Proteorhodopsin genes are distributed among divergent marine bacterial taxa

José R. de la Torre[†‡], Lynne M. Christianson[†], Oded Béjà[†§], Marcelino T. Suzuki[†¶], David M. Karl[‖], John Heidelberg[**], and Edward F. DeLong[†,††]

[†]Monterey Bay Aquarium Research Institute, 7700 Sandholdt Road, Moss Landing, CA 95039; [§]Department of Biology, Technion-Israel Institute of Technology, Haifa 32000, Israel; [¶]Chesapeake Biological Laboratory, University of Maryland, Solomons, MD 20688; [‖]Department of Oceanography, University of Hawaii, Manoa, HI 96822; and [**]Institute for Genomic Research, Rockville, MD 20850

Edited by Sallie W. Chisholm, Massachusetts Institute of Technology, Cambridge, MA, and approved August 21, 2003 (received for review June 10, 2003)

Proteorhodopsin (PR) is a retinal-binding bacterial integral membrane protein that functions as a light-driven proton pump. The gene encoding this photoprotein was originally discovered on a large genome fragment derived from an uncultured marine γ-proteobacterium of the SAR86 group. Subsequently, many variants of the PR gene have been detected in marine plankton, via PCR-based gene surveys. It has not been clear, however, whether these different PR genes are widely distributed among different bacterial groups, or whether they have a restricted taxonomic distribution. We report here comparative analyses of PR-bearing genomic fragments recovered directly from planktonic bacteria inhabiting the California coast, the central Pacific Ocean, and waters offshore the Antarctica Peninsula. Sequence analysis of an Antarctic genome fragment harboring PR (ANT32C12) revealed moderate conservation in gene order and identity, compared with a previously reported PR-containing genome fragment from a Monterey Bay γ-proteobacterium (EBAC31A08). Outside the limited region of synteny shared between these clones, however, no significant DNA or protein identity was evident. Analysis of a third PR-containing genome fragment (HOT2C01) from the North Pacific subtropical gyre showed even more divergence from the γ-proteobacterial PR-flanking region. Subsequent phylogenetic and comparative genomic analyses revealed that the Central North Pacific PR-containing genome fragment (HOT2C01) originated from a planktonic α-proteobacterium. These data indicate that PR genes are distributed among a variety of divergent marine bacterial taxa, including both α- and γ-proteobacteria. Our analyses also demonstrate the utility of cultivation-independent comparative genomic approaches for assessing gene content and distribution in naturally occurring microbes.

marine bacteria | picoplankton | genomics | evolution | ecology

Light-absorbing proteins covalently bound to retinal are found throughout the three domains of extant life: Eucarya, Bacteria, and Archaea. These pigments can serve a variety of functions, from photoreception in the visual system and related phototactic sensory responses to ion transport and energy generation. One of the best-studied categories of light-absorbing pigments is the rhodopsins, retinylidene proteins of the opsin superfamily (1). In higher Metazoa, rhodopsins are found primarily in the photoreceptor cells of the retina and are responsible for the detection and discrimination of photons of different wavelengths. Microbial rhodopsins were originally thought to occur exclusively in extremely halophilic archaea, and these archaeal photoproteins share structural and mechanistic similarities. They are all integral membrane proteins, containing seven helical plasma membrane-spanning domains that form a channel between the cytoplasm and the extracellular environment. Key residues lining the inner surface of the channel bind and coordinate a single molecule of *all-trans* retinal, which, on absorption of a photon, changes to a 13-cis isomer. The corresponding change in protein conformation enables ion transport through the central channel, or a signal transduction event via accessory effector proteins that interact with the sensory rhodopsin family of proteins.

Previously, rhodopsins were not known to occur in the domain Bacteria. Recent cultivation-independent surveys of genome fragments recovered directly from marine plankton revealed a broader phylogenetic distribution of these photoproteins (2). This genomics-enabled discovery of a new bacterial rhodopsin, proteorhodopsin (PR), suggested that retinal-bound photoproteins that function as light-driven proton pumps exist in planktonic marine γ-proteobacteria of the SAR86 group (2). Biochemical analysis of this PR protein demonstrated its ability to bind retinal and to function as a light-driven proton pump when expressed heterologously in *Escherichia coli* (2). Biophysical analysis of the γ-proteobacterial PR verified its function as an ion pump and not a sensory rhodopsin as judged by its relatively fast photocycle, comparable to that found in haloarchaeal bacteriorhodopsin (2). Later studies showed that native PR was present in seawater of Monterey Bay, and that the native photoprotein was similar in its spectral properties and photocycle to the expressed photoprotein encoded on the original, cloned PR gene (3).

