions. Cultures or environmental samples were pelleted, then resuspended in RNAlater, and stored at various temperatures for various times. Total RNA was extracted from stored samples (Fig. 2B) and compared with RNA extracted immediately from fresh samples (Fig. 2A) following gel electrophoresis. RNA isolated from *Acanthamoeba* was stable for at least 10 days at room temperature, as shown in Fig. 2B, lane 2. No evidence of degradation was observed with yeast stored for 10 days at 4°C or activated sludge stored for 3 months at −20°C (data not shown). All environmental samples were placed in RNAlater upon collection, refrigerated within 4 h, and put at −20°C within 8 days.

Reverse transcription. The synthesis of cDNA was carried out by reverse transcription followed by LD-PCR, as detailed in the Clontech protocol. Figure 3 shows the products for *A. polyphaga* (panel B), algal mat LP4 (panel C) and activated sludge (panel D) compared with that for the control human placental mRNA provided in the kit (panel A). Sufficient ma-

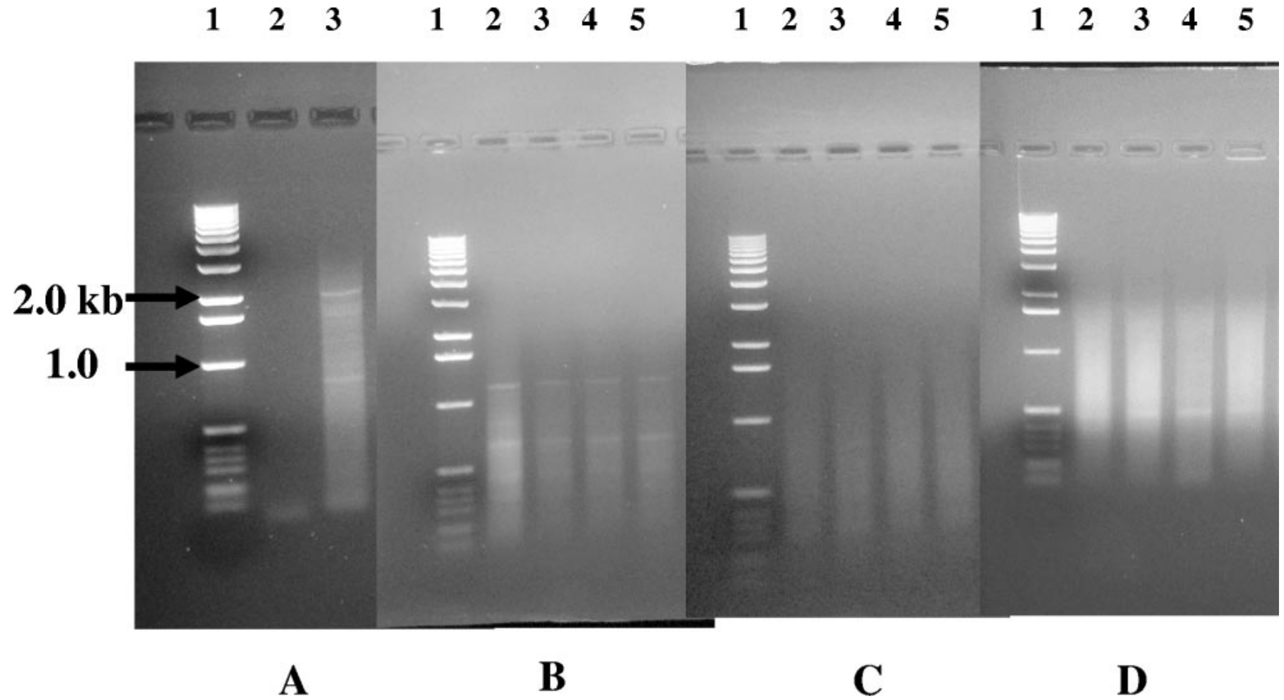

FIG. 3. Results of reverse transcription and LD-PCR on RNA from different sources to produce cDNAs run out on TBE 1.2% agarose gels. (A) Lane 1, 1-kb DNA ladder (Invitrogen); lane 2, negative control; lane 3, product from the human placental mRNA provided as a control in the Smart cDNA Library construction kit. (B) Lane 1, 1-kb ladder; lanes 2 to 5, product from total RNA extracted from *Acanthamoeba polyphaga.* (C) Same as panel B, showing product from the LP4 Chinese algal mat. (D) Same as panel B, showing product from activated sludge total RNA.

terial for subsequent cloning was obtained by pooling the products of four to six LD-PCRs. Much more cDNA was made from the activated sludge sample. No LD-PCR product was obtained if the reverse transcription step was omitted, confirming that the template for amplification was indeed RNA and not residual DNA.

Library properties. The properties of the five metagenomic cDNA libraries constructed are shown in Table S1 in the supplemental material. Primary titers and amplified titers were all similar, 2.5×10^5 to 7.5×10^5 and 0.35×10^9 to 8×10^9, respectively. Blue/white screening showed that the TC2 library had the highest number of clones with inserts (94%), while the activated sludge libraries had the lowest number (70% and 73%, respectively).

