Sequence and Expression Analyses of Cytophaga-Like Hydrolases in a Western Arctic Metagenomic Library and the Sargasso Sea†

Matthew T. Cottrell, Lifying Yu, and David L. Kirchman*

University of Delaware, College of Marine Studies, Lewes, Delaware 19958

Received 2 May 2005/Accepted 9 September 2005

Sequence analysis of environmental DNA promises to provide new insights into the ecology and biogeochemistry of uncultured marine microbes. In this study we used the Sargasso Sea Whole Genome Sequence (WGS) data set to search for hydrolases used by Cytophaga-like bacteria to degrade biopolymers such as polysaccharides and proteins. Analysis of the Sargasso WGS data for contigs bearing both the 16S rRNA genes of Cytophaga-like bacteria and hydrolase genes revealed a cellulase gene (celM) most similar to the gene found in Cytophaga hutchinsonii. A BLAST search of the entire Sargasso Sea WGS data set indicated that celM was the most abundant cellulase-like gene in the Sargasso Sea. However, the similarity between CelM-like cellulases and peptidases belonging to metalloprotease family M42 led us to question whether CelM is involved in the degradation of polysaccharides or proteins. PCR primers were designed for the celM genes in the Sargasso Sea WGS data set and used to identify celM in a fosmid library constructed with prokaryotic DNA from the western Arctic Ocean. Expression analysis of the Cytophaga-like Arctic CelM, which is 63% identical and 77% similar to CelM in C. hutchinsonii, indicated that there was peptidase activity, whereas cellulase activity was not detected. Our analysis suggests that the celM gene plays a role in the degradation of protein by Cytophaga-like bacteria. The abundance of peptidase genes in the Cytophaga-like fosmid clone provides further evidence for the importance of Cytophaga-like bacteria in the degradation of protein in high-molecular-weight dissolved organic matter.

The hydrolysis of biopolymers is potentially a rate-limiting step in the degradation of high-molecular-weight dissolved organic matter (DOM) by microbial communities in the oceans (2). The consumption of high-molecular-weight DOM requires a hydrolysis step because only compounds that have molecular masses that are less than about 500 Da can be transported across the bacterial membrane (33). Despite the requirement for hydrolysis, high-molecular-weight DOM is highly bioactive (1) and plays an important role in bacterial metabolism in the ocean. Polysaccharides and proteins are the most abundant known constituents of high-molecular-weight DOM (4) and are readily utilized by bacteria. In the Sargasso Sea, for example, protein supports about one-half of the bacterial nitrogen requirement (20), and polysaccharides, such as chitin, can support as much as 10% of the bacterial production in estuaries (22).

Cytophaga-like bacteria are hypothesized to be important in the hydrolysis and mineralization of biopolymers in the oceans. Cultured isolates of Cytophaga-like bacteria are proficient in degrading carbohydrate biopolymers, such as cellulose and chitin (21), which are constituents of high-molecular-weight DOM (3). Efficient utilization of biopolymers in high-molecular-weight DOM and in detritus particles might explain the high levels of free-living Cytophaga-like bacteria and their especially high levels on detritus particles in the ocean (12–14, 29). The hypothesis that high-molecular-weight DOM is consumed by Cytophaga-like bacteria is supported by radiotracer studies that examined the consumption of protein and chitin by uncultured bacteria (8). Finally, Cytophaga-like bacteria grow more rapidly in seawater incubation mixtures supplemented with concentrated high-molecular-weight DOM (11). Examining hydrolase genes in environmental DNA could be another approach to link uncultured bacterial biopolymer degradation and other biogeochemical processes.

In this study we examined genes in the Sargasso Sea Whole Genome Sequence (WGS) data set (32) encoding hydrolases potentially used by marine Cytophaga-like bacteria for degrading biopolymers in high-molecular-weight DOM. Our focus on Cytophaga-like bacteria was motivated by the desire to understand the role of these bacteria in the utilization of high-molecular-weight DOM. PCR primers were designed for the most abundant type of endoglucanase identified in the Sargasso data set and used to screen a fosmid library constructed with prokaryotic DNA from the western Arctic Ocean. The cloned hydrolase was expressed in Escherichia coli and assayed for various hydrolase activities because gene function could not be determined with confidence using amino acid similarity alone due to the similarity of the hydrolase to both cellulases and peptidases. Our results highlight the value of experimental data that support actual gene function. The complete sequence of the fosmid bearing Arctic Cytophaga-like DNA supports the hypothesis that Cytophaga-like bacteria are especially adapted to utilizing biopolymers in high-molecular-weight DOM.