Subsequent PCR-based gene surveys have identified a wide variety of PR genes in samples from Monterey Bay (Eastern Pacific), the Hawaii Ocean Time-series (HOT) ALOHA (A Long-term Oligotrophic Habitat Assessment) station (Central North Pacific), the Antarctic Peninsula (Southern Ocean), and most recently the Mediterranean Sea and Red Sea (3–5). Spectroscopic analyses indicate that some PR variants are spectrally tuned to absorb different wavelengths of light: PRs from Monterey Bay and surface waters of the Central North Pacific absorbed maximally at 520 nm [green-absorbing proteorhodopsin (G-PR)] whereas PRs from Antarctica and from greater depths in the Central North Pacific were blue shifted and had a peak absorption at 480 nm [blue-absorbing proteorhodopsin (B-PR)] (3). Recently, similar results have been found for PR genes PCR amplified from the Eastern Mediterranean and Red Sea bacterioplankton (4, 5). Recent work has further shown that a single amino acid residue change can shift the PR absorption maximum dramatically (5). This same amino acid residue change has been observed in natural PR variants having different absorption spectra (5). Although the relationships between PR spectral properties, sequence characteristics, and environmental distributions are complex (4, 5), in the Central North Pacific, there was a correlation between depth and PR distributions, with the blue-absorbing types being found at greater depth (3).

Despite the considerable sequence diversity reported among naturally occurring PR genes (3–5), whether these photoproteins

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: PR, proteorhodopsin; B-PR, blue-absorbing PR; G-PR, green-absorbing PR; BAC, bacterial artificial chromosome.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. AF279106 and AY372452–AY372455).

[‡]Present address: Department of Civil and Environmental Engineering, University of Washington, Seattle, WA 98195.

[††]To whom correspondence should be addressed. E-mail: delong@mbari.org.

© 2003 by The National Academy of Sciences of the USA

originate from a single taxonomic lineage, or instead are derived from divergent bacterial taxa, is unknown. PR was originally discovered on a contiguous genomic fragment that also contained an rRNA operon, identifying the host organism as a γ-proteobacterium related to the SAR86 group (2). To date, however, the SAR86 group is the only phylogenetically identified source of bacterial PR.

To better assess the biology of the bacterial PRs, we compared the genomic context of divergent PR genes (recovered from large-fragment genome libraries prepared from bacterioplankton in Antarctic and Hawaiian waters) with the original SAR86 PR-containing genome fragment. The results of our comparative genomic analyses provide insight into the origins and organismal distributions of PR genes in the environment, and demonstrate the existence of PR genes in widely different taxa of marine bacteria.

Materials and Methods

Large-Insert Genomic Library Construction from Environmental DNA. Three different large-insert environmental genomic libraries were screened to identify the clones used in the present study. The first is a fosmid library prepared from Antarctic surface water collected in August 1996, and has been described (6). This library was screened for PR-containing clones by multiplex PCR amplification of PR genes, as described below. The second library consisted of bacterial artificial chromosome (BAC) clones constructed from a plankton sample collected from surface waters at the HOT station (7) in December 2001. BAC library construction was performed as described (8). Briefly, 2,000 liters of sea water was collected from the surface, prefiltered through a 1.8-μm glass fiber filter, and concentrated by tangential flow filtration by using an Amicon DC-10 unit equipped with a 30,000-Da-cutoff polysulfone hollow-fiber filter cartridge (6). Bacterioplankton cells were further concentrated shipboard by a second ultrafiltration of the Amicon retentate on a Millipore Pellicon 30,000-Da-cutoff 50-cm^2 filter, by using TFF tangential flow filtration system. The bacterioplankton pellet was embedded in agarose plugs, and DNA was extracted and cloned into the pIndigoBAC536 vector as described (8, 9). The HOT BAC library was screened by both multiplex PCR and macroarray colony hybridization by using nick-translated PR probes, as described below. Finally, BAC clone EB000-45B06 (a contig to the original PR-containing clone, EBAC31A08; ref. 2) was identified by multiplex PCR screening using PCR primers targeting the 5′ terminus of EBAC31A08 (8).

Multiplex PCR Screening for PR Genes. Multiplex PCR screening of the Antarctic fosmid library was carried out as described (9). PCR primers used to amplify B-PR genes were RhodPal-fwd (5′-TCC AGC AGG ATA AAT TGC CC-3′) and RhodPal-rev (5′-TTT AAG AAG CTT CTA GCT GG-3′). In addition, PCR primers were designed to amplify ≈500 bp at each end of the PR-containing genomic clones ANT32C12 and EBAC31A08, and used to screen for overlapping clones. In each case, the initial amplification products were sequenced by using BIGDYE v.3 dideoxy-nucleotide terminator chemistry on an ABI3100 automated capillary sequencer (Applied Biosystems).

