Comparative analysis of a genome fragment of an uncultivated mesopelagic crenarchaeote reveals multiple horizontal gene transfers

Purificación López-García,1*† Céline Brochier,2*† David Moreira,1 and Francisco Rodríguez-Valera3
1UMR CNRS 8079, Ecologie, Systématique et Evolution, Université Paris-Sud, bâtiment 360, 91405 Orsay Cedex, France.
2Atelier de Bioinformatique, Université Pierre et Marie Curie, 16 rue Cuvier, 75005 Paris, France.
3División de Microbiología, Universidad Miguel Hernández, Campus de San Juan, 03550 San Juan de Alicante, Spain.

Summary

Marine planktonic crenarchaeota have escaped all cultivation attempts to date, all crenarchaeota growing in pure culture so far being hyperthermophiles. Here, we present a comparative genomic analysis of a 16S- plus 23S-rDNA-containing fragment of a crenarchaeote retrieved from an environmental genomic library constructed from picoplankton collected at 500 m depth in the Antarctic Polar Front. The clone DeepAnt-EC39 contained an insert of 33.3 kbp, which was completely sequenced. DeepAnt-EC39 appears to represent a lineage specific to deep-sea waters but widespread geographically, as revealed by the analysis of the 16S-23S-rDNA intergenic spacer region. A comparison with previously sequenced marine crenarchaeotal genomic clones also containing an rrn operon (74A4, 4B7 and Cenarchaeum symbiosum strains A and B) revealed a highly variable structure involving gene rearrangements and insertions/deletions. The surroundings of the rrn operon and the contiguous glutamate-1-semialdehyde aminotransferase gene appear hot spots for recombination. Phylogenetic analyses of all individual predicted proteins revealed the existence of several likely cases of horizontal gene transfer both, between the two archaeal kingdoms and between the two prokaryotic domains. The most frequent horizontal transfers appear to involve genes from mesophilic methanogenic euryarchaeota related to Methanosarcinales. We hypothesise that the acquisition of genes from mesophilic bacteria and euryarchaeota has played a major role in the adaptation of Group I crenarchaeota to life at lower temperatures.

Introduction

The use of molecular methods based on the analysis of 16S rRNA genes present in the environment was a major breakthrough in studying prokaryotic diversity in natural populations (Olsen et al., 1986; Giovannoni et al., 1990; Pace, 1997; Hugenholtz et al., 1998). In the past 15 years, we have come to realize that the microbial world is much more diverse and complex than ever suspected, and thousands of environmental 16S rRNA sequences are accumulating in databases. However, the information contained in 16S rRNA genes (basically phylogenetic) is limited, and it does not generally allow making inferences about the physiology or the mode of life of the respective microorganisms. This is particularly so in the case of phylogenetically, very divergent lineages having escaped all cultivation attempts. Recently, another molecular-based source of information for environmental lineages has raised new hopes among microbiologists. This is the so-called environmental genomic, eco-genomic or metagenomic approach (Stein et al., 1996; Rondon et al., 1999; Béjà et al., 2000a; DeLong, 2001; Rodríguez-Valera, 2002). Cloning and sequencing large environmental DNA fragments in BACs, fosmids or cosmids, followed by the screening of clones to search for a particular marker, for instance, the presence of 16S rRNA genes, would provide additional knowledge about uncultivated lineages.

Environmental genomics thus facilitates the access to information covering a wide range of aspects. These go from basic studies on genomic organization (gene density, presence of operons or clusters, genomic rearrangements, etc.) to purely applied objectives, such as the identification of enzymatic activities with biotechnological potential or the detection of antibacterial or antitumoural agents (Rondon et al., 1999; Osburne et al., 2000). It can also give clues about the metabolism and physiology of uncultured microorganisms. A paradigmatic example is the discovery of a previously unknown type of phototrophy in members of marine bacterioplankton (Béjà et al.,...
Environmental genomics can also provide phylogenetic information additional or alternative to that contained in 16/18S rRNA genes via the identification of conserved protein genes, such as the EF-2 (Stein et al., 1996). It can also allow the detection of divergent rRNA genes that would be easily missed by PCR amplification approaches (López-García et al., 2002). The phylogenomic study of environmental clones, that is, the comparison of gene arrangement coupled to phylogenetic analysis of individual genes, may potentially lead to the identification of horizontal gene transfers and, eventually, the direction in which they occurred (Brochier et al., 2000). Finally, the analysis of 16S rRNA genes in environmental genomic libraries can subsidiarily complement diversity studies, as this approach is not biased by PCR amplification.

Marine picoplankton was the preferred choice for the first 16S rRNA diversity studies (Stahl et al., 1984; Giovannoni et al., 1990; DeLong, 1992), and it has also so for pioneered environmental genomic studies (Stein et al., 1996; Schleper et al., 1998; Béjà et al., 2000a, b, 2001, 2002a, b). Marine archaea have been, and still are, focus of special interest. Marine crenarchaeota were first identified by DeLong (1992). Their discovery attracted much attention because they appeared to be the first non-thermophilic members of the Crenarchaeota described; their psychrophilic nature was demonstrated later (Schleper et al., 1997). All marine crenarchaeota cluster in a single group (Group I) of apparently fast-evolving (as deduced from their long 16S rRNA branches in phylogenetic trees) lineages, which includes the psychrophilic sponge symbiont *Cenarchaeum symbiosum* (Preston et al., 1996). All non-thermophilic members of the Crenarchaeota detected in other environments, such as soils, also cluster within this group, suggesting that the adaptation to mesophilic conditions occurred only once in this archaeal kingdom. Also a few clusters of non-thermophilic and as yet non-cultivable marine planktonic euryarchaeota have been identified (Fuhrman and Davis, 1997; López-García et al., 2001a). Marine archaea appear ubiquitous in the ocean, as they have been detected in different oceanic regions (Massana et al., 2000), but their distribution and abundance vary with depth. Different studies point out to a higher abundance of crenarchaeota with depth. This has been suggested by 16S rRNA surveys in Antarctic coastal waters from 0 to 50 m (Murray et al., 1998), and other oceanic regions sampled at the surface and 200 m depth (Massana et al., 1997, 2000). A fluorescent in situ hybridization (FISH) study using specific probes carried out in the equatorial Pacific from surface to 4750 m depth showed that crenarchaeota accounted for as much as 39% of total picoplankton counts at deep-sea areas (Karner et al., 2001). However, euryarchaeota could dominate, at least in terms of diversity, some oceanic regions at depth, such as the Antarctic Polar Front (3000 m depth) (López-García et al. 2001a, b).