Sequences of inserts in Lambda clones. The sequences of inserts in individual clones from each of the libraries were determined after PCR amplification from isolated plaques. Most ranged in size from about 500 bp to 2 kb. Larger clones were not found, probably because small DNA fragments tend to be preferentially cloned into the vector. Sequencing was carried out using the 5′ forward primer, resulting in up to 800 bases of sequence, depending on the size of the insert. Sequencing from the 3′ end of the insert was unsuccessful, probably because of the poly(A) tails. Clone sequences were therefore completed by primer walking. All were found to have sequence at the 5′ end corresponding to part (shown in boldface type) of the 3′ end of the SMART IV primer (5′-AAG CAGTGGTATCAACGCAGAGT**GGCCATTACGGCCGGG-**3′) used for first-strand cDNA synthesis. This indicates that cloning into the vector was indeed directional. It also indicates that the cloned products contain a sequence incorporated during the synthesis of cDNA, i.e., that we are cloning RNA sequences. This latter point is important and is confirmed by the absence of a PCR product if the reverse transcription step of the protocol was omitted, as described above. poly(A) tails at least 30 bases in length were evident for all fully sequenced clones, corresponding to the CDS III/3′ PCR primer.

Sequence analysis. If mRNA sequences had been cloned, the expectation would be that they would have an ORF over most of the insert length in a forward reading frame from the 5′ end. In subsequent analysis, therefore, only ORFs in the 5′ to 3′ direction were considered. Where the complete insert up to the poly(A) region has been sequenced, a stop codon would also be expected.

The prediction of ORFs in sequences derived from the metagenome is complicated by a significant variation in codon usage

TABLE 1. Results of sequencing clones

Library	Total no. of sequences	No. of possible protein sequences	No. of ribosomal RNA sequences	No. of sequences not identified
001 *Acanthamoeba polyphaga*	5	4	0	1
002 Algal mat LP4	34	6	17	11
003 Algal mat TC2	19	0	15	4
004 Activated sludge total RNA	24	13	5	6
005 Activated sludge poly(A) RNA	23	16	1	6

TABLE 2. Activated sludge total RNA cDNA clones with possible matches to database proteins[a]

Clone	Accession no.	Read length (bp)	ORF length (aa[b])	Closest match
ASt-7	AJ879845	578	164	E = 2e − 63 (129/167, 77%) to actin binding protein (281 aa) of *Physarum polycephalum* (S32566); poly(A)s reached; has TAA stop codon
ASt-19	AJ879846	690	99	E = 1e − 12 (39/101, 38%) to putative ribosomal protein L10a of *Oryza sativa* (BAD28853); poly(A)s reached; has TGA stop codon BLASTX E = 1e − 32 (78/213, 36%) 60S ribosomal protein L10a (214 aa) of *Entamoeba histolytica* (EAL48615)
ASt-28	AJ879847	867	87 85	Both ORFs in same reading frame, match to eukaryotic cysteine proteases; BLASTX E = 6e − 47(104/263, 39%) cysteine protease (474 aa) of *Daucus carota* (BAD29954); poly(A)s, ~950; has TAA stop codons at end of both ORFs
ASt-30	AJ879848	807	259	E = 1e − 07 (37/95, 38%) Arg, Leu-rich conceptual translation (114 aa) of *Paramecium bursaria Chlorella* virus 1 (NP_048556); poly(A)s reached; has TAA stop codon
ASt-33	AJ879849	1,052	315	E = e − 135 (217/311, 70%) to CPC2 protein (316 aa) of *Neurospora crassa* (CAA57460); CD WD40 modules; poly(A)s reached; has TGA stop codon
ASt-49	AJ879850	823	257	E = 1e − 17 (81/295, 27%) to Lembadion factor *Lembadion bullinum* (CAA70420, 350 aa); poly(A)s reached; has TAA stop codon
ASt-57	AJ879851	1,052	322	E = 5e − 16 (45/120, 37%) surface protein Sdr1 (length, 1,893 aa) of *Staphylococcus saprophyticus* (AAM90673); CD ankyrin repeats; poly(A)s not reached (~1,750 bp), therefore, no stop codon
ASt-60	AJ879852	776	188	E = 1e − 28 (62/188, 32%) to COG2202: FOG: PAS/PAC domain (991 aa) of *Methanococcus burtonii* (ZP_00147582); CD histidine kinase-like ATPases; poly(A)s, ~975 bp; has TAG stop codon
ASt-61	AJ879853	803	200	E = 5e − 57 (108/180, 60%) to ribosomal protein S2 (201 aa) of *Branchiostoma lanceolatum* (AAN77880); poly(A)s reached; has TAA stop codon
ASt-65	AJ879854	780	199	E = 2e − 53 (94/159, 59%) unnamed protein (179 aa) of *Tetraodon nigroviridis* (CAF90670); conserved domain RAB GTPase; poly(A)s reached; has TGA stop codon
ASt-73	AJ879855	1,064	37 **270**	E = 0.003(18/36, 50%) probable serine-type carboxypeptidase (2,105 aa, T18968) of *Caenorhabditis elegans;* TAA stop; BLASTX E = 7e − 25(99/324, 30%) putative serine carboxypeptidase II of *Arabidopsis thaliana* (AAM91708, 479 aa); poly(A)s reached; conserved domain peptidase_S10
ASt-74	AJ879856	764	44 102 **233**	Both ORFs in same reading frame, TAA stop codons; both with conserved domains to pfam 00996, GDI GDP dissociation inhibitor; BLASTX E = 4e − 71 (135/239, 56%) putative rab GDI alpha protein (473 aa) of *Cryptosporidium parvum* (EAK87451); poly(A)s reached
ASt-86	AJ879857	588	168	E = 3e − 09 (43/119, 36%) neurofilament medium subunit (487 aa) of *Serinus canaria* (AAC06245); poly(A)s reached; has TAA stop codon