Hybridization Screening for Deeply Divergent PR Genes. To identify genomic fragments containing deeply divergent PR genes, the HOT library was also screened by low stringency hybridization with a B-PR probe. First, the entire HOT library was arrayed onto nylon filters by Amplicon Express (Pullman, WA). Each 22 × 22-cm filter contained 9,216 individual clones arrayed in duplicate. To generate the probe, the B-PR gene from ANT32C12 was first amplified by using the PCR primers described above. The resulting amplification product was then labeled with fluorescein by using the Random Prime Labeling Kit from Amersham Pharmacia Biosciences according to the manufacturer's instructions. The filters were then probed by adding the entire labeling reaction to 50 ml of the recommended hybridization buffer and hybridizing in roller bottles overnight at 45°C. After hybridization, filters were washed first in 1× and then in 0.5× wash buffer for 15 min each at 45°C according to the manufacturer's instructions. Positive clones were detected by using the ECF Signal Amplification module (Amersham Pharmacia Biosciences) according to the manufacturer's instructions, and visualized on a Molecular Dynamics FluorimagerSI (Amersham Pharmacia Biosciences). After hybridization, positive clones were grown, and BAC DNA was prepared as described (8). BAC clones that failed to amplify with PR primers were then restricted by using *Hin*dIII, and the resulting restriction fragments were separated by electrophoresis on a 1% agarose gel. Restriction fragments containing the PR gene were identified by Southern blotting (10) using the conditions described above, subcloned into pBluescript II and sequenced by using BIGDYE v.3 dideoxy-nucleotide terminator chemistry on an ABI3100 automated sequencer (Applied Biosystems).

Sequencing and Annotation of PR-Containing Genomic Fragments. For sequencing of the ANT32C12 and ANT8C10 genomic fragments, fosmid DNA was prepared from a 100-ml liquid culture using a Qiagen (Valencia, CA) midiprep kit. Plasmid DNA was treated with an ATP-dependent Exonuclease (Epicentre Technologies, Madison, WI) according to the manufacturer's instructions for 24 h at 37°C to remove contaminating *E. coli* chromosomal DNA. A shotgun subclone library was then prepared from the plasmid DNA by using the Invitrogen TOPO Shotgun cloning kit (Invitrogen). Fosmid subclones were purified by using the Montage plasmid DNA kit (Millipore) according to the manufacturer's instructions, and sequenced by using BIGDYE v.3 dideoxy-nucleotide terminator chemistry on an ABI3100 automated sequencer (Applied Biosystems). Sequence assembly and editing were performed by using SEQUENCHER software (Gene Codes, Ann Arbor, MI). BAC clone HOT2C01 was sequenced by The Institute for Genomic Research (TIGR, Rockville, MD) production sequencing facility (http://www.tigr.org/tdb/MBMO/closure_method.shtml). Analysis of the potential genes and protein-coding regions was performed by using a combination of the BLAST (11), GLIMMER 2.02 (TIGR) (12, 13), FGENESB (Softberry, Mount Kisco, NY), and ARTEMIS (Sanger Center, Cambridge University, U.K.) (14) software packages.

Phylogenetic Analysis. Phylogenetic trees were constructed by using the maximum parsimony (MP) and minimal evolution (ME) phylogenetic inference methods with the PAUP* version 4.0b10 (15) software package. All heuristic searches were unrooted and performed with random, stepwise addition of taxa with the tree bisection/reconnection branch-swapping algorithm. The validity of inferred topologies was determined by bootstrap analysis with resampling performed on 100 or 1,000 replicates in each analysis. Consensus trees were determined from non-resampled data by majority rule.

Nucleotide Sequence Accession Numbers. Nucleotide sequences from these studies were deposited in GenBank with accession numbers AF279106 and AY372452–AY372455.

Results

Identification of a Monterey Bay Bacterioplankton BAC Clone Contiguous to EBAC31A08. To expand the amount of genomic sequence information available for PR-containing γ-proteobacteria in the SAR86 group, we screened a BAC library derived from Monterey Bay surface water bacterioplankton (2, 8). Clones that overlapped with the PR-containing BAC clone EBAC31A08 were identified by multiplex PCR (9) using primers that amplified the EBAC31A08 insert termini. A single BAC clone (EB000-45B06) was identified and sequenced and was shown to