Despite intense effort, neither marine planktonic crenarchaeota nor euryarchaeota have been isolated to date, and there is no indication about their possible metabolism, although recent data showing bicarbonate uptake by marine crenarchaeota in darkness suggest that they might fix carbon chemosyntheticly (Wuchter et al., 2003). Environmental genomic data on marine archaea could thus help to get clues on their physiology and ecological role. To date, four marine crenarchaeotal genomic clones containing 16S + 23S rRNA genes have been sequenced: fosmid 4B7 (38.5 kbp, 200 m deep, Eastern North Pacific) (Stein et al., 1996), two fosmid clones of close variants of the sponge symbiont *Cenarchaeum symbiosum* (33 and 42.4 kbp respectively) (Schleper et al., 1998) and fosmid 74A4 (43.6 kbp, surface waters off the Antarctic peninsula) (Béjà et al., 2002a). Additionally, a marine Group II euryarchaeotal BAC clone from the eastern Pacific (60 kbp) and a crenarchaeotal fosmid clone from soil (33.9 kbp) are also available (Béjà et al. 2000a; Quaiser et al., 2002). In this study, we analysed the sequence of a 33.3 kbp genomic fragment of a crenarchaeote from 500 m deep picoplankton at the Antarctic Polar Front, and we compared its organization with those of other marine crenarchaeota coming from more superficial waters. Despite certain gene conservation, we observed considerable gene shuffling when comparing all the crenarchaeotal environmental clones available to date. Phylogenetic analyses of individual genes suggest the existence of multiple horizontal gene transfer (HGT) events between the archaeal kingdoms and also between the two prokaryotic domains. In certain cases, HGT seems to have been involved in genomic shuffling.

**Results and discussion**

**Library construction and phylogenetic screening**

The crenarchaeotal clone DeepAnt-EC39 was retrieved from a cosmid genomic library generated from 500 m deep marine plankton in the size fraction 0.2–5 µm, which contained ~200–250 Mbp of environmental DNA (see *Experimental procedures*). A total of seven archaeal clones were identified by PCR amplification of 16S rRNA genes using specific primers. Six of them clearly affiliated to the Group II of the Euryarchaeota, whereas only one clone, DeepAnt-EC39, belonged to the Crenarchaeota. This supports the idea that Euryarchaeota are more abundant than Crenarchaeota in deep-sea water at the Antarctic Polar Front, corroborating previous suspicions derived from PCR-based diversity surveys (López-García et al., 2001b), which is in apparent contrast to other regions such as the Equatorial Pacific (Karner et al., 2001).
The 16S rRNA of the crenarchaeotal clone DeepAnt-EC39 was closely related to marine environmental sequences retrieved from the deep ocean, such as the clones pIVWA (hydrothermal field at the Iheya Basin, Middle Okinawa Trough, 972 m deep) (Takai and Horikoshi, 1999) and the clone ST-12K2A (brine–seawater interface at the Shaban Deep, Red Sea, 1,331 m deep) (Eder et al., 2002) or from upper layers in the aphotic region such as the clone SB95-57 (Santa Barbara basin, 200 m deep) (Massana et al., 1997). These sequences form a cluster that is clearly separated from other marine crenarchaeota including the genomic clones that have been sequenced to date, those of *C. symbiosum* strains A and B, and fosmids 4B7 and 74A4 (Fig. 1A). Sequencing our clone DeepAnt-EC39 would allow, on the one hand, to have an insight of the genome organisation in this deep-sea Group I crenarchaeota and, on the other hand, to carry out a genomic comparison with the available marine crenarchaeota clones. In particular, it would be interesting to see whether our clone shared higher organizational and gene content similarities with fosmid 74A4, also coming from the Antarctic area but retrieved from surface, or with fosmid 4B7 coming from the North-Eastern Pacific but from deeper waters (200 m).

Analysis of the intergenic transcribed spacer (ITS) region

Marine crenarchaeota appear particularly conserved and ubiquitous in the ocean, as identical (or nearly identical) 16S rRNA gene sequences are easy to detect in very different locations (Fuhrman and Davis, 1997; Massana et al., 2000; and pers. obs.). This restricted diversity at the 16S rRNA level might not reflect differences occurring at the species or strain level. The intergenic spacer region between 16S and 23S rRNA genes offers a possibility to study microdiversity by revealing differences between closely related species and strains, as this is a region highly variable in size, sequence and presence/absence of RNA genes (Gurtler and Stanisich, 1996). In fact, this approach has been already used to explore marine crenarchaeota, which resulted in the detection of a number of environmental ITS clusters including the sequences contained in the crenarchaeotal cosmids available at the time (García-Martínez and Rodríguez-Valera, 2000). In order to better characterize differences at a microdiversity scale, and as soon as we got the ITS region of DeepAnt-EC39 during the cosmid sequencing process, we carried out a comparative analysis with other available environmental marine crenarchaeota ITS sequences. The DeepAnt-EC39 ITS region is 141 bp long, which is comparable in size to ITS regions from the rest of marine crenarchaeota (sizes between 130 and 146 bp), and lacks tRNA genes. However, this short ITS region contrasts with that found in the soil crenarchaeota 29i4, which is 828 bp long (Quaiser et al., 2002), which could be an indicator of a less compact genome.

We then carried out a phylogenetic analysis including selected ITS sequences representative of the different marine crenarchaeota clusters defined previously (García-Martínez and Rodríguez-Valera, 2000) (Fig. 1B). The

![Fig. 1. Phylogenetic trees of 16S rRNA archaeal genes (A) and internal transcribed spacers (ITS) between the 16S and the 23S rRNA genes (B) showing the position of the sequences present in the crenarchaeotal cosmid clone DeepAnt-EC39. In (B), the number in the second position of environmental sequence names corresponds to the depth they were obtained from; M: Mediterranean; A: Antarctic; B: North Atlantic samples. Grey boxes indicate sequences from other genomic clones whose sequence is available. Accession numbers of environmental sequences are in brackets. Only bootstrap values above 50 are shown. The number of substitutions is indicated on the scale bars.](image-url)
828 bp ITS sequence of the soil crenarchaeote was excluded from our analysis because the lack of similarity rendered the alignment with the marine sequences impossible. Surprisingly, our sequence was found to be 100% identical to various sequences coming from Mediterranean waters at depths of 400–450 m (Crena-D group in García-Martínez and Rodríguez-Valera, 2000). This correspondence at the level of a usually highly variable region, such as the ITS, between clones coming from very distant geographical locations validates robustly the ubiquitous nature of this particular lineage of crenarchaeota in mesopelagic waters. The forces, selective and/or related to dispersal, operating to maintain such a remarkable level of conservation remain unknown. In any case, the phylogenetic relationships derived from the ITS analysis, despite the smaller number of positions compared, are overall congruent with those observed using the 16S rRNA gene as a phylogenetic marker (Fig. 1). DeepAnt-EC39 forms a separate cluster with other sequences, and is clearly separated from the clusters containing the C. symbiosum, 4B7 and 74A4 genomic clones.