[a] The expected BLASTP value (E) is shown along with the percentage of amino acid homology to a database protein with the given accession number. CD indicates a match in the NCBI conserved domains database; COG indicates a match with the NCBI Clusters of Orthologous Groups of proteins. Boldface type in the ORF column indicates a code 6 translation.

[b] aa, amino acids.

in different organisms, both for translational initiation and termination (detailed for many species in the Codon Usage database tabulated) from GenBank (http://www.kazusa.or.jp/codon) (22), leading to a choice of 22 genetic codes. Initiation is most efficient from AUG, but in rare cases, other codons are utilized, e.g., in the yeast *Candida albicans*, molds, protozoans, coelenterate mitochondria, and mycoplasmas (1, 6, 29; NCBI taxonomy browser). Similarly, differences in the usage of termination codons have been observed in ciliated protozoa (13), and in some ciliates, stop codons are reassigned to sense codons (19). Clearly, the appropriate code can be selected when sequencing clones from a known organism, but choice is more problematic for environmental isolates. Accordingly, we used the standard genetic code in the sequence analysis of the clones in this study reported in Tables 1, 2, and 3 and in Table S2 in the supplemental material. Different results were obtained using genetic code 6 (ciliate, dasycladacean, and *Hexamita* code), where TAA and TAG stop codons are suppressed and read as glutamine. These results are included in Tables 2 and 3, see Discussion for more details. E values in Tables 2 and 3 and Table S2 in the supplemental material are derived from BLASTP unless referred to as BLASTX.

(i) *Acanthamoeba polyphaga* cDNA library. The inserts in only a few clones from the *Acanthamoeba polyphaga* library were sequenced to validate the methodology. Using the standard code in ORF finder, BLASTN, and BLASTX searches, four of five clones were found to have some identity to proteins in the databases and are designated as possible proteins in Table 1. One clone showed no significant identity to any database entry and is designated as not identified. The complete results are presented in the supplemental material. TBLASTX was no more informative than BLASTX when analyzing this library or the ones described below.

(ii) Algal mat libraries. Thirty-four LP4 and 19 TC2 cDNA clones were sequenced and analyzed using ORF finder, BLASTN, and BLASTX searches (Table 1). For the LP4 library, 17 clones of various lengths had high similarity to prokaryote rRNA gene sequences in the databases. Their identities are reported in the supplemental material. Only 6 clones from LP4 appeared to be related to protein sequences in the databases. The matches to proteins in the databases are generally quite high (e^{-30} to e^{-102}). However, in 4 of 6 clones, no stop codon is evident at the end of the ORF, which reaches the end of the cloned sequence. In addition, these ORFs are short

TABLE 3. Activated sludge mRNA cDNA clones with possible matches to database proteins[a]