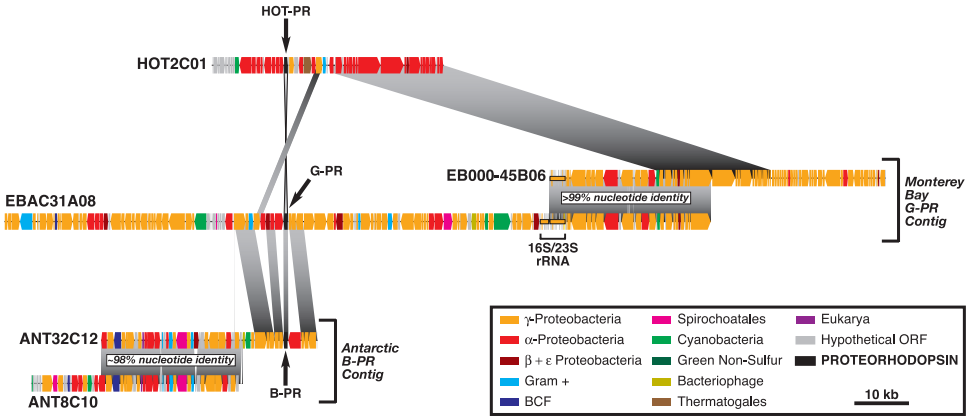


Fig. 1. Gene content and organization of PR gene-containing genomic fragments recovered from environmental samples. The alignment of the genome fragments is based on the position of the PR gene. Regions of similarity in gene content and organization between different genome fragments are connected by shaded areas. BCF, *Bacteroides-Cytophagales-Flavobacterium* group; Gram +, Gram-positive bacteria. Each colored bar represents one ORF. The color of each bar depicts the phylogenetic origin of the highest ranking BLAST match to each predicted ORF (see Tables 1–5). Small insertions/deletions in the highly conserved overlaps between clones ANT32C12 and ANT8C10, and also between EBAC31A08 and EB000-45B06, are indicated. Differences in the lengths of homologous regions between the different clones are due to variations in the size of the orthologous genes. HOT2C01, HOT BAC clone that encodes a PR gene; EBAC31A08, Monterey Bay BAC clone that encodes both a PR gene and an rRNA operon; EB000-45B06, Monterey Bay BAC clone that overlaps EBAC31A08; ANT32C12, Antarctic fosmid clone that encodes a PR gene; ANT8C10, Antarctic fosmid clone that overlaps with ANT32C12; HOT-PR, HOT proteorhodopsin gene. (Scale bar = 10 kbp.)

contain overlapping sequence with EBAC31A08 (Fig. 1). Sequence comparisons between EBAC31A08 and EB000-45B06 showed >99% nucleotide sequence identity over a >30-kb region of sequence overlap (Fig. 1).

Characterization of a PR-Containing Fosmid from the Antarctic Peninsula. To identify genome fragments containing bacterial B-PR genes, a fosmid library derived from bacterioplankton collected in surface water (−1.8°C) near Palmer Station in the Antarctic Peninsula (6) was screened via multiplex PCR (9) using nondegenerate B-PR-specific oligonucleotide primers. Five individual Antarctic fosmid clones containing a B-PR gene were identified, and one (ANT32C12) was fully sequenced. Fosmid ANT32C12, which contained a B-PR gene 97% identical at both the nucleotide and amino acid levels to previously described Antarctic B-PR sequences (3), was selected for complete sequencing. In addition, multiplex PCR screening of the Antarctic bacterioplankton library using primers amplifying the ends of the ANT32C12 insert identified another fosmid clone (ANT8C10) with significant overlap to ANT32C12. Sequence comparison of the overlapping region between ANT32C12 and ANT8C10 revealed complete synteny, with the exception of a single gene insertion in ANT32C12 (Fig. 1). These clones were ≈98% identical over this entire 25-kb region, which terminates ≈10 kb upstream of the PR gene (Fig. 1).

Sequence analysis identified 40 potential protein-coding sequences on the 39.4-kb insert of ANT32C12 and 39 potential genes on the 37.9-kb insert of ANT8C10 (Tables 1 and 2, which are published as supporting information on the PNAS web site, www.pnas.org.). The 25-kb region of overlap between ANT32C12 and ANT8C10 contained 29 predicted ORFs. Two small insertions (134 bp and 241 bp) in this region in ANT8C10 do not change the predicted gene order or identity (Fig. 1). However, a 778-bp deletion in ANT8C10 eliminates one potential ORF present in ANT32C12 in the region of overlap (ANT32C12.23; Tables 1 and 2). Predicted protein-coding sequences shared by both fosmids in the region of overlap are 98–100% identical to each other at both the nucleotide and amino acid levels. In total, 53 individual ORFs were predicted from the composite ≈52-kb genome fragment composed of ANT32C12 and ANT8C10. Of these putative genes, 30 (57%) can be assigned a clear function based on their similarity to functionally characterized protein sequences in GenBank. Another 14 (26%) share high sequence similarity to conserved hypothetical proteins in other organisms, but could not be assigned a biochemical function. Finally, 9 putative coding sequences (17%) had no significant matches (E values $> 1 \times 10^{-4}$) to protein sequences in the public databases.

None of the predicted proteins encoded by ANT32C12/ANT8C10 could be confidently used to determine the phylogenetic identity of the B-PR containing Antarctic bacterium. However, based on the highest scoring BLASTP match (11), 58% of the predicted proteins encoded by ANT32C12/ANT8C10 were most highly similar to proteobacterial homologues (Fig. 1 and Tables 1 and 2). Thirty-six percent of the predicted protein sequences were most similar to γ-proteobacterial proteins whereas 18% bore highest similarity to α-proteobacterial protein sequences. In contrast, a higher proportion of the proteins encoded on EBAC31A08 (that also encoded the SAR86-like rRNA operon), shared greatest sequence similarity with proteins derived from γ-proteobacteria (69%), compared with α-proteobacteria (11%) (Fig. 1). The remaining ORFs from ANT32C12/ANT8C10 shared highest similarity to proteins from a number of different bacterial phylogenetic divisions, including Gram-positive bacteria (11%), ε-proteobacteria (4%), and cyanobacteria (2%).