Sequence, genes and genomic organization of the crenarchaeotal cosmid DeepAnt-EC39

The 33.3 kbp genome fragment contained in the cosmid DeepAnt-EC39 was completely sequenced. The average G+C content was 34%, a value similar to those calculated for the marine crenarchaeotal clones 74A4 and 4B7 (32% and 34% respectively) (Stein et al., 1996; Béjà et al., 2002a), but far from the 55.6% and 57.1% estimated for the two C. symbiosum strains (Schleper et al., 1998).

A total of 41 predicted open reading frames (ORF) longer than 50 amino acids were identified. From these, 18 exhibited significant sequence or structural similarity to gene products with known function, and six were homologous to uncharacterized proteins (Table 1). The 17 remaining predicted ORFs showed a separate cluster with other sequences, and is clearly separated from the clusters containing the C. symbiosum, 4B7 and 74A4 genomic clones.

A detailed analysis of gene density and distribution in our sequence revealed a compact organization with short intergenic regions. The only exception corresponds to the flanking regions of the ORF 35 (Fig. 2). As in the other marine crenarchaeotal cosmids analysed, the rrn operon is composed of the 23S and 16S rRNA genes, but the 5S rRNA gene is located elsewhere in the genome. No tRNA gene was identified in DeepAnt-EC39. The region located upstream of the 5′ end of the rrn operon exhibited no similarity to that of the two other marine planktonic crenarchaeota, 74A4 and 4B7 (the corresponding region in C. symbiosum strain B contains four ORFs that are not homologous to the other crenarchaeote ORFs; the region is not available for strain A). However, 74A4 and 4B7 share a few homologues with similar disposition in this area (Fig. 2). Nevertheless, ORFs 20 and 21 from our clone (menaquinone biosynthesis and hypothetical conserved proteins) had two corresponding adjacent homologues in C. symbiosum but located here downstream of the rrn region, suggesting that a chromosomal rearrangement has taken place around the rrn operon (area shown by a grey circle in Fig. 2).

The region located downstream of the rrn 3′ end was more conserved. In particular, DeepAnt-EC39 and 74A4 showed strong conservation of gene order. Eight out of 11 predicted genes in this DeepAnt-EC39 fragment had corresponding homologues in 74A4: glutamate-1-semialdehyde aminotransferase (GSAT), tetratrico-peptide repeat (TPR) protein, two double-stranded beta-helix fold enzymes, a two-domain (DnaJ-like J-domain and ferredoxin) protein, a HIT superfamily hydrolase, an uncharacterized membrane protein and the 3-hydroxyacyl-CoA dehydrogenase (Table 1 and Fig. 2). The association of a J domain with a ferredoxin domain (ORF 37) represents a unique feature that had only been reported in the 74A4 clone so far. The J domain is a 70 aa stretch at the N-terminal region of the canonical DnaJ that mediates the specific interaction with Hsp70, and constitutes the key feature of the Hsp40 chaperone family (Cyr et al., 1994; Kelley, 1998). In the crenarchaeote clone 4B7, a type I membrane protein containing a DnaJ C-terminal cytoplasmic domain had been detected by Béjà et al. (2002), but it does not correspond to the true J domain that is found in the DeepAnt-EC39 and 74A4 DnaJ-ferredoxin protein. This predicted protein is 248-amino-acids long, the J domain being located at the N-terminal (60 aa) and the ferredoxin domain located at the C-terminal (72 aa) end. In between, there is a 116 aa region without recognizable homology (Fig. 2). An interesting hypothesis that can be advanced from this observation is the occurrence of Hsp70 in Group I crenarchaeota. Hsp70 is absent from the genomes of all hyperthermophilic archaea sequenced to date, and it has been most probably transferred from bacteria to mesophilic euryarchaeota (Philippe et al., 1999). The presence of Hsp70 in mesophilic crenarchaeota, if our prediction is confirmed, would imply that this chaperone has been recruited also by these archaea, helping them to adapt to life at lower temperatures.

Coming back to the genomic structure of the region downstream of the rrn operon, despite an overall similar organization, a few differences exist between the DeepAnt-EC39 and 74A4 clones. These include: (i) a TPR repeat and a type I membrane protein gene in 74A4 (Béjà et al., 2002a) that have been inserted between the homologues of glutamate-1-semialdehyde aminotrans-
## Table 1. Predicted RNA and protein-coding genes in the archaeal cosmid DeepAnt-EC39.