MATERIALS AND METHODS

Contigs in the Sargasso WGS database (32) were screened for Cytophaga-like 16S rRNA genes using BLASTN and a data set for bacterial 16S rRNA genes (http://rdp.cme.msu.edu/download/philrdp.16S..arb) (19). Open reading frames of
the Cytophaga-like contigs were then examined for endogluccanase genes using BLASTP and the GenBank database (http://www.ncbi.nlm.nih.gov/). This analysis identified a celM-like endoglucanase gene most similar (63%) to that found in Cytophaga hutchinsonii. The C. hutchinsonii gene was then used in a second iteration of BLASTP to find similar genes in the Sargasso WGS data set.

**fosmid library screening.** The Arctic fosmid library was constructed using DNA isolated from cells in the <0.8-μm size fraction of a seawater sample collected at a depth of 10 m in the Chukchi Sea (72°19.33′N, 151°59.07′W) in July 2004. For DNA isolation and library construction we used procedures described previously (10), except that the bacterial biomass was collected by vacuum filtration rather than by tangential flow filtration. Pools of 96 fosmid clones were screened for 16S rRNA genes by denaturing gradient electrophoresis (DGGE) of PCR amplicons generated with primers GC358F and 517R (28). Selected bands resolved on an 8% polyacrylamide gel containing a 25% to 55% denaturant gradient (13.6% to 22% formamide and 10.5% to 23% urea) were reamplified and sequenced. Phylogenetic classification was performed using BLAST and the ARB sequence analysis tool (26).

PCR primers were designed to amplify the celM-like genes identified in the Sargasso WGS data set using the tool for identifying consensus primers in the Oligo primer analysis software package (Molecular Biology Insights, Inc.). The celM gene primers CelM298F (5′-GGCTTCTCTCAAAATGCGG-3′) and CelM559R (5′-AAC CTCAGCAATCTAAATACCC-3′) were selected based on successful in silico amplification of four of the most complete (>990-bp) Sargasso celM-like genes. These primers were then used to screen the Arctic fosmid library. Pools of 96 fosmid clones were screened by performing a celM gene PCR with primers at a concentration of 200 nM in a buffer containing 0.7 mM MgCl₂ provided with the Taq polymerase (Promega). The thermal cycling conditions consisted of 35 cycles of denaturation at 96°C, primer annealing at 50°C and DNA polymerization at 72°C. The celM-bearing clones in PCR-positive plate pools were identified by further PCR screening of the 96 pooled clone rows and columns of the 96-well plate.

The celM-bearing Arctic clone was completely sequenced by the Joint Genome Institute. The sequence was analyzed using the annotation tools available in the Artemis DNA sequence viewer obtained from the Welcome Trust Sanger Institute (http://www.sanger.ac.uk/Software/Artemis/) and using the FGENSEB website (http://www.softberry.ru/) (Softberry, Inc.).

**Phylogenetic analysis.** 16S rRNA gene sequences were aligned using the ARB fast aligner (26), and celM gene sequences were aligned using a CelM amino acid sequence alignment generated using ClustalW (6). Distance matrices and neighbor-joining trees were constructed using PHYLIP (15).

**Expression analysis of celM.** The celM gene of the Arctic fosmid clone (Arctic CelM) was subcloned into an expression vector for activity analyses. The entire celM gene of the Arctic CelM clone was amplified using PCR primers CelM516U (5′-CACCATGGCAACAAAAAAAATACTT-3′) and CelM1648L (5′-AGA TGTTTCTACTTATCTAAACCAGA-3′) and was cloned into the PET101 expression vector and E. coli TOP10 (Invitrogen) by following the procedure provided by the manufacturer. The celM PET101 construct was subsequently transformed into E. coli BL21(DE3) for expression analysis. In order to examine the size and amino acid sequence of Arctic CelM, overnight cultures (500 ml) of E. coli bearing the celM PET101 construct or the pET101 vector alone were pelleted by centrifugation and washed with phosphate-buffered saline (PBS) (8 g of NaCl per liter, 0.2 g of KCl per liter, 1.44 g of Na₂HPO₄ per liter, 0.24g of KH₂PO₄ per liter; pH adjusted to 7.4). The pellet was resuspended in 5 ml of PBS, and the cells were lysed using sonication. Cell debris was removed by centrifugation at 100,000 × g for 60 min at 4°C. The molecular weight of the expressed protein was estimated by sodium dodecyl sulfate-polyacrylamide gel electrophoresis and compared to that expected from the 1,094-bp celM gene. Amino acid sequencing by liquid chromatography-mass spectrometry of the expressed protein was performed by the Campus Chemical Instrument Center, Ohio State University, using a procedure described by Hanson and Tabita (18). The Arctic CelM was digested with trypsin, and the sequences of 21 peptides covering 63% of the original protein were then matched to the sequence of Arctic CelM.