Characterization of a PR-Containing BAC from Central North Pacific Surface Waters. To broaden the search for bacterial PRs, a BAC library was prepared from surface waters at the HOT station in the subtropical Central North Pacific Ocean. Subsequently, high-density colony arrays containing all of the clones in this library were screened by using low-stringency hybridization conditions with a B-PR targeted probe. Three clones yielded positive hybridization signals with the B-PR probe. None of these clones could be amplified by using standard PR PCR primers whereas one clone (HOT2C01) was tested positive in Southern blot experiments using the labeled B-PR probe. Subcloning and sequencing of the HOT2C01 PR gene revealed that its DNA sequence was significantly different from the originally described PR sequences (Fig. 2). Sequence analysis of the 42.2-kb insert of clone HOT2C10 indicated the presence of 41 predicted ORFs (Fig. 1). Of these potential genes, 28 (68%) could be assigned a function based on sequence homology to proteins in public

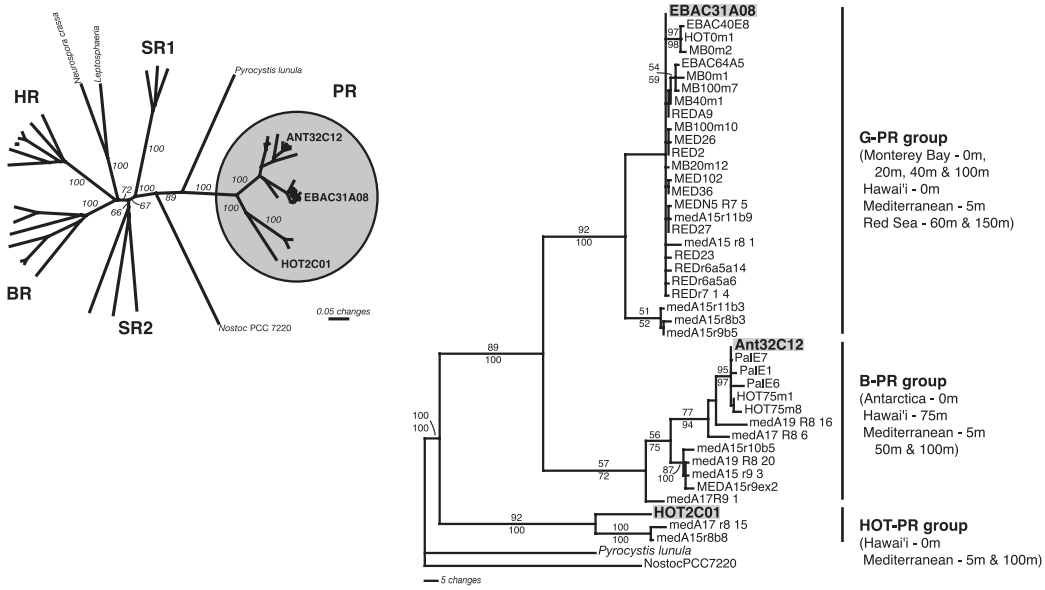


Fig. 2. Phylogenetic analysis of archaeal and bacterial rhodopsin family amino acid sequences. (*Left*) Phylogenetic analysis of PR, and archaeal [bacteriorhodopsin (BR), halorhodopsin (HR), and sensory rhodopsin (SR)], bacterial (*Nostoc* sp. PCC 7220), and eucaryal (*Neurospora crassa*, *Leptosphaeria maculans*, and *Pyrocystis lunula*) rhodopsin amino acid sequences. Minimal evolution inference method based on an analysis of 167 characters illustrates the relationship of the PR sequences to other rhodopsin sequences. Sequences indicated in bold represent PR sequences derived from the environmental genome fragments analyzed in this study. Bootstrap values, given as percentages from 1,000 replicate trees, are indicated for branches supported by >50% of the trees. Shown are SR groups 1 and 2, BR, and HR. Sequences used for the construction of this tree were derived from *Halorubrum sodomense*, *Halobacterium salinarum*, *Haloarcula vallismortis*, *Natronomonas pharaonis*, *Halobacterium* sp., *Neurospora crassa*, *Leptosphaeria maculans*, *Nostoc* PCC 7220, and *Pyrocystis lunula*. (*Right*) Phylogenetic analysis of 167 characters from selected PR amino acid sequences showing the detailed relationship among the three PR variants examined in this study (indicated in bold) and those previously reported (refs. 2–5). Bootstrap values for maximum parsimony (upper) and minimal evolution (lower) analysis, given as percentages from 1,000 replicate trees, are indicated for branches supported by >50% of the trees. Amino acid sequences for the rhodopsin genes from *Nostoc* sp. PCC 7220 and *Pyrocystis lunula* were used as the outgroup.

databases, 3 (7%) show homology to hypothetical proteins of unknown function in other microbial genomes, and 10 (24%) have no significant similarity to any known protein sequence.