<table>
<thead>
<tr>
<th>ORF</th>
<th>Nucleotide range (aa)</th>
<th>Protein size (aa)</th>
<th>Putative function</th>
<th>Most similar homologue (E-value)</th>
<th>Apparent phylogenetic affinity</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2–163</td>
<td>+ 53</td>
<td>Uncharacterized protein</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>2</td>
<td>190–342</td>
<td>+ 50</td>
<td>Uncharacterized membrane protein</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>3</td>
<td>530–1447</td>
<td>+ 305</td>
<td>Multitransmembrane protein</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>4</td>
<td>1627–2628</td>
<td>– 333</td>
<td>Putative sugar kinase COG0061&lt;sup&gt;+&lt;/sup&gt;</td>
<td><em>Pyrobaculum aerophilum</em> NP_560502.1 (4e-30)</td>
<td>Archaea/Bacteria</td>
<td>Missing in other Crenarchaeota</td>
</tr>
<tr>
<td>5</td>
<td>2658–3530</td>
<td>– 290</td>
<td>Putative ribokinase COG0524&lt;sup&gt;+&lt;/sup&gt;</td>
<td><em>Chloroflexus aurantius</em> ZP_00019243.1 (4e-6)</td>
<td>Bacteria/Archaea</td>
<td>None</td>
</tr>
<tr>
<td>6</td>
<td>3583–3870</td>
<td>+ 95</td>
<td>Ribosomal protein S26 COG4830&lt;sup&gt;+&lt;/sup&gt;</td>
<td><em>Sulfobolus solfataricus</em> NP_342095.1 (7e-15)</td>
<td>Crenarchaeota</td>
<td>Present only in Eukarya and Crenarchaeota, missing in Euryarchaeota and Bacteria</td>
</tr>
<tr>
<td>7</td>
<td>3872–4444</td>
<td>– 190</td>
<td>CDP-diacylglycerol-glycerol-3-phosphate 3-phosphatidyltransferase COG0558&lt;sup&gt;+&lt;/sup&gt;</td>
<td><em>Pyrococcus furiosus</em> NP_578191.1 (1e-27)</td>
<td>Archaea</td>
<td>Multiple copies in Euryarchaeota</td>
</tr>
<tr>
<td>8</td>
<td>4444–5319</td>
<td>– 291</td>
<td>Agmatinase COG0010&lt;sup&gt;+&lt;/sup&gt;</td>
<td><em>Thermoplasma volcanium</em> NP_111057.1 (8e-46)</td>
<td>Archaea</td>
<td>None</td>
</tr>
<tr>
<td>9</td>
<td>5324–7174</td>
<td>– 616</td>
<td>threonyl-tRNA synthetase COG0441&lt;sup&gt;+&lt;/sup&gt;</td>
<td><em>Pyrococcus furiosus</em> NP_579080.1 (0.0)</td>
<td>Archaea</td>
<td>Possible interkingdom HGT (tree shown in Fig. 3A)</td>
</tr>
<tr>
<td>10</td>
<td>7231–7500</td>
<td>– 89</td>
<td>Uncharacterized protein</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>11</td>
<td>7667–7996</td>
<td>– 109</td>
<td>Uncharacterized protein</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>12</td>
<td>8082–9050</td>
<td>+ 322</td>
<td>Deoxyhypusine synthase COG1899&lt;sup&gt;+&lt;/sup&gt;</td>
<td><em>Nostoc sp.</em> NP_487844.1 (4e-53)</td>
<td>Archaea</td>
<td>Independent interkingdom HGT events. Gene duplication in Archaeoglobus and Methanosarcinales (tree shown in Fig. 5)</td>
</tr>
<tr>
<td>13</td>
<td>9182–9409</td>
<td>– 75</td>
<td>Uncharacterized protein</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>14</td>
<td>9509–9814</td>
<td>+ 101</td>
<td>Uncharacterized protein</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>15</td>
<td>9797–10177</td>
<td>– 126</td>
<td>Uncharacterized protein</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>16</td>
<td>10230–10736</td>
<td>+ 168</td>
<td>DNA-directed RNA polymerase lib COG1758&lt;sup&gt;+&lt;/sup&gt;</td>
<td><em>Caenorhabditis elegans</em> NP_496278.1 (5e-12)</td>
<td>Archaea</td>
<td>None</td>
</tr>
<tr>
<td>17</td>
<td>10786–11091</td>
<td>+ 101</td>
<td>Uncharacterized protein</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>ORF</td>
<td>Nucleotide range</td>
<td>Protein size (aa)</td>
<td>Putative function (COG Accession Number if available)</td>
<td>Most similar homologue (E-value)</td>
<td>Apparent phylogenetic affinity</td>
<td>Comments</td>
</tr>
<tr>
<td>-----</td>
<td>------------------</td>
<td>-------------------</td>
<td>--------------------------------------------------------</td>
<td>---------------------------------</td>
<td>-------------------------------</td>
<td>----------</td>
</tr>
<tr>
<td>18</td>
<td>11281–11523</td>
<td>– 60</td>
<td>Conserved hypothetical protein COG2522</td>
<td>Sulfolobus solfataricus NP_342420.1 (3e-5)</td>
<td>Crenarchaeota</td>
<td>Exclusive of some, but not all, Crenarchaeota</td>
</tr>
<tr>
<td>19</td>
<td>11587–12534</td>
<td>+ 315</td>
<td>Asparagine synthase (asnB) COG0367</td>
<td>Wigglesworthia brevipalpis NP_715136.1 (2e-22)</td>
<td>Archaea</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>12721–13608</td>
<td>– 295</td>
<td>Menaquinone biosynthesis protein (1,4-dihydroxy-2-naphthoate octaprenyltransferase) COG1575</td>
<td>Cenarchaeum symbiosum T31325 (4e-78)</td>
<td>Bacteria</td>
<td>Bacterial protein probably transferred independently to marine crenarchaeota and Halobacterium sp. (tree shown in Fig. 4A)</td>
</tr>
<tr>
<td>21</td>
<td>13608–14009</td>
<td>– 133</td>
<td>Hypothetical conserved protein COG4911</td>
<td>Cenarchaeum symbiosum AAC62693.1 (2e-39)</td>
<td>Crenarchaeota</td>
<td>Present only in Crenarchaeota (except in Aeropyrum pernix), possible HGT to the bacteria Aquilax aeolicus and Leptospira interrogans (tree shown in Fig. 4B)</td>
</tr>
<tr>
<td>22</td>
<td>14045–15397</td>
<td>– 450</td>
<td>Uncharacterized membrane protein COG2114</td>
<td>None</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>15390–16136</td>
<td>– 248</td>
<td>Probable adenylate cyclase COG2114</td>
<td>Nostoc punctiforme ZP_0010893.1 (7e-13)</td>
<td>Bacteria</td>
<td>Present only Bacteria and Eukarya. Likely transferred from bacteria to the crenarchaeote DeepAnt-EC39</td>
</tr>
<tr>
<td>24</td>
<td>16252–17052</td>
<td>– 266</td>
<td>Predicted transcriptional regulator COG4742</td>
<td>Methanosarcina acetivorans NP_619475.1 (5e-17)</td>
<td>Euryarchaeota</td>
<td>Present in three copies in Archaeoglobus fulgidus and Methanobacterium thermosautotrophicum, and in multiple copies in Methanosarcinales (tree shown in Fig. 3B)</td>
</tr>
<tr>
<td>25</td>
<td>17180–17425</td>
<td>+ 81</td>
<td>Uncharacterized protein</td>
<td>None</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>17501–17962</td>
<td>+ 153</td>
<td>Uncharacterized protein</td>
<td>None</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>18022–18225</td>
<td>– 67</td>
<td>Uncharacterized membrane protein</td>
<td>None</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>18456–18713</td>
<td>+ 85</td>
<td>Uncharacterized protein</td>
<td>None</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>19150–20121</td>
<td>– 323</td>
<td>Thioredoxin reductase (TnxB-2) COG0492</td>
<td>Sulfolobus solfataricus 344086.1 (2e-61)</td>
<td>Archaea</td>
<td>Widely distributed in the three domains</td>
</tr>
<tr>
<td>30</td>
<td>20125–20397</td>
<td>– 90</td>
<td>Putative transcriptional regulator</td>
<td>Pyrococcus abyssi CA055655.1 (3e-06)</td>
<td>Archaea</td>
<td>Probably transferred to Thermobifida fusca and Streptomyces coelicolor</td>
</tr>
<tr>
<td>rRNA</td>
<td>20650–23658</td>
<td>+ 23S ribosomal RNA (large subunit)</td>
<td>Uncultured crenarchaeote 4B7 U40238 (GI:14548123) (90% identity)</td>
<td>None</td>
<td>Group I Crenarchaeota</td>
<td></td>
</tr>
<tr>
<td>ORF</td>
<td>Nucleotide range</td>
<td>D</td>
<td>Protein size (aa)</td>
<td>Putative function</td>
<td>Most similar homologue</td>
<td>Apparent phylogenetic affinity</td>
</tr>
<tr>
<td>-----</td>
<td>------------------</td>
<td>---</td>
<td>-----------------</td>
<td>-----------------</td>
<td>----------------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td>31</td>
<td>25346–25555</td>
<td>− 69</td>
<td>Uncharacterized membrane protein glutamate-1-semialdehyde aminotransferase COG0001&lt;sup&gt;c&lt;/sup&gt;</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>32</td>
<td>25542–26855</td>
<td>+ 437</td>
<td>Uncultured crenarchaeote 74A4</td>
<td>Uncultured crenarchaeote 74A4</td>
<td>EAA93183.2 (e-176)</td>
<td>Crenarchaeota</td>
</tr>
<tr>
<td>33</td>
<td>26881–27699</td>
<td>+ 275</td>
<td>TPR-repeat protein (Tetratricopeptide repeat domain)</td>
<td>Uncultured crenarchaeote 74A4</td>
<td>EAF39466-23 (5e-73)</td>
<td>Archaea</td>
</tr>
<tr>
<td>34</td>
<td>27702–28415</td>
<td>− 238</td>
<td>Uncharacterized conserved protein</td>
<td>Uncultured crenarchaeote 74A4</td>
<td>EAAK96087.1 (2e-50)</td>
<td>Crenarchaeota</td>
</tr>
<tr>
<td>35</td>
<td>29263–29496</td>
<td>− 77</td>
<td>Uncharacterized protein</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>36</td>
<td>30014–30481</td>
<td>155</td>
<td>Uncharacterized conserved protein, contains a double-stranded beta-helix fold domain COG1917</td>
<td>Uncultured crenarchaeote 74A4</td>
<td>EAA96088.1 (3e-25)</td>
<td>Archaea/Bacteria</td>
</tr>
<tr>
<td>37</td>
<td>30527–31273</td>
<td>+ 248</td>
<td>J domain from the molecular chaperone of the DnaJ class (Hsp40), with a C-terminal 3Fe-4S ferredoxin domain COG0484/COG1141</td>
<td>Uncultured crenarchaeote 74A4</td>
<td>EAAK96090.1 (2e-35)</td>
<td>Group I Crenarchaeota</td>
</tr>
<tr>
<td>38</td>
<td>31274–31684</td>
<td>− 136</td>
<td>HIT superfamily hydrolase COG0537</td>
<td>Uncultured crenarchaeote 74A4</td>
<td>EAAK96091.1 (4e-41)</td>
<td>Archaea</td>
</tr>
<tr>
<td>39</td>
<td>31722–32021</td>
<td>− 99</td>
<td>Uncharacterized membrane protein</td>
<td>Uncultured crenarchaeote 74A4</td>
<td>EAAK96092.1 (e-155)</td>
<td>Archaea</td>
</tr>
<tr>
<td>40</td>
<td>32048–32230</td>
<td>− 60</td>
<td>Uncharacterized protein</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>41</td>
<td>32260–33345</td>
<td>+ 361</td>
<td>3-hydroxyacyl-CoA dehydrogenase COG1250&lt;sup&gt;c&lt;/sup&gt;</td>
<td>Uncultured crenarchaeote 74A4</td>
<td>EAAK96092.1 (e-155)</td>
<td>Archaea</td>
</tr>
</tbody>
</table>