Cell lysates were assayed for glutamyl aminopeptidase activity using glu- tamine-p-nitroanilide (Glu-pNA) (Peptide Institute, Inc.) as described by L’Anson et al. (25). Twenty microliters of 5 mM Glu-pNA was added to 180 μl of lysate and incubated for 18 h at 37°C or 60°C (cloned Cytophaga hutchinsonii CelM). The reaction was stopped by addition of 100 μl of 30% (vol/vol) acetic acid. The sample was then centrifuged at 10,000 × g for 5 min, and the absorbance at 410 nm was determined.

Hydrolysis of carboxymethylcellulose (CMC) was assayed by monitoring the production of reducing sugars using the procedure outlined by Wu et al. (34). The reaction mixture contained 20 μl of 1% CMC (Polysciences, Inc.) and 180 μl of lyses and was incubated at 37°C for 18 h. The reaction was terminated by addition of 0.2 ml of 2% sodium carbonate and 1 ml of cyanide-carbonate solution (10 mM KCN; 50 mM Na₂CO₃). Two milliliters of 0.05% potassium ferricyanide was added, and the solution was vortexed and boiled for 30 min. The tubes were cooled, and the absorbance at 420 nm was determined.

Cellulase activity was assayed by measuring the hydrolysis of fluorogenic and chromogenic substrates analogs, including 4-methylumbelliferyl-β-D-gluco- pyranoside, 4-nitrophenyl-β-D-cellobioside, and 4-nitrophenyl-β-D-cellottetra- side (Sigma-Aldrich). The reaction mixture contained 20 μl of 5 mM substrate and 180 μl of lyses. The reaction mixtures were incubated for 18 h at 37°C or 60°C (cloned C. thermocellum CelM) and terminated by addition of glycine carbonate buffer (pH 9.7). The sample was then centrifuged at 10,000 × g for 5 min, and the absorbance at 410 nm was determined (nitrophenyl-linked sub- strate) or the preparation was assayed for blue fluorescence under UV light (methylumbelliferyl-linked substrate).

Protease activity was measured using fluorescein isothiocyanate-conjugated bovine albumin. The reaction mixture contained 20 μl of a 1-mg/ml labeled albumin solution and 180 μl of lyses. The mixture was incubated at 37°C for 18 h, and the reaction was terminated by addition of 5% trichloroacetic acid. The sample was centrifuged at 10,000 × g for 15 min, and the green fluorescence fluorescence in the supernatant was assayed under UV light. Controls were treated with proteinase K.

**Nucleotide sequence accession number.** The nucleotide sequence of the Arctic fosmid was deposited in GenBank under accession number DQ272742.

**RESULTS AND DISCUSSION**

Cytophaga-like contigs in the Sargasso Sea WGS data set. BLASTN analysis of the Sargasso WGS data set using a database of 16S rRNA genes revealed 26 contigs having 16S rRNA genes from Cytophaga-like bacteria. In contrast, in the original annotation of the Sargasso WGS data set only 11 contigs were identified as having Cytophaga-like 16S rRNA genes (32), which represents less than 5% of the prokaryotes in the Sargasso WGS data set. Such a low abundance of Cytophaga-like bacteria is somewhat surprising for a marine system because Cytophaga-like bacteria are one of the abundant groups of bacteria in the ocean (21). Our BLASTN analysis indicates that the abundance of Cytophaga-like bacteria in the Sargasso WGS data set might be closer to 15%. The contribution of this bacterial group to bacterial communities has been underestimated by studies using clone libraries of 16S rRNA genes (7). Although problems with the general PCR primers that might miss Cytophaga-like bacteria would not affect the shotgun sequen- cing approach, DNA extraction efficiency could be a factor. Our results suggest that 16S rRNA genes of other bacterial groups may have been missed in the original annotation as well.