Clone	Accession no.	Read length (bp)	ORF length (aa[b])	Closest match
ASm-4	AJ879870	859	224	E = 2e − 08 (51/147, 34%) to Lembadion factor (350 aa) of *Lembadion bullinum* (CAA70420); poly(A)s reached; has TAA stop codon
ASm-10	AJ879872	786	44 87 34 **244**	All three ORFs in same reading frame, match to 14-3-3 homologues; TAA stop codon for first two, third no stop; BLASTX E = 2e − 74 (145/235, 61%) 14-3-3 protein (244 aa) of *Tetrahymena pyriformis* (BAA83080); poly(A)s reached
ASm-13	AJ879873	806	244	E = 2e − 28 (83/247, 33%) hypothetical protein (263 aa) of *Plasmodium falciparum* (NP_700790); conserved domain DER1 family; poly(A)s reached; has TGA stop codon
ASm-16	AJ879875	696	56	E = 1e − 13 to cathepsin H (366 aa, AAO61485) of *Sterkiella histriomuscorum*; BLASTX E = 3e − 60 (120/192, 62%) cysteine protease (350 aa, CAA92583) of *Pisum sativum* poly(A)s reached; TAA stop codon
ASm-21	AJ879876	1,086	276	E = 2e − 13 (54/235, 22%) similar to CG6004-PB (1,542 aa) of *Mus musculus* (XP_488010); poly(A)s reached; has TAA stop codon
ASm-25	AJ879877	1,783	180 375	Contiguous ORFs with TAA stop codon between and TAG at end; BLASTX E = 4e − 09 (52/137, 37%) predicted protein of *Entamoeba histolytica* (298 aa, EAL42803). Conserved domain ABC ATPase; poly(A)s reached
ASm-31	AJ879878	509	142	E = 0.20 (33/98, 33%) to putative cytochrome P450 (521 aa) of *Oryza sativa japonica* (XP_464369); poly(A)s reached; has TAA stop codon
ASm-32	AJ879879	379	95	E = 4e − 11 (34/92, 36%) to cystatin (98 aa, CAD20980) of *Lepidoglyphus destructor* (CAD20980); poly(A)s reached; has TAA stop codon
ASm-36	AJ879880	566	102	E = 1e − 08 (35/102, 34%) hypothetical protein (131 aa) of *Yarrowia lipolytica* (CAG82780); conserved domain for ribosomal protein L27e; poly(A)s reached; has TAA stop codon
ASm-45	AJ879883	1,090	323	E = 0.002 (24/59, 40%) to latency-associated antigen-related protein (316 aa) of *Plasmodium yoelii* (EAA19832); poly(A)s reached; has TAA stop codon
ASm-46	AJ879884	728	33	Conserved domain found to ubiquitin homologs; BLASTX E = e − 110 (134/146, 91%) ubiquitin (229 aa, AAF00920) of *Oxytricha trifallax*; poly(A)s reached; has TAA stop codon
ASm-59	AJ879887	1,376	378	E = 0.0 (351/375, 93%) to beta tubulin (conserved domain found) of *Toxoplasma gondii* (449 aa, S16340); poly(A)s reached; has TAA stop codon
ASm-60	AJ879888	462	146	E = 0.014 (33/114, 28%) NtEPc-like protein (179 aa) of *Nicotiana tabacum* (BAC53926); poly(A)s reached; has TAA stop codon
ASm-61	AJ879889	562	153	E = 3e − 46 (91/143, 63%) to 40S ribosomal protein S16 (145 aa) of *Gossypium hirsutum* (CAA53567); poly(A)s reached; has TAA stop codon
ASm-62	AJ879890	659	123	E = 0.001 (23/64, 35%) putative lipocalin (189 aa) of *Acinetobacter* sp. strain ADP1 (accession no. YP_046294); conserved domain bacterial lipocalin; poly(A)s reached; has TAA stop codon
ASm-63	AJ879891	460	137	E = 6e − 32 (66/133, 49%) to ribosomal protein L32 (134 aa, conserved domain found) of *Branchiostoma belcheri* (AAO31774); poly(A)s reached; has TAA stop codon

[a] Expected BLASTP value (E) is shown along with the percentage of amino acid homology to a database protein with the given accession number. CD indicates a match in the NCBI conserved domains database; COG indicates a match with the NCBI Clusters of Orthologous Groups of proteins. Boldface type in the ORF column indicates a code 6 translation.

[b] aa, amino acids.

compared to the proteins they most closely match. The results are reported in Table S2 in the supplemental material. The remaining 11 clones sequenced from the LP4 library contained sequences with low relatedness to both nucleotide and protein database sequences. In most cases, ORFs in a forward reading frame were short, 35 to 80 amino acids, with only low similarity to known or hypothetical proteins. BLASTX searches similarly gave low matches over short stretches of the inserts. These are designated as not identified in Table 1.

For the TC2 library, most of the cloned inserts (15/19) gave the highest identity in BLASTN searches to rRNA gene sequences. Their identities are reported in the supplemental material. The remaining four sequences in the TC2 library have been designated as not identified in Table 1. ORFs in the forward direction were again short, with low matches to proteins in the databases.

(iii) Activated sludge libraries. The inserts in 24 clones from the activated sludge total RNA library and 23 from the one constructed from poly(A)-enriched RNA were sequenced. As summarized in Table 1, the majority of inserts in both libraries (13 and 16, respectively) showed significant similarity to proteins in the databases. Both gave 6 inserts which had only short ORFs and low matches in BLASTN or BLASTX searches, designated as not identified. Five sequences from the total RNA library and only one from the poly(A)-enriched RNA library were related to rRNA sequences, all most closely matched eukaryotic large subunit 26-28S rRNA genes, mostly from alveolates.