<sup>a.</sup> COG functional class 'Translation/Transcription/DNA replication, recombination and repair.'
<sup>b.</sup> COG functional class 'Cellular processes'.  
<sup>c.</sup> COG functional class 'Transport and energy metabolism'.  
ORFs are numbered 1–41 following the sense 5' to 3' in Fig. 2. E-values were obtained from psi-BLAST searches. D: direction of transcription.
ferase gene (GSAT) and the second TPR repeat protein that is also found in DeepAnt-EC39; (ii) the absence in DeepAnt-EC39 of the third double-stranded beta-helix fold enzyme found in 74A4; and (iii) the insertion in DeepAnt-EC39 of a large intergenic region containing the ORF 35 between the two double-stranded beta-helix fold enzymes. Part of this genomic disposition shared by DeepAnt-EC39 and 74A4 is also common to that found in *C. symbiosum* (see grey underlying boxes in Fig. 2). The fact that the order and arrangement of the three ORFs 32–34 in DeepAnt-EC39 is completely conserved in *C. symbiosum* indicates that the two ORFs present in 74A4 between GSAT and the TPR protein correspond to an insertion specific to this clone. It is interesting to note here that the first TPR protein in 74A4, which corresponds to an insertion, is exclusively present in this clone and in euryarchaeota, as deduced from BLAST searches. This suggests that it was acquired by horizontal gene transfer (HGT). Therefore, HGT may have been one of the possible causes triggering genome reorganization in these archaea. In the case of 4B7, only the GSAT gene is conserved downstream of the *rrn*. Indeed, there is a systematic association of the *rrn* operon and the GSAT genes, which are contiguous in all marine planktonic crenarchaeota analysed so far (Béjà *et al.*, 2002a). However, this is not the case in the only *rrn*-containing genome fragment available from Group I soil crenarchaeota (Quaiser *et al.*, 2002). The rest of the region is not conserved in 4B7.

In spite of a close proximity in the 16S rRNA phylogenetic tree, the genomic organization of the region containing the *rrn* operon appeared very divergent in all marine crenarchaeota. The poor conservation of gene order, already emphasized by Béjà *et al.* (2002), suggested that multiple rearrangements occurred in this region. Indeed, the flanking regions of the *rrn* operon + GSAT appear to have undergone recombination events frequently. The region upstream of the *rrn* operon shares some organizational similarities between 4B7 and 74A4, the two planktonic clones from more superficial waters, whereas the downstream area is conserved in *C. symbiosum*, and in DeepAnt-EC39 and 74A4, the two Antarctic clones. This indicates that the rearrangement of the downstream region took place in 4B7. A possible explanation could be that recombination is facilitated by the presence of a second *rrn* operon in the chromosome. Although all cultivated crenarchaeota, which are hyperthermophiles, are known to have a single *rrn* operon, the possibility of a second *rrn* operon being present in mesophilic crenarchaeota cannot be excluded. In fact, a trend of this kind can be observed in euryarchaeota, as all isolated hyperthermophilic species have a single *rrn* operon, whereas some halophiles, such as *Halofex* or *Haloarcula* spp. carry two operons (Mylvaganam and Dennis, 1992; López-García *et al.*, 2002).
16S rDNA-containing genome fragment of mesopelagic crenarchaeote

and some mesophilic methanogens carry even three operons (Deppenmeier et al., 2002; Galagan et al., 2002). In addition, horizontal gene transfer could also play an active role in genome rearrangement as can be hypothesized from the phylogenetic analysis of the DeepAnt-EC39 ORFs, which revealed in several cases the existence of HGT (see below).