The Cytophaga-like bacteria identified in the Sargasso WGS data set were related to other Cytophaga-like bacteria seen previously in marine environments, based on 16S rRNA gene phylogeny. The Sargasso Cytophaga-like sequences were grouped into nine clusters of marine Cytophaga-like bacteria (Fig. 1), although one Sargasso gene (contig AAC01458001) remained separate from the other Cytophaga-like bacterial sequences (Fig. 1). Because the shotgun sequencing approach has no PCR step, we might have expected the Sargasso WGS data set to include completely new types of Cytophaga-like bacteria not seen previously in PCR-based studies. However, the Cyto- phaga-like bacteria in the Sargasso study were closely related to bacteria already identified in PCR clone libraries.

Cytophaga-like hydrolases in the Sargasso WGS data set. BLASTP analysis identified two Cytophaga-like contigs bearing genes encoding hydrolases most similar to the celM gene in C. hutchinsonii. Contig AAC01014206 encodes a protein that
FIG. 1. Phylogenetic relationships of *Cytophaga*-like bacterial 16S rRNA genes in the Sargasso Sea WGS data set (Sargasso Sea) and in a fosmid clone library of Arctic environmental DNA (Arctic Ocean). Clusters 1 and 3 are expanded to reveal the Sargasso contigs and Arctic fosmid having both 16S rRNA and *celM* genes. The other *Cytophaga*-like clusters contained one to three Sargasso *Cytophaga*-like 16S rRNA genes and *Cytophaga*-like genes of uncultured bacteria from various marine environments.
is 63% identical and 77% similar to *C. hutchinsonii* CelM (Table 1). This protein is also 33% to 39% identical to the CelM in *Geobacter metallireducens*, a putative hydrolase in *Geobacter sulfurreducens*, and endoglucanases in three bacteria. The protein encoded by this *Cytophaga*-like hydrolase gene is also similar to peptidases in three species of *Bacillus* (Table 1).

A protein encoded on a second *Cytophaga*-like contig (AACY01060866) also is similar to CelM. BLASTP analysis revealed that the hydrolase encoded on the second contig is 53% identical and 79% similar to CelM in *C. hutchinsonii*. In contrast, there is no similarity between the hydrolase encoded on this *Cytophaga*-like contig and peptidases or proteases.

**Abundance of CelM in the Sargasso WGS.** CelM is the most abundant cellulase-like hydrolase in the Sargasso data set (Fig. 2) based on a BLASTP analysis using 113 cellulases (EC 3.2.1.4) available from the Swiss-Prot database (March 2005). CelM occurs 30 times in the Sargasso data set, outnumbering all other cellulases classified in glycosyl hydrolase families, according to the nomenclature used by the Carbohydrate-Active Enzymes (CAZY) database (http://afmb.cnrs-mrs.fr/CAZY/) (Fig. 2). Family 5 glycosyl hydrolase is the second most abundant type of cellulase and occurs on 27 contigs. Cellulases in families 8, 9, 10, and 12 occur on eight or fewer contigs. No contig contained more than one cellulase.

**Phylogeny of celM.** The celM gene has been annotated in a broad range of *Archaea* and *Bacteria*, including enteric bacteria, methanogens, thermophiles, and high-G+C-content gram-positive bacteria, as well as *Cytophaga*-like bacteria. Phylogenetic analysis suggested that some of the genes are true orthologs of celM, but others may be distantly related genes that have been mistaken for celM. There does not seem to be a simple relationship between these putative celM genes and bacterial phylogeny. For example, celM genes of *Archaea* and *Bacteria* are not clearly separated from each other (Fig. 3).

Some bacteria apparently possess several copies of genes annotated as celM, which complicates the picture of the relationship between the celM gene tree and bacterial phylogeny. For example, *C. hutchinsonii* has three copies of the celM gene. The levels of amino acid identity for proteins encoded by *C. hutchinsonii* genes range from 20% to 30%, which are comparable to the levels of similarity among celM genes in different bacteria. As a result, the *C. hutchinsonii* genes appear in different parts of the celM tree (Fig. 3). Similarly, celM genes in different strains of *E. coli* do not group together, and the four *Clostridium* celM genes appear in different parts of the celM tree, although they are from different *Clostridium* species. One possibility is that bacteria received celM genes with different evolutionary histories through lateral gene transfer. On the other hand, closer scrutiny may reveal that misidentification of distantly related genes is responsible for the observed difference between the bacterial phylogeny and the celM gene tree.