The results of the searches for clones having matches to database proteins are shown in Table 2 (total RNA) and Table 3 [poly(A) RNA library]. In most cases, the sequence of the complete insert was obtained, using primer walking for the larger clones. The first criterion for inclusion of a sequence in Tables 2 and 3 is that an ORF in the forward direction has been found which extends for most of the length of the insert. In some cases, the match to proteins in the databases is quite high (e.g., ASt-7, ASt-33, ASm-59, and ASm-61), while others are lower (e.g., ASt-30, ASt-49, ASm-4, and ASm-60). Lower matches might be expected given the diversity of organisms likely to be in the sample and the fact that, for many of the organisms present in the algal mat, there may be no sequences in the databases.

For some sequences, the length of the ORF does seem to correlate with the length of the protein it most closely matches (e.g., ASt-61, ASt-65, ASm-13, ASm-32, ASm-61, and ASm-63), suggesting that we have identified complete reading frames, but for others, this is not the case. Where the similarities are low, this might be expected, but it might also suggest that incomplete cDNA products have been cloned.

Also included as possible proteins are sequences for which only short forward ORFs appear (sometimes two or three in the same reading frame). However, these have reasonable matches to known proteins, sometimes having conserved domains, and the BLASTX results show that the match is in fact over most of the length of the insert. The stop codons curtailing these ORFs could be genuine or introduced by errors in reverse transcription or during PCR. Examples in Tables 2 and 3 are ASt-19, ASt-28, ASt-73, ASm10, ASm25, and ASm46. For some of these inserts (e.g., ASt-19 and ASm-10), the BLASTX analysis does identify a protein of comparable length to the closest match, even though the corresponding ORF(s) seem prematurely terminated. An alternative explanation, at least for some of these sequences, e.g., ASt-73, ASt-74 and Asm-10, is that the termination signal is suppressed and a longer protein is made (Tables 2 and 3; see also Discussion).

Overall, analysis of the longest ORFs in clones from the environmental cDNA libraries also shows qualitative differences between the activated sludge and the Chinese algal mat libraries. In the case of TC2, only 3 sequences have the longest putative ORF in the forward direction, whereas 16 had the longest in the reverse orientation. For LP4, only 8 were forward and 26 were reverse. The activated sludge libraries both had many more clones with the longest ORF in the forward direction, 18 for the library made from total RNA and 17 for the mRNA-derived library. Both had only 6 sequenced clones where the longest ORF was in a reverse reading frame. This supports the conclusion that more of the inserts in the activated sludge libraries were directionally cloned and derived from mRNA.

18S rRNA gene amplification. The cDNA product yield obtained following reverse transcription (Fig. 3) and the library clone sequencing (Tables 1, 2, 3; see Table S2 in the supplemental material) suggested that there were low levels of eukaryotic mRNA in the Chinese algal mat samples. Accordingly, 18S rRNA gene PCR amplification of the DNA samples was carried out to independently establish the presence or absence of eukaryotic material in the samples. The resulting PCR products are shown in Fig. 4; products of about 550 bp in size were expected. The *Acanthamoeba* material served as a control, yielding a single major band of the expected size (panel A, lane 3). A similarly sized PCR product was observed using activated sludge DNA as a template (lane 4). However, both LP4 and TC2 gave very weak signals, with multiple bands and little material of the expected size (panel B, lanes, 4, 5, 6). Amplification products from the activated sludge LP4 and TC2 samples were cloned and sequencing was carried out. Sequencing confirmed the presence of 18S gene products in the activated sludge libraries and LP4; no 18S sequences were identified in the TC2 library. The sequencing results are presented in the supplemental material.

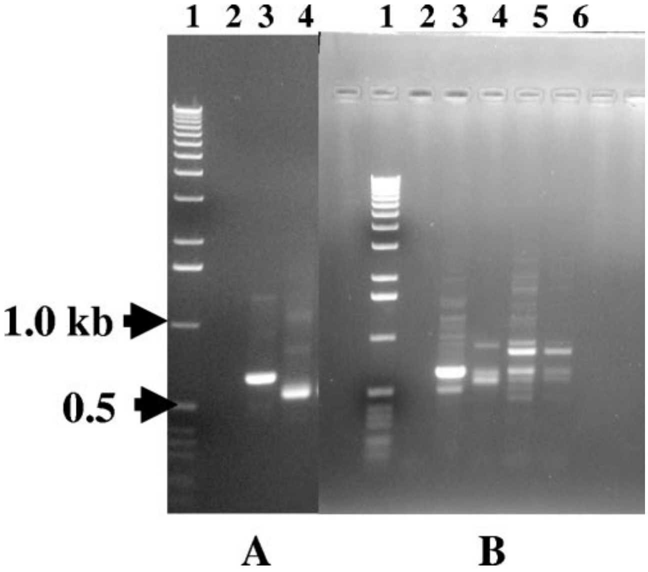

FIG. 4. 18S rRNA gene amplification from algal mat samples and activated sludge. (A) Lane 1, 1-kb ladder; lane 2, negative control; lane 3, *Acanthamoeba polyphaga* positive control; lane 4, product from activated sludge. (B) Lane 1, 1-kb ladder; lane 2, negative control; lane 3, *Acanthamoeba polyphaga* positive control; lane 4, algal mat LP4; lanes 5 and 6, product from two different DNA extracts from algal mat TC2.