**Phylogenetic analyses of protein genes**

Phylogenetic analysis of all individual ORFs in DeepAnt-EC39 showed that most of them have undergone complex different evolutionary histories (all trees and alignments are available upon request). Only the ribosomal protein S26 gene (ORF 6) allowed inferring a phylogeny fully congruent with the 16S rRNA tree, with the DeepAnt-EC39 sequence branching early within a robust cluster grouping all available crenarchaeotal sequences (not shown). Interestingly, the ribosomal protein S26 appears to be present only in crenarchaeota, including *Aeropyrum pernix* (where the gene has not been annotated, but is identifiable by TBLASTN search), and eukaryotes. It is not striking that the 16S rRNA tree and the ribosomal protein S26 produce congruent phylogenies, given their co-evolution in the ribosomal machinery. However, it is interesting to note that all the remaining predicted genes in DeepAnt-EC39 rendered quite disparate tree topologies. The majority of the predicted genes belonged to multigenic families that showed complex phylogenies with multiple paralogues resulting from duplications events. In addition various genes failed to recover the monophyly of the domain Archaea, and several likely cases of horizontal gene transfer were identified. Of course, explanations other than HGT are conceivable, such as duplication and differential gene loss. However, in most cases the number of duplication and loss events would be so high that the most parsimonious explanation is HGT. Here, we show a few representative examples of genes that appear to have undergone horizontal transfer (i) between the two archaeal kingdoms, and (ii) between the two prokaryotic domains.

The first example concerns the threonyl tRNA synthetase gene (ORF 9). The phylogeny inferred showed two separated groups very well supported by a bootstrap proportion (BP) of 100% (Fig. 3A). Each of these two groups was characterized by various specific signatures (insertions/deletions and amino-acid sequences) that defined two distinct types of sequences (indicated by I and II in Fig. 3A). One type of sequences (II) grouped all bacteria and eukaryotes, as well as some euryarchaeota (halophiles and Thermoplasmales) and two *Sulfolobus* spp., whereas the other type (I) clustered most euryarchaeota, both hyperthermophilic lineages and methanogens, the DeepAnt-EC39 sequence and the very divergent.

---

**Fig. 3.** Possible cases of horizontal gene transfer (HGT) between the two archaeal kingdoms, Euryarchaeota and Crenarchaeota, involving DeepAnt-EC39 genes.

A. Phylogenetic tree of threonyl tRNA synthetase genes (DeepAnt-EC39 ORF 9). The region of the tree containing triangles has been reduced in scale, each triangle representing the number of species that is indicated inside. The two groups of sequences labelled I and II are distinguishable by specific amino acid signatures.

B. Phylogenetic tree of homologues to DeepAnt-EC39 ORF 24, a predicted transcriptional regulator of the Asr family. Only bootstrap values above 50 are shown. The number of substitutions is indicated on the scale bars.

---

Pyrococcus furiosus and Aeropyrum pernix sequences. DeepAnt-EC39 emerged clearly amongst the euryarchaeota (BP 100%), after a group including thermophilic methanogens (Methanopyrus kandleri, Methano- c Calderococcus jannaschii and Methanothermophobacter thermautotrophicus), and as a sister-group of the hyper- thermophilic Archeoglobus fulgidus and Pyrococcus spp. and the mesophilic Methanosarcinales. The atypical phy- logeny of this marker in archaea is suggestive of horizon- tal gene transfer between the two archaeal kingdoms. In particular, the DeepAnt-EC39 threonyl tRNA synthetase gene was most probably acquired from euryarchaeota. This might not be surprising, as aminoacyl-tRNA syn- thetases are particularly prone to horizontal transfer (Wolf et al., 1999; Woese et al., 2000).

Another very likely case of gene transfer from eur- yarchaeota, possibly methanogens, to our clone DeepAnt- EC39 was that of the ORF 24, a predicted protein related to transcription regulators of the ArsR family (Table 1). The gene is absent from all other crenarchaeotal genomes, but it is present in many copies in *M. thermoautotrophicus* (three copies), *A. fulgidus* (three copies), and various Methanosarcina spp. (*M.arkeri, M. mazei and M. ace- tivorans*, numerous copies). As can be inferred from the phylogenetic tree shown in Fig. 3B, successive duplica- tion events occurred independently in *M. thermoau- totrophicus, A. fulgidus* and Methanosarcinales. The number of duplications in the ancestor of Methanosarcina spp. was particularly impressive. However, the study of the genomic context of those genes did not provide a clear scenario to explain their distribution in the three Metha- nosarcina genome sequences available, the genes being duplicated but not the respective surrounding genomic regions (not shown). Although the evolutionary history of this gene seems complex in these methanogens, involving duplications, losses and genome rearrangement events, its acquisition by HGT by DeepAnt-EC39 appears convincing.

Likely examples of gene transfer between the two prokaryotic domains were observed in both directions, from bacteria to archaea, and from archaea to bacteria, sometimes involving various independent transfers. For instance, the putative DeepAnt-EC39 menaquinoine bio- synthesis protein homologous to MenA (ORF 20) was also found in one euryarchaeote (*Halobacterium* sp.) and in the two *C. symbiosum* strains. MenA is present in various bacterial divisions (low and high GC Gram positives, some proteobacteria, cyanobacteria, Thermotogales, etc.), but no bona fide homologues are identifiable in eukaryotes. Menaquinone (vitamin K2) plays an essential role in sev- eral anaerobic transport systems being a major electron carrier during anaerobic growth with various electron acceptors. Menaquinones are present in at least some archaea (Thurl et al., 1985; Shimada et al., 2001), but their biosynthetic pathways remain to be elucidated. In bacteria, six genes are implied in its biosynthesis, MenA to MenF. MenB to MenF are contiguous in the *Escherichia coli* chromosome, but MenA is located elsewhere (Suvarna et al., 1998). Since this gene is widespread in bacteria, and the archaeal sequences appeared interspersed within them, these archaea most probably acquired these genes from bacteria by HGT (Fig. 4A). Furthermore, because the *Halobacterium* sp. sequence emerged far from the cluster of crenarchaeotal sequences (the two *C. symbiosum* and the DeepAnt-EC39 sequences form a robust group with a BP 100%), the most parsimonious explanation is the independent acquisition from two different bacterial donors. The absence of the gene in the clones 74A4 and 4B7 might reflect either secondary loss or genomic rearrangements resulting in gene translocation to another part of the genome. In addition to MenA, the genes encoding MenB, MenD and MenF are annotated in the *Halobacterium* sp. genome (Ng et al., 2000). MenC is clearly missing, but a MenE homologue can be detected by BLAST (annotated as long-chain fatty acid-CoA ligase). This suggests that a *Halobacterium* ancestor acquired the entire menaquinone biosynthesis operon, and that MenC was subsequently lost. A similar situation, i.e. that the entire operon was acquired by HGT, could exist in group I crenarchaeota. Whether this operon is functional and whether it was acquired to facilitate menaquinone synthesis under mesophilic conditions in these archaea remain open questions. Other sequenced archaeal genomes lack MenA, MenC and MenD, although distant homologues to MenB, MenE and MenF belonging to multigene families can be identified. Therefore, other archaea possessing menaquinones lacking the MenA-F genes in their genomes must have another way of synthesis.