Despite possible problems with the deep branches in the celM gene phylogeny, the presence of closely related clusters of celM genes does provide some insight into the types of bacteria in the Sargasso Sea that have celM. Although celM genes from related bacteria, such as *Clostridium* sp. strains, do not always group together, when celM genes do group together, they come

---

**TABLE 1. BLASTP analysis of a Cytophaga-like celM gene in the Sargasso WGS data set (contig AACY01014206)**

<table>
<thead>
<tr>
<th>Accession no.</th>
<th>Description</th>
<th>Organism</th>
<th>% Identity</th>
<th>% Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>ZP_00309918</td>
<td>Cellulase M and related proteins (COG1363)</td>
<td><em>Cytophaga hutchinsonii</em></td>
<td>63</td>
<td>77</td>
</tr>
<tr>
<td>ZP_00301131</td>
<td>Cellulase M and related proteins (COG1363)</td>
<td><em>Geobacter metallireducens</em></td>
<td>38</td>
<td>58</td>
</tr>
<tr>
<td>NP_953245</td>
<td>Hydrolase, putative</td>
<td><em>Geobacter sulfurreducens</em></td>
<td>39</td>
<td>58</td>
</tr>
<tr>
<td>AAV45786</td>
<td>Endoglucanase</td>
<td><em>Halocella marismortui</em></td>
<td>36</td>
<td>55</td>
</tr>
<tr>
<td>YP_176856</td>
<td>Endoglucanase</td>
<td><em>Bacillus clausii</em></td>
<td>36</td>
<td>53</td>
</tr>
<tr>
<td>NP_693252</td>
<td>Endoglucanase</td>
<td><em>Oceanobacillus iheyensis</em></td>
<td>34</td>
<td>53</td>
</tr>
<tr>
<td>ZP_00185964</td>
<td>Cellulase M and related proteins (COG1363)</td>
<td><em>Rubrobacter xylanophilus</em></td>
<td>31</td>
<td>51</td>
</tr>
<tr>
<td>NP_618720</td>
<td>Cellulase</td>
<td><em>Methanosarcina acetivorans</em></td>
<td>32</td>
<td>54</td>
</tr>
<tr>
<td>NP_632697</td>
<td>Hypothetical protein MM0673</td>
<td><em>Methanosarcina mazei</em></td>
<td>32</td>
<td>53</td>
</tr>
<tr>
<td>NP_868423</td>
<td>Probable endoglucanase</td>
<td><em>Rhodopirellula baltica</em></td>
<td>31</td>
<td>52</td>
</tr>
<tr>
<td>ZP_00149447</td>
<td>Cellulase M and related proteins (COG1363)</td>
<td><em>Methanococcoides</em></td>
<td>34</td>
<td>52</td>
</tr>
<tr>
<td>YP_083617</td>
<td>Glucanase/aminopeptidase</td>
<td><em>Bacillus thuringiensis</em></td>
<td>33</td>
<td>55</td>
</tr>
<tr>
<td>NP_834277</td>
<td>Aminopeptidase</td>
<td><em>Bacillus cereus</em></td>
<td>33</td>
<td>55</td>
</tr>
<tr>
<td>YP_021461</td>
<td>Peptidase, M42 family</td>
<td><em>Bacillus anthracis</em></td>
<td>32</td>
<td>55</td>
</tr>
<tr>
<td>YP_085892</td>
<td>Glucanase/aminopeptidase</td>
<td><em>Bacillus cereus</em></td>
<td>32</td>
<td>55</td>
</tr>
<tr>
<td>NP_693064</td>
<td>Endo-1,4-beta-glucanase</td>
<td><em>Oceanobacillus iheyensis</em></td>
<td>33</td>
<td>55</td>
</tr>
</tbody>
</table>

* The 16 genes in the GenBank database with the highest levels of similarity to the Sargasso celM gene are listed. The levels of identity and similarity are the levels of identity and similarity between the amino acid sequence of the Sargasso CelM and the amino acid sequences of the homologous gene products in the organisms listed.
from related bacteria. For example, all 14 Firmicutes celM genes belong to a single clade, as do the six Pyrococcus celM genes (Fig. 3). Similarly, all 15 of the Sargasso celM genes cluster together with celM from C. hutchinsonii (Fig. 3), suggesting that all of these genes are associated with Cytophaga-like bacteria. Although most of the celM genes in the Sargasso data are not linked to 16S rRNA genes, the phylogenetic analysis suggests that despite the broad diversity of bacteria that potentially have this gene, celM seems to be restricted to Cytophaga-like bacteria in the Sargasso Sea.