DISCUSSION

Here we describe the construction of metagenomic cDNA libraries from environmental RNA samples. The success of the procedure with respect to identification of eukaryotic sequences depends strongly on the sample material. PCR amplification of the algal mat samples using primers for the eukaryote 18S ribosomal subunit RNA indicated only limited eukaryotic content in the LP4 sample and little or none in TC2. It is hardly surprising, therefore, that these algal mat samples yielded clones predominantly derived from prokaryotic RNA. None of the 19 clones from the TC2 library contained sequences readily identified as coding for proteins, whereas 6/34 clones from LP4 contained sequences that may code for bacterial proteins (see Table S2 in the supplemental material). The T-rich oligonucleotide used to initiate reverse transcription of RNAs in this study should enrich for poly(A)-containing sequences characteristic of eukaryotic mRNAs. So why have we identified prokaryotic sequences? It is now clear that polyadenylation of mRNA is widespread in prokaryotes (27). The LP4 clones, which may code for proteins, therefore, could be derived from prokaryotic mRNA rather than the more limited eukaryotic material.

Libraries produced from the algal mat samples also contained a large proportion of clones (17/34 in LP4 and 15/19 in TC2) (Table 1) matching prokaryotic rRNA sequences. These matches were mainly to cyanobacterial 23S rRNA sequences (see the supplemental material). Their abundance could simply be the result of mispriming during reverse transcription because of the large amounts of prokaryotic rRNA evidently present in the samples. Alternatively, polyadenylation of the prokaryotic rRNAs in these algal mats could account for the large percentage of such sequences in these libraries. Polyadenylation has also been shown to be involved in the degradation of ribosomal RNAs in prokaryotes, and in *E. coli*, the 23S rRNA is the major polyadenylated RNA (20).

RNA isolated from activated sludge contained substantial amounts of both prokaryotic and eukaryotic RNA, evident from the presence of 16S/18S and 23S/28S rRNA gene doublets (Fig. 1, lanes 5 and 6). Libraries made from this RNA were qualitatively different than the libraries made from the algal mat RNA described above (Table 1). There were many more clones matching possible protein sequences and many fewer matching rRNA sequences. This difference was even more pronounced with the library made from poly(A)-enriched RNA compared to total RNA (Table 1). The database matches were also mostly against eukaryotic sequences. Both the total RNA and the poly(A)-enriched RNA libraries had the same proportion of clones designated as not identified (6/24 and 6/23, respectively). If these clones were not derived from mRNA, we would have expected the proportion to decline in the library made from poly(A)-enriched RNA. As they did not, it seems likely that these sequences are in fact derived from mRNAs with no closely matching protein sequences present in the databases.

For most of the activated sludge clones detailed in Tables 2 and 3, complete sequences were obtained that reached the poly(A) tail, except for ASt-57 and ASt-60. In most cases, the apparent protein-encoding ORFs have a stop codon (usually TAA) when using the standard genetic code for the search. For some clones, e.g., ASm-21 and ASt-60, the encoded ORF protein matches part of a larger protein in the database. These could be domain matches. Other clone sequences appear to match database proteins closely in size and are probably full length. For example, ASm-32 has an ORF of 95 amino acids, with 36% identity to a cystatin ORF of 98 amino acids from the dust mite *Lepidoglyphus destructor.* In ASm-61, an ORF of 153 amino acids has 63% identity to the 40S ribosomal protein S16 (145 amino acids) of *Gossypium hirsutum*. Similarly, a 137-amino-acid ORF in ASm-63 has 49% identity to ribosomal protein L32 (134 amino acids) of *Branchiostoma belcheri*. An example of a larger full-length ORF would be ASt-33 at 315 amino acids, corresponding to a protein of 316 amino acids from *Neurospora*.

The presence of clone sequences with good BLASTX matches to known proteins but having only short corresponding ORFs or two or three short ORFs with high identity to the same protein (Tables 2 and 3) may be due to the introduction of erroneous stop codons during reverse transcription or the subsequent amplification steps. Alternatively, it could be due to suppression of the apparent stop codon in the unknown organism from which the sequence was derived. Some organisms reassign stop codons to code for particular amino acids. In the ciliate, dasycladacean, and *Hexamita* nuclear code, both TAA and TAG code for glutamine instead of chain termination (code 6; NCBI taxonomy browser). Since most of these ORFs end with a TAA codon, using code 6 results in a longer single ORF. Using the standard code, ASm-10, for example, has three short ORFs in the same reading frame, all with matches to 14-3-3 proteins. Translation using code 6 gives an ORF of 244 amino acids with a high identity (E = 9e-85) to a 14-3-3 protein of 244 amino acids from *Tetrahymena pyriformis.* Similar results using code 6 are obtained for other clones, e.g., ASt-73, which instead of an ORF of only 37 amino acids then has an ORF of 270 amino acids, with a match to serine-type carboxypeptidases. With ASt-74, a standard code 44-amino-acid ORF lengthens to 233 amino acids with code 6. Although caution is needed in interpreting these results, in these cases, because of the continuity of the ORF match to database proteins, it does seem likely that the TAA codon is not being used for chain termination when these sequences are translated in the organisms from which they derived.