Conversely, some likely transfers from crenarchaeota to bacteria were observed, as is illustrated by the phylogeny of ORF 21, a hypothetical protein of unknown function (Fig. 4B). This conserved protein appears specific to cre- narchaeota, including three marine planktonic crenarchaeota (the gene appears only absent in *A. pernix*), but it is also found in the two unrelated bacteria *Aquilexi aelolicus* and *Leptospira interogens*. The taxonomic distribution of this gene suggests that these bacteria gained the gene from crenarchaeota from one or, possibly, two indepen- dent HGTs. At least the acquisition of archaeal genes by *A. aelolicus* appears well documented (Deckert et al., 1998; Forterre et al., 2000).

Finally, we also detected far intricate HGT cases involving inter- and intradomain transfers. One such example is that of the deoxypusine synthase gene (ORF 12). We detected homologues of this ORF in eukaryotes, archaea and a few bacteria (*Cytophaga Hutchinsonii, Chloroflexus aurantiacus, Chlorobium tepi-
Deoxyhypusine synthase intervenes in the post-translational synthesis of hypusine, a modified lysine residue that is present in a single protein in eukaryotes and archaea, the initiation factor eIF5A (aIF5A in archaea) (Bartig et al., 1992; Park et al., 1993). Its expression is essential for yeast viability and for the archaeal cell cycle (Park et al., 1998; Jansson et al., 2000). The presence of deoxyhypusine synthase in bacteria has not been described previously. The phylogenetic analysis of this gene shows a clear separation between eukaryal and archaeal sequences (BP 96%) (Fig. 5). The history of this gene appears complex in archaea, with some interkingdom transfers and duplication events, although specific insertions/deletions and amino acid motifs confirm the robustness of several groups in the tree (not shown). The presence of two paralogues in *A. fulgidus* and the Methanosarcinales certainly resulted from a duplication event in a common ancestor of this group. Thermoplasmales appear to have acquired the deoxyhypusine synthase gene from crenarchaeota, although the less likely possibility that the gene was originally from euryarchaeota and transferred from Thermoplasmales to thermophilic crenarchaeota cannot be completely ruled out. The marine crenarchaeote DeepAnt-EC39 sequence branches in a robust (BP 97%) group with three cyanobacterial sequences that emerges as a sister group to the Archaeoglobales and Methanosarcinales. It is possible then that DeepAnt-EC39 acquired the euryarchaeotal gene indirectly through cyanobacteria or, alternatively, that cyanobacteria gained the euryarchaeotal gene from marine crenarchaeota. Unfortunately, the poor resolution (reflected in low BP) after the divergence of the *M. thermoautotrophicus* and *M. kandleri* in the tree does not allow inferring unambiguously the number and order of the gene transfers towards these lineages. At any rate, this would be a case of secondary HGT, an event by which a gene has been passed to two different phylogenetic lineages successively, whose origin can be still traced back to the euryarchaeota. The original phylogenetic signal has been therefore retained despite the fact that the acquisition of the deoxyhypusine synthase gene by these cyanobacteria and by the ancestor of DeepAnt-EC39 was probably a very ancient event, predating the divergence of *Archaeoglobus* and the Methanosarcinales.
Conclusions

We have analysed here the sequence of the 33.3 kbp genome fragment of a crenarchaeote retrieved from 500 m deep Antarctic plankton, DeepAnt-EC39. A comparison of gene content and gene organization with the other available 16S rDNA-containing genome fragments from marine Group I crenarchaeota shows a highly variable genomic structure. This is characterized by the shared presence of different sets of genes and by multiple rearrangements. DeepAnt-EC39 resembles the surface planktonic Antarctic clone 74A4 on one side of the rrn operon but, in turn, 74A4 has similarities in gene content and organization with the 200 m deep North Atlantic clone 4B7 on the other side. Therefore, there is no apparent correlation between gene organization and geographical location in these planktonic crenarchaeota. Furthermore, the analysis of the ITS region, a usually highly variable marker used for microdiversity studies, suggests that DeepAnt-EC39 belongs to a geographically widespread marine lineage, its ITS sequence being identical to other crenarchaeotal ITSs retrieved from deep Mediterranean waters. This, however, reinforces the idea that this lineage is specific of mesopelagic waters, and that vertical stratification is a further more important parameter influencing microbial distribution in oceans than geographical location. From the limited data available, it appears that the immediate surroundings of the rrn operon are particularly prone to recombination. The fact that one clone resembles a first one on the one side, and a second on the other side could imply the existence of more that one rrn operon in these crenarchaeota that would recombine easily within the chromosome, being a source of genomic reorganization. Indeed, recombination events between rrn operons resulting in the inversion of large chromosomal segments have been previously observed in halophilic euryarchaeota (López-García et al., 1995). Of course, the presence of multiple rrn operons in Group I crenarchaeota needs to be demonstrated in the future. In addition, gene insertions and/or deletions are observed. Taken together, these data offer a highly dynamic vision of the genome in these crenarchaeota, an image that is further supported by the detection of multiple horizontal gene transfers as suggested from phylogenetic analyses of individual genes.

The phylogenetic analysis of the 41 ORFs predicted in DeepAnt-EC39 revealed a remarkable percentage of putative HGT cases involving different lineages, but not necessarily the marine crenarchaeota (16 genes, 39%). The proportion of genes likely acquired by DeepAnt-EC39/marine crenarchaeota (11 genes, 26%) appeared also notably high. These include basically operational genes, which are thought to be more prone to this type of exchange. Therefore, this value cannot be extrapolated to the whole genome of this crenarchaeon, as regions with higher density of informational genes will be probably less affected by HGT. For this genome fragment, transfers are observed both between the two prokaryotic domains and between the two archaeal kingdoms, involving sometimes multiple-independent and/or secondary transfers. However, despite extensive HGT, transfer direction and gene origin were often identifiable. Our results support the increasingly substantiated idea that HGT plays a major role in the environment, particularly that of genes involved in metabolic and adaptive physiological functions. These genes would superimpose to a core of conserved genes involved in informational processes whose transfer, although possible, appears less favourable (Jain et al., 1999; Daubin et al., 2003). As data accumulate, this hypothesis would become more statistically testable. Interestingly, the possible gene transfers observed in DeepAnt-EC39 involve as partners either mesophilic bacteria or mesophilic euryarchaeota. In the latter case, methanogens, most often relatives to Methanosarcinales, appear to be the original source for several of the marine crenarchaeotal sequences analysed. The acquisition of genes from mesophilic euryarchaeota may have therefore played a crucial role in the secondary adaptation of Group I crenarchaeota to life under mesophilic conditions.
Experimental procedures