**Gene content of an Arctic fosmid bearing celM.** To test the hypothesis that Cytophaga-like bacteria having celM also possess genes encoding other hydrolases, we examined a celM-containing genome fragment from an uncultivated Cytophaga-like bacterium. This analysis was performed using a fosmid library of prokaryotic DNA collected from the Arctic Ocean. The Arctic fosmid library had 4,800 clones (average insert size, 40 kb) that were largely derived from prokaryotic DNA. Twenty-seven clones carried 16S rRNA genes, as determined by DGGE screening. BLASTN analysis of the DGGE band sequences revealed that 23 clones contained prokaryotic DNA and four clones carried DNA from the photosynthetic picoeukaryote Mantoniella squamata. The clones carrying prokaryotic DNA included 10 clones bearing Cytophaga-like 16S rRNA and four clones identified as clones carrying DNA from Gammaproteobacteria. The remaining clones with 16S rRNA genes belonged to the Alphaproteobacteria, Betaproteobacteria, and Actinobacteria (three each). No archaeal clones were detected with DGGE primers for this group.

One clone in the Arctic fosmid library (Arctic CelM) was positive as determined by the celM gene PCR. Similar to our results with the Sargasso data set, the Arctic CelM clone also carried a Cytophaga-like 16S rRNA gene detected in the DGGE analysis. A sequence analysis extending upstream and downstream of the celM gene PCR priming sites revealed a 1,089-bp open reading frame. A BLASTP analysis indicated that this open reading frame encodes a protein that is 63% identical and 77% similar to CelM in C. hutchinsonii. The amino acid sequence of Arctic CelM does not appear to have a signal peptide found in the secreted proteins examined so far, based on an analysis using SignalP (http://www.cbs.dtu.dk/services/SignalP/).

Sequence analysis of the Arctic CelM fosmid clone revealed 28 genes, including 25 open reading frames, 16S and 23S rRNA genes, and a 16S-23S rRNA internal transcribed spacer (Table 2; see Table S1 in the supplemental material). Phylogenetic analysis placed the 16S rRNA gene in marine Cytophaga-like clade 1, which also has genes from the Sargasso Sea and other marine environments (Fig. 1). Many of the protein-encoding genes were also most similar to genes in Cytophaga-
like bacteria, which further indicates that the Arctic CelM fosmid clone contains DNA from a Cytophaga-like bacterium. Sixty percent of the genes on the Arctic fosmid were most similar to genes in *Cytophaga*-like bacteria. Seven (30%) of the fosmid genes were most similar to genes in *C. hutchinsonii*. The celM gene exhibited the highest level of similarity (65% identity) to a gene in *C. hutchinsonii*. Other genes similar to genes in *C. hutchinsonii* include genes encoding acyl coenzyme A synthetase (COG1363), dihydroorotate dehydrogenase (COG0167), and a hypothetical protein, which were 62 to 78% identical to genes in the Arctic fosmid (see Table S1 in the supplemental material). Two genes were 50% identical to peptidyl-prolyl cis-trans isomerase and glycerol kinase genes.

Several of the genes in the Arctic CelM fosmid clone appear to encode hydrolases. In fact, two genes encoding peptidases, an aminopeptidase gene (gene 7) and a zinc carboxypeptidase gene (gene 24), were identified in addition to the celM gene (gene 16) (Table 2). Such a combination of proteolytic and glycolytic potential is consistent with the hypothesis that Cytophaga-like bacteria are well adapted to using biopolymers such as those in high-molecular-weight DOM. Genes involved in organic nutrient uptake were also well represented. The Arctic fosmid had genes encoding a glycerol uptake facilitator protein (gene 3) and an ABC-type transporter (gene 21) that is potentially involved in the uptake of peptides (see Table S1 in the supplemental material).

**Peptidase activity of the Cytophaga-like celM gene.** Most studies of genes cloned from environmental DNA have relied on sequence analysis to infer what metabolic function the genes might mediate. Function is commonly assigned according to the highest level of similarity in a BLASTN analysis using a large database, such as the GenBank database. Although this approach is easy and widely used, it can be misleading because there is often little or no experimental evidence for the function of many, if not all, of the genes identified by the search. This was the case in our analysis of celM. The only experimental evidence for endoglucanase activity of CelM is evidence for the celM gene in *C. thermocellum* (24). The Arctic CelM was 29% identical and 48% similar to the *C. thermocellum* CelM (Table 1). Other genes similar to the Arctic celM gene were genes encoding family M42 peptidases. The peptidase most similar to the Arctic CelM was a protein in *Bacillus cereus*, which was 32% identical and 55% similar to the Arctic CelM.