Many of the predicted proteins found on HOT2C01 contained significant phylogenetic information (Table 3, which is published as supporting information on the PNAS web site). These predicted proteins included several large-subunit and small-subunit ribosomal proteins, homologues of the translation elongation factors EF-G and EF-Tu, and two subunits (RpoB and RpoC) of RNA polymerase (Table 3, and see Tables 4 and 5, which are published as supporting information on the PNAS web site). Phylogenetic analyses of the predicted ribosomal proteins (analyzed as a single concatamer of 13 ribosomal proteins; Fig. 3*A*) and of the RNA polymerase beta′ subunit sequence (Fig. 3*B*) indicated that the bacterium harboring the HOT2C01 PR was an α-proteobacterium. In addition, the majority of highest scoring BLASTP matches (11) of the entire suite of predicted protein sequences on HOT2C01 were affiliated with α-proteobacterial homologues ($\approx$68%) whereas only a small percentage ($\approx$7%) grouped with γ-proteobacterial homologues (Fig. 1 and Tables 3, 4, and 5). In several cases, protein orthologues present on HOT2C01 and EBAC31A08-EBAC45B06 allowed direct phylogenetic comparisons, demonstrating the disparate origins of these bacterial genome fragments. Together, these results strongly suggest that the HOT2C01 genome fragment is derived from a planktonic marine α-proteobacterium.

Phylogenetic Analyses of PR Sequences. Comparison of the PR sequence contained on ANT32C12 with previously reported PR sequences (3, 4) indicated a close relationship to other PCR-amplified B-PR sequences from Antarctic surface waters and deeper waters of the North Pacific. ANT32C12 PR shared 97–99% sequence identity in both the amino acid and nucleotide sequence to known B-PR sequences whereas it shared only $\approx$80% identity to G-PR sequences. In contrast, HOT2C01 PR shared only $\approx$60% amino acid sequence identity with either G-PR or B-PR. This sequence divergence suggests that the HOT2C01-encoded PR may have dramatically different spectral and/or biological properties compared with other characterized G-PR and B-PR proteins.

Phylogenetic analyses of PR sequences (Fig. 2) further indicated the divergent nature of the HOT2C01 PR. In phylogenetic analyses of both amino acid and nucleotide sequences, ANT32C12 PR associated at high confidence with known B-PR sequences (3, 4). In contrast, phylogenetic analysis revealed that HOT2C01 PR represents one of the most divergent PR sequences identified to date. In phylogenetic analyses that included sequences representing the entire superfamily of microbial rhodopsin proteins, the HOT2C01 PR was affiliated with one of the deepest branching nodes in the PR group (Fig. 2*A*). The HOT2C01 PR branches deeply in phylogenetic trees of known PR sequences (Fig. 2*B*) and is related to a group of PRs recently identified in PCR-based PR gene surveys in the Mediterranean and Red Seas (4). Comparisons between these new PR sequences and HOT2C01 PR, however, revealed that they share only $\approx$70% sequence identity. These data suggest that this deeply branching PR clade may encompass proteins as different from one another other as the G-PRs and B-PRs are.

Although nearly all other PR sequences reported so far display evidence for the presence of a signal sequence, the predicted HOT2C01 PR amino acid sequence revealed no such signal. Such sequences are required in Eucaryotes for the proper insertion of proteins into the plasma membrane. However, significant evidence suggests that such signals are not an absolute requirement for secretion and targeting of proteins to the bacterial membrane (16). Comparison of the amino acid sequences of HOT2C01 PR and known PRs reveals that the