Longer ORFs are also found, as would be expected, for most of the sequenced clones when translated with code 6. ASt-2 (accession no. AJ879863), for example, which with standard code has no ORF in a forward reading frame and is assigned to the not identified category, with code 6 has an ORF of 283 amino acids with an E value of 3e-12 to an SAP DNA-binding domain-containing protein from *Dictyostelium discoideum* (see Pfam accession number PF02037). Other clones in this group, e.g., ASt-38 (accession no. AJ879865), when translated using code 6 (presumably inappropriately) still show only short ORFs, the longest being 92 amino acids, with low similarity to database proteins.

Another observation on the correct length off ORFs as predicted by the NCBI facility can be illustrated using clone ASt-60 (Table 2). This putative ORF of 188 amino acids starts with a methionine residue and runs from base 164 to 730 of the 776-bp sequence. However, the BLASTX result shows similarity of the translated sequence to the closest matching protein starting from base 23 of the clone. Of the 47 amino acids added to the N terminus by this analysis, 45% are identical and 60% are positively related. This level of similarity compares well with the overall value of BLASTP for the ORF itself of 32% identity, with 61% positive over the 188-amino-acid length. It would seem possible in this case that the true start codon is upstream of the methionine given by ORF finder. A similar situation is found for ASt-33, where ORF finder places a leucine as the start codon at nucleotide 126 of the insert. However, the BLASTX results show putative alignment of the translated sequence to the highest match protein from nucleotide 33. This could be the result of alternative initiation codons being used, as discussed previously.

This study has shown that RNA from environmentally complicated and diverse samples can be stabilized under field conditions for subsequent laboratory analysis. cDNA libraries containing both prokaryotic and eukaryotic sequences can be made. Preliminary screening for 18S rRNA genes, as shown in Fig. 4, would help to determine whether a sample is likely to yield a library containing eukaryotic mRNA sequences. In the case of the activated sludge, good results were obtained even for the library made from total RNA. The methods we describe clone cDNAs directionally in a vector capable of expressing all three reading frames of an ORF. Our libraries can be expression screened for enzyme activity. A screen of 50,000 clones for esterase activity using methods previously described by us (24) was not successful. Improvements in our methodology are undoubtedly possible. Our RNA extraction method (QIAGEN RNeasy mini kit) is reported in its promotional literature to be effective on thick-walled structures such as bacterial spores and yeast, but we have made no attempt to compare different extraction techniques on model eukaryotes. The second strand of cDNA synthesis should only occur if the 5′ end of the mRNA is copied, but other protocols may be better. A proofreading *Taq* enzyme would minimize possible mutational errors during PCR amplification of the cDNA. Even so, this work is a starting point for eukaryotic cDNA metagenomics. It is possible with reasonable efficiency to identify eukaryotic ORFs likely to code for full-length proteins. Comparing levels of environmental RNAs in response to time or changing conditions may also find uses in microbial ecology and physiology.

ACKNOWLEDGMENT

This research, MultigenomeAccess Technology for Industrial Catalysts, was supported by the European Commission research program "Quality of life and management of living resources" (QLRT-2001-01972).