To clarify the true identity of Arctic CelM, PCR primers were designed to amplify the entire gene and then were used to subclone the celM gene into an expression vector. The molecular mass of the protein expressed from the Arctic CelM clone determined by polycrylamide gel electrophoresis was 34 kDa, which was consistent with the size of the Arctic celM open reading frame (1,089 bp). The amino acid sequence of the expressed protein determined by liquid chromatography-mass spectrometry (21 peptide fragments and 63% coverage) was identical to that of the expected protein based on the celM sequence in the Arctic CelM fosmid clone.

CelM encoded in the Arctic fosmid tested positive for glutamyl aminopeptidase activity when it was assayed with Glu-pNA (Fig. 4). The peptidase activity of the Arctic CelM in

![FIG. 4. Aminopeptidase activity of Arctic CelM, Clostridium CelM, and pET101 control assayed with and without addition of Zn using glutamine-ω-nitroanilide. The error bars indicate standard errors.](http://aem.asm.org/)

**TABLE 2. Analysis of select open reading frames and rRNA genes of the Arctic CelM fosmid clone**

<table>
<thead>
<tr>
<th>Open reading frame</th>
<th>Location</th>
<th>Length (bp)</th>
<th>Accession no.</th>
<th>Gene function</th>
<th>Organism</th>
<th>% Identity</th>
<th>% Similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>c(3976)</td>
<td>735</td>
<td>AAD36499</td>
<td>Glycerol uptake facilitator protein</td>
<td><em>Thermotoga maritima</em></td>
<td>47</td>
<td>66</td>
</tr>
<tr>
<td>7</td>
<td>c(10446)</td>
<td>1,233</td>
<td>CAD78184</td>
<td>Probable aminopeptidase</td>
<td><em>Pseudomonas sp.</em></td>
<td>38</td>
<td>52</td>
</tr>
<tr>
<td>8</td>
<td>10478</td>
<td>1,065</td>
<td>CAE79899</td>
<td>Probable aminopeptidase</td>
<td><em>Brevibacillus bacteriovorus</em></td>
<td>36</td>
<td>51</td>
</tr>
<tr>
<td>13</td>
<td>c(16837)</td>
<td>1,858</td>
<td>AY03305</td>
<td>23S rRNA</td>
<td>Uncultured Cytophagaales</td>
<td>98</td>
<td>NA</td>
</tr>
<tr>
<td>14</td>
<td>c(17336)</td>
<td>500</td>
<td>AY03305</td>
<td>16S-23S rRNA internal transcribed spacer</td>
<td>Uncultured Cytophagaales</td>
<td>93</td>
<td>NA</td>
</tr>
<tr>
<td>15</td>
<td>c(18840)</td>
<td>1,505</td>
<td>AY03305</td>
<td>16S rRNA</td>
<td>Uncultured Cytophagaales</td>
<td>97</td>
<td>NA</td>
</tr>
<tr>
<td>16</td>
<td>19781</td>
<td>1,092</td>
<td>ZP_00309918</td>
<td>Cellulase M and related proteins (COG1363)</td>
<td><em>Cytophagacella tetani</em></td>
<td>63</td>
<td>77</td>
</tr>
<tr>
<td>21</td>
<td>25715</td>
<td>1,758</td>
<td>AAM24093</td>
<td>ABC-type multidrug/protein/lipid transport system ATPase</td>
<td><em>Thermoanaerobacter tengcongensis</em></td>
<td>37</td>
<td>55</td>
</tr>
<tr>
<td>24</td>
<td>27957</td>
<td>2,508</td>
<td>AAO35142</td>
<td>Zinc carboxypeptidase</td>
<td><em>Clostridium tetani</em></td>
<td>25</td>
<td>41</td>
</tr>
</tbody>
</table>