Sampling and genomic library construction

The plankton sample used to construct the genomic library was collected at a depth of 500 m in the Antarctic Polar Front at the Drake Passage (59° 19′ S, 55° 45′ W) in December 1998, during the Spanish oceanographic campaign DHARMA 98 (Heo52). Five hundred metre deep water was collected in Niskin bottles attached to a rosette and a conductivity-temperature-depth sensor. Seawater (~30 l) was pre-filtered through a nylon mesh, and then filtered through a 5 μm pore size polycarbonate filter. The remaining microbial biomass was collected with 0.2 μm pore size Sterivex filter units (Durapore, Millipore). The Sterivex units were filled with 1.8 ml of lysis buffer (40 mM EDTA, 50 mM Tris-HCl, 0.75 M sucrose) and stored at −20°C. After a first lysis step with 1 mg ml⁻¹ lysozyme at 37°C for 30 min, lysis proceeded with the addition of 50 ng ml⁻¹ proteinase K and 1% sodium dodecyl sulphate (final concentrations) at 55°C for 2 h. Nucleic acids were extracted twice with phenol-chloroform-isooamyl-alcohol, and once with chloroform-isooamyl-alcohol, and then desalted and concentrated using a Centricon-100 system as previously described (Massana et al., 1997).

Our environmental genomic library was constructed with DNA corresponding to the 0.2−5 μm plankton biomass fraction. We started from an initial amount of ~3.5 μg of DNA, as estimated by comparison with different amounts of linear T7 DNA in agarose gels. DNA was broken by mechanical shearing and fractionated in a 0.8% low-melting point agarose gel after DNA ends were repaired (Epicentre pWEB™ Cloning System). The band corresponding to 35–45 kbp DNA fragments was cut off the gel and digested with gelase. After concentration, the yield of 40 kbp DNA was ~0.2 μg. Half of this amount (~0.1 μg) was cloned in the 8179 bp-cosmid vector of the pWEB Cloning System (Epicentre). For this study, only half of the ligated cosmids were packaged in phage particles and used to transfect E. coli EPI305. A total of 6107 cosmid clones was obtained in this way, i.e. an initial library of 200–250 Mbp assuming an average insert size of 35−40 kbp.

Library screening and sequence of cosmid clone DeepAnt-EC39

The library was pooled in groups of 20 clones that served for PCR screening. A small amount of biomass from each clone was suspended in 10 mM Tris pH 8 to a final volume of 150 μl. Pools were lysed by heat (3 min 98°C) and centrifuged at high speed for 3–5 min. Next, 3 μl of supernatant was used for PCR reactions. The genomic library was PCR screened using the 16S rDNA archaeal-specific primer 23FLP (gcggagtacgcccggcgtcgaYCTGCGTGATYTCTGGC; lowercase letters correspond to a polylinker region) and the prokaryotic reverse primer 1492R (GTTTAACCTGTGTTAACGATT). Seven pools gave expected-size (~1.5 kbp) amplification products. Further PCR screening of individual clones integrating these pools led to the identification of seven archaeal 16S rDNA-containing cosmid clones. The seven amplicons were sequenced using primer 1492R. From these, only clone DeepAnt-EC39 affiliated to the Crenarchaeota.

DeepAnt-EC39 was entirely sequenced (Genome Express). DNA was purified from a 350 ml overnight culture. A shotgun library was generated by subcloning mechanically sheared DNA (2.5−5 kbp) into pUC18. Inserts were sequenced with vector primers using a MegaBase 4000 capillary sequencer (Amershan Biosciences). After sequence assembly, two major contigs were obtained. The remaining gaps were closed by primer walking using sequence-derived primers. The sequence of the crenarchaeotal cosmid DeepAnt-EC39 is available at GenBank with Accession Number AY316120.

Open reading frame and tRNA gene search

The detection of ORFs was performed with the ORF finder software at the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov). Several criteria were used to define bona fide ORFs among all the potential ORFs detected. First, only non- (or very little) overlapping sequences longer than 50 amino acids were retained. Second, when putative ORFs were detected in different reading frames, we selected those that had known homologues (as detected by BLAST searches) and third, in the case of various putative ORFs in different reading frames with no known homologues, we selected those ORFs that were longer. The presence of tRNA genes was explored with the program tRNAscan (Lowe and Eddy, 1997).

Protein identification

We used the programs PSI-BLAST, TBLASTN and BLASTP to look for sequence similarity with published protein sequences (Altschul et al., 1997). The analyses of predicted known conserved domains and transmembrane segments were carried out using the PROSITE and TMHRED software respectively (Bairoch, 1991; Hofmann and Stoffel, 1993; Hofmann et al., 1999). Predicted proteins encoded in cosmid DeepAnt-EC39 were classified according to the clusters of orthologous groups (COGs) database (Tatusov et al., 1997, 2000).

Phylogenetic analysis

In the case of 16S rRNA genes, the closest relatives to the DeepAnt-EC39 as identified by BLASTN were incorporated into a large alignment containing ~16 000 prokaryotic sequences. Sequences were aligned automatically using the program BABA (H. Philippe, pers. comm.), and the alignment was refined manually using the program ED of the MUST package (Philippe, 1993). After removal of gaps and unambiguously aligned positions, a total of 1285 positions were used for the analysis. The tree was constructed by neighbour-joining (NJ) applying the Kimura 2-parameter model of sequence evolution (Kimura, 1980) using the MUST package. For analysis of internal transcribed spacer regions (ITS), sequences from marine crenarchaeota representing various subgroups (García-Martínez and Rodríguez-Valera, 2000) were retrieved from GeneBank and aligned using CLUSTALX (Thompson et al., 1997). A NJ tree was generated after removal of non-aligned ITS ends; internal informative gaps...
were included in the analysis, and a total of 144 positions were considered. For both, 16S rRNA gene and ITS analyses, bootstrap values were calculated from 1000 replicates using the MUST package.

In the case of proteins, the sequences homologous to DeepAnt-EC39 ORFs identified by BLAST were retrieved from GeneBank and aligned using CLUSTALW (Thompson et al., 1994). Amino acid alignments were inspected manually using the MUST package. All individual alignments and the corresponding phylogenetic trees are available upon request.

Acknowledgements

This project was financed by the Generalitat Valenciana (Grant GV01-16), the European Commision Project GEMINI (QLK3-CT-2002-02056) and the French CNRS – Ministère de la Recherche program Séquençage à Grande Echelle. We also acknowledge L. Aggerbeck and the IFR Séquençage à Grande Echelle for access to the DNA-chip platform facilities at Gil-sur-Yvette, France.

References


Tatusov, R.L., Koonin, E.V., and Lipman, D.J. (1997) A