REFERENCES

1. **Abramczyk, D., M. Tchorzewski, and N. Grankowski.** 2003. Non-AUG translation initiation of mRNA encoding acidic ribosomal P2A protein in *Candida albicans*. Yeast **20:**1045–1052.
2. **Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25:**3389–3402.
3. **Amann, R. I., W. Ludwig, and K. H. Schleifer.** 1995. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol. Rev. **59:**143–169.
4. **Amaral Zettler, L. A., F. Gomez, E. Zettler, B. G. Keenan, R. Amils, and M. L. Sogin.** 2002. Microbiology: eukaryotic diversity in Spain's River of Fire. Nature **417:**137.
5. **Barns, S. M., C. F. Delwiche, J. D. Palmer, and N. R. Pace.** 1996. Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. Proc. Natl. Acad. Sci. USA **93:**9188–9193.
6. **Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, B. A. Rapp, and D. L. Wheeler.** 2000. GenBank. Nucleic Acids Res. **28:**15–18.
7. **Botero, L. M., S. D'Imperio, M. Burr, T. R. McDermott, M. Young, and D. J. Hassett.** 2005. Poly(A) polymerase modification and reverse transcriptase PCR amplification of environmental RNA. Appl. Environ. Microbiol. **71:**1267–1275.
8. **Cottrell, M. T., J. A. Moore, and D. L. Kirchman.** 1999. Chitinases from uncultured marine microorganisms. Appl. Environ. Microbiol. **65:**2553–2557.
9. **DeLong, E. F., K. Y. Wu, B. B. Prezelin, and R. V. Jovine.** 1994. High abundance of Archaea in Antarctic marine picoplankton. Nature **371:**695–697.
10. **Fuhrman, J. A., K. McCallum, and A. A. Davis.** 1992. Novel major archaebacterial group from marine plankton. Nature **356:**148–149.
11. **Giovannoni, S. J., T. B. Britschgi, C. L. Moyer, and K. G. Field.** 1990. Genetic diversity in Sargasso Sea bacterioplankton. Nature **345:**60–63.
12. **Handelsman, J., M. R. Rondon, S. F. Brady, J. Clardy, and R. M. Goodman.** 1998. Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chem. Biol. **5:**R245–R249.
13. **Harper, D. S., and C. L. Jahn.** 1989. Differential use of termination codons in ciliated protozoa. Proc. Natl. Acad. Sci. USA **86:**3252–3256.
14. **Henne, A., R. Daniel, R. A. Schmitz, and G. Gottschalk.** 1999. Construction of environmental DNA libraries in *Escherichia coli* and screening for the presence of genes conferring utilization of 4-hydroxybutyrate. Appl. Environ. Microbiol. **65:**3901–3907.
15. **Hughes, R., and S. Kilvingon.** 2001. Comparison of hydrogen peroxide contact lens disinfection systems and solutions against *Acanthamoeba polyphaga*. Antimicrob. Agents Chemother. **45:**2038–2043.
16. **Hurt, R. A., X. Qiu, L. Wu, Y. Roh, A. V. Palumbo, J. M. Tiedje, and J. Zhou.** 2001. Simultaneous recovery of RNA and DNA from soils and sediments. Appl. Environ. Microbiol. **67:**4495–4503.
17. **Lopez-Garcia, P., F. Rodriguez-Valera, C. Pedros-Alio, and D. Moreira.** 2001. Unexpected diversity of small eukaryotes in deep-sea Antarctic plankton. Nature **409:**603–607.
18. **Lorenz, P., K. Liebeton, F. Niehaus, and J. Eck.** 2002. Screening for novel enzymes for biocatalytic processes: accessing the metagenome as a resource of novel functional sequence space. Curr. Opin. Biotechnol. **13:**572–577.
19. **Lozupone, C. A., R. D. Knight, and L. F. Landweber.** 2001. The molecular basis of nuclear genetic code change in ciliates. Curr. Biol. **11:**65–74.
20. **Mohanty, B. K., and S. R. Kushner.** 2000. Polynucleotide phosphorylase, RNase II and RNase E play different roles in the *in vivo* modulation of polyadenylation in *Escherichia coli*. Mol. Microbiol. **36:**982–994.
21. **Moon-van der Staay, S. Y., R. De Wachter, and D. Vaulot.** 2001. Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. Nature **409:**607–610.
22. **Nakamura, Y., T. Gojobori, and T. Ikemura.** 2000. Codon usage tabulated from the international DNA sequence databases: status for the year 2000. Nucleic Acids Res. **28:**292.
23. **Poretsky, R. S., N. Bano, A. Buchan, G. LeCleir, J. Kleikemper, M. Pickering, W. M. Pate, M. A. Moran, and J. T. Hollibaugh.** 2005. Analysis of microbial gene transcripts in environmental samples. Appl. Environ. Microbiol. **71:**4121–4126.
24. **Rees, H. C., S. Grant, B. Jones, W. D. Grant, and S. Heaphy.** 2003. Detecting cellulose and esterase enzyme activities encoded by novel genes present in environmental DNA libraries. Extremophiles **7:**415–421.
25. **Rondon, M. R., P. R. August, A. D. Bettermann, S. F. Brady, T. H. Grossman, M. R. Liles, K. A. Loiacono, B. A. Lynch, I. A. MacNeil, C. Minor, C. L. Tiong, M. Gilman, M. S. Osburne, J. Clardy, J. Handelsman, and R. M. Goodman.** 2000. Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. Appl. Environ. Microbiol. **66:**2541–2547.
26. **Sambrook, J., and D. W. Russell.** 2002. Molecular cloning: a laboratory manual, 3rd ed., p. 11.8, 11.26–11.35. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.
27. **Sarkar, N.** 1997. Polyadenylation of mRNA in prokaryotes. Annu. Rev. Biochem. **66:**173–197.
28. **Ward, D. M., R. Weller, and M. M. Bateson.** 1990. 16S rRNA sequences reveal numerous uncultured microorganisms in a natural community. Nature **345:**63–65.
29. **Wheeler, D. L., C. Chappey, A. E. Lash, D. D. Leipe, T. L. Madden, G. D. Schuler, T. A. Tatusova, and B. A. Rapp.** 2000. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. **28:**10–14.