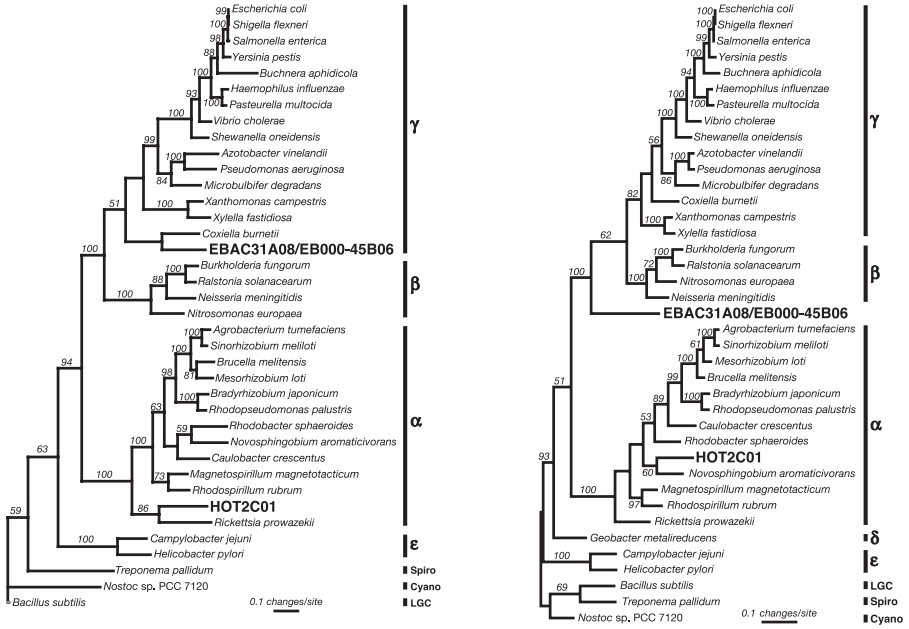


Fig. 3. Phylogenetic analyses of ribosomal proteins and RNA polymerase genes indicate that HOT2C01 is derived from an α-proteobacterium. (*Left*) Maximum parsimony analysis of 13 concatemerized ribosomal protein sequences demonstrating the relationship of the HOT2C01-encoded sequences to other α-proteobacterial genes. Amino acid sequences for ribosomal proteins L1, L10, L11, L2, L22, L23, L3, L4, L7/L12, S10, S12, S3, and S7 from HOT2C01, the EBAC31A08/EB000-45B06 contig, and selected organisms from completed genome projects were aligned individually, masked to include only homologous residues, and then concatenated to create a single sequence with 2,427 characters. Bootstrap values, given as percentages from 1,000 replicate trees, are indicated for branches supported by >50% of the trees. Sequences from *Bacillus subtilis*, *Treponema pallidum*, and *Nostoc* PCC 7120 were used as outgroups. (*Right*) Maximum parsimony analysis of RNA polymerase subunit beta′ (RpoC) amino acid sequences indicates that HOT2C01 is derived from the genome of an unknown marine α-proteobacterium. Bootstrap values, given as percentages from 1,000 trees based on the resampling of 1,420 characters, are indicated for branches supported by >50% of the trees. Sequences from *Bacillus subtilis*, *Treponema pallidum*, and *Nostoc* PCC 7120 were used as outgroups.

variable residues in HOT2C01 PR are spread throughout the protein, including transmembrane domains. Nevertheless, there is strong conservation of the residues predicted to form the retinal binding pocket, with 18 of the 20 residues identical between HOT2C01 PR and EBAC31A08 G-PR. One important exception is the glutamine residue at position 88 in HOT2C01 PR (leucine 105 in EBAC31A08 G-PR), which has been associated with the blue-shifted absorption spectrum of B-PRs (5) and archaeal bacteriorhodopsins (17). The presence of glutamine at position 88 strongly indicates that the HOT2C01 PR protein will have a blue-shifted absorption spectrum, similar to other BPRs ($Abs_{max} \approx 490$ nm) when bound to retinal (3, 5).

Comparative Analyses of PR-Containing Genome Fragments. There was no universal synteny in regions directly flanking the PR gene between all three genome fragments. Comparison of the sequences of ANT32C12 to EBAC31A08 revealed an ≈10-kb region flanking the PR gene in which gene identity and order seem to be conserved (Fig. 1). The region of similarity contained 12 predicted genes on EBAC31A08, and 10 on ANT32C12. Two large insertions in EBAC31A08 encompassed three predicted genes not present in ANT32C12 whereas an insertion in ANT32C12 encoded a predicted protein sequence not found in EBAC31A08. Nine predicted genes are shared between the two genome fragments. Comparison of the predicted amino acid sequences of the shared ORFs revealed sequence identities ranging from 42% to 86%. Given that the region of similarity between ANT32C12 and EBAC31A08 is found on the 3′ terminus of ANT32C12, it remains possible that this genomic synteny extends beyond the ≈10-kb overlap reported here. No similarity of gene order or gene content was detected between ANT32C12/ANT8C10 and HOT2C01.

Very little similarity was observed between HOT2C01 and EBAC31A08 in the region flanking the PR gene (Fig. 1). In addition to the PR gene, the only evident homologue was a putative NAD-dependent formate dehydrogenase gene (HOT2C01.21; Table 3) found ≈5 kb upstream of the HOT2C01 PR locus (Fig. 1). In EBAC31A08, the equivalent gene was located ≈4 kb upstream of the G-PR gene and was oriented in the opposite direction compared with the HOT2C01 homologue orientation (Fig. 1). The predicted amino acid sequence for this gene is 79% identical to the similar protein-coding sequence on EBAC31A08 (11), indicating that these two predicted proteins are more similar to each other than to any other homologous sequence currently in GenBank. Just beyond the NAD-dependent formate dehydrogenase gene lies a large region of conserved genes found in all bacterial genomes that facilitate inference of the phylogenetic relationships between HOT2C01 and EBAC31A08. Most of the genes in this region belong to operons of large- and small-subunit ribosomal proteins, that share a high conservation of gene content and order in the genomes of all currently sequenced proteobacteria. Also included in this region are genes encoding RNA polymerase subunits β and β′, two translation elongation factors, and a transcriptional anti-terminator.