* Gene functions were assigned based on the highest levels of similarity in BLASTP analyses (GenBank database, March 2005). rRNA genes were identified using BLASTN. The levels of identity and similarity are the levels of identity and similarity between the amino acid sequences encoded by the open reading frames and the amino acid sequences of the homologous gene products in the organisms listed. The level of nucleotide identity is indicated for rRNA genes. The location indicates where the gene starts; c indicates genes encoded on the complementary strand. For a complete list of all genes in the fosmid clone see Table S1 in the supplemental material. **NA,** not applicable to rRNA genes.
crude E. coli extracts was much higher than the activity in extracts of E. coli expressing the C. thermocellum CelM; the activity of the C. thermocellum CelM was not different from that of the negative control, the expression vector alone (Fig. 4). Because of the amino acid similarity of the Arctic CelM and the C. thermocellum CelM to zinc metalloproteases, we expected that the peptidease activity of the Arctic CelM would be higher with the addition of Zn (30). In fact, Zn addition actually reduced the activity of the Arctic CelM and had no effect on the activity of the C. thermocellum CelM (Fig. 4). The absence of peptidease activity in the C. thermocellum CelM is noteworthy because based solely on amino acid similarity one might predict that this protein would have peptidease activity due to the presence of a putative peptidase M42 domain (PFAM 05343).

No cellulase activity was detected in Arctic CelM extracts assayed with CMC, 4-methylumbelliferyl-β-D-glucopyranoside, 4-nitrophenyl-β-D-cellobioside, and 4-nitrophenyl-β-D-cellotetraoside. As we expected, the cloned C. thermocellum CelM tested positive for CMC hydrolysis but did not hydrolyze any of the cellulose analogs.

**Conclusions.** Sequence analysis of environmental DNA from uncultivated microbes is a potentially powerful tool for uncovering the metabolic capabilities of microbes in the environment and broadening our view of microbial ecology. However, gene sequences must be translated into biological functions, and this is a limiting step in using environmental sequence data to address ecological questions. This study exposed the shortcomings of assigning gene function based on sequence similarity alone. The difficulty in distinguishing CelM from the M42 family of peptidases based on sequence similarity was overcome only by determining the activity of the expressed protein. While expression analysis is labor-intensive, it extracts the most complete and accurate information from environmental sequence data.

Functional assays circumvent the problem of assigning biological functions to genes recovered from metagenomic libraries, but this approach has its shortcomings. Sophisticated approaches based on complementation of genes in mutant E. coli host cells have proven to be useful in detecting genes that are active in the fermentation of glycerol (23). In contrast, rather simple assays using selective media (31) and looking for distinctive coloration of colonies (17) have been effective for detecting antibiotic resistance genes. Similarly straightforward assays are available to detect genes encoding hydrolases, such as xylanase (5) and chitinase (9). However, the effectiveness of any functional assay depends on successful expression in E. coli or alternative hosts, such as those proposed for studying uncultured soil bacteria (16, 27).

The combination of sequence and expression analyses should allow us to start to examine the role of CelM in biopolymer degradation by *Cytophaga*-like bacteria, although we need more information about signal peptides used by these bacteria and other marine microbes. If CelM is exported using a signal peptide unlike those studied to date, then it may liberate low-molecular-weight peptides that either are transported directly into the cell or are hydrolyzed further to even smaller peptides or free amino acids. However, if CelM remains in the cell, it may play a role analogous to that of the PepA glutamyl aminopeptidase in *Lactococcus lactis* (25), which acts on intracellular peptides. Regardless, CelM is probably one part of a multicomponent system used by *Cytophaga*-like bacteria for the consumption of biopolymers in high-molecular-weight DOM and particulate detritus. CelM and other peptidases probably work together with proteases and peptide transporters in the degradation of protein-containing DOM by *Cytophaga*-like bacteria. Characterizing the genes encoding components of biopolymer utilization systems should be useful in linking uncultured *Cytophaga*-like bacteria and other bacterial groups to biopolymer degradation and DOM cycling.

**ACKNOWLEDGMENTS**

This study was supported by grants from the U.S. Department of Energy (DE-FG02-01ER63142) and the National Science Foundation (OFP 0124733). The sequencing and assembly of the fosmid clone was performed by the production sequencing group at the DOE/Joint Genome Institute through the Sequence-For-Others Program under the auspices of the U.S. Department of Energy’s Office of Science, Biological and Environmental Research Program and the University of California Lawrence Livermore National Laboratory.

We thank Rex Malmstrom for his assistance and the Chief Scientists of the Shelf Basin Interaction project, Jackie Grebmeier and Lee Cooper, for their support during sample collection in the western Arctic Ocean. David Wilson kindly provided the clone of *CelM* from *C. thermocellum*.

**REFERENCES**