Discussion

Cultivation-independent molecular surveys previously revealed a wide variety of PR sequences in marine planktonic bacteria (2–4). Biochemical characterization of several of these PR proteins has shown them to be spectrally tuned to the light available in different environments, demonstrating a significant degree of functional diversification as well (2–5). These findings raise interesting questions about the organismal distribution of these diverse PR genes among planktonic marine bacteria. To date, only the first PR sequence to be characterized (EBAC31A08) can be assigned to a specific bacterial group

because it is contiguous with a robust phylogenetic marker, an rRNA operon. All other PR sequences reported have been recovered as PCR amplicons, and so the origins of these sequences remain unknown. To better understand the organismal distribution of diverse PRs, we compared large genome fragments recovered from marine bacterial populations that contained different classes of PR genes.

Analyses of these naturally occurring genome fragments provided strong evidence that PRs are widely distributed among several different bacterial taxa. Even though the PR-containing genome fragments reported here do not contain rRNA genes, analyses of flanking protein-encoding genes allowed preliminary assessment of the phylogenetic affiliation of different PR-containing bacteria. In the case of ANT32C12, analysis of gene content, order, and arrangement, as well as examination of the origin of the highest scoring BLASTP matches (11), provided relevant information toward determining the organism's phylogenetic affiliation (Figs. 1 and 3). The limited, but nonetheless significant synteny between ANT32C12 and the SAR86-derived PRs, as well as the affiliation of most ORFs to homologues from the γ-proteobacterial subdivisions, indicates that the ANT32C12/ANT8C10 PR is most likely derived from a γ-proteobacterium only distantly related to the SAR86 group. Although not definitive, these data provide evidence that G-PRs and B-PRs originate from different organisms, and also point out the strengths and weaknesses of the environmental genomic approach. On some genomic fragments, the phylogenetic signature (such as that derived from ribosomal proteins) may be sufficient to infer taxonomic origins of the DNA. In other genome fragments, the encoded genes may provide little cumulative phylogenetic signal, so the DNA's origin will remain a matter of speculation until more data are acquired.

The HOT2C01 PR-containing BAC and the EBAC31A08 PR-containing BAC both encoded a variety of more phylogenetically informative ORFs. These conserved genes included several ribosome-associated proteins, translation elongation factors, and subunits of RNA polymerase. The concatenated ribosomal protein phylogeny for EBAC31A08 placed it in the γ-proteobacteria, consistent with its rRNA-inferred phylogenetic origins (2). The RNA polymerase beta subunit phylogeny for EBAC31A08 was less consistent, probably an artifact of the small taxon samples available for these genes, compared with 16S rRNA. Phylogenetic analyses of the predicted proteins of HOT2C01 (Fig. 3) suggest that this PR-bearing genome fragment is derived from an abundant, planktonic α-proteobacterium. To date, the most abundant marine α-proteobacteria have been characterized almost exclusively by their 16S rRNA sequences, and therefore could not be included in the analyses. As most of these abundant marine α-proteobacteria are not closely related to cultivated representatives, we were unable to determine definitively whether the HOT2C01 organism might belong to previously identified α-proteobacteria abundant in marine plankton, such as the *Pelagibacter* (SAR11) or *Roseobacter* groups (18). As whole genome sequences for the growing number of cultivable microorganisms become more available (19), the resolving power of these comparative genomics approaches will continually improve.

These genomic data demonstrate that PRs are widely distributed among divergent bacterial taxa in the marine environment. It was possible to determine the general phylogenetic affiliation of microbes harboring some of these PR genes by virtue of genetically linked, highly conserved ribosomal protein genes. The variety of PR sequences previously observed in PCR-based surveys, and the different organismal origins of these photoproteins reported here, may be the result of speciation events, functional diversification of paralogous genes, or lateral gene transfer (or some combination of these processes). Additional genomic studies of PR genes in uncultivated marine bacteria, as well as those brought into cultivation through advanced cultivation techniques (19), should help to better define the mechanisms underlying PR diversification.

More definitive physiological experiments with PR-containing bacteria still remain to be performed. However, the available evidence suggests that retinal-binding photoproteins may have substantial adaptive significance for marine bacteria inhabiting sunlit waters of the global ocean. The evidence includes the following: (*i*) PR-bearing bacteria are themselves ubiquitous and abundant in the marine environment; (*ii*) a considerable variety of PRs have evolved and are maintained in the planktonic marine bacterial assemblages; (*iii*) at least some characterized PR genes function as efficient light-driven proton pumps and are capable of generating a chemiosmotic potential, analogous to haloarchaeal bacteriorhodopsin; and (*iv*) we show here that PR genes are distributed among divergent bacterial lineages that dwell in the marine photic zone. All of the above facts strongly suggest that potent selective forces, most likely acting on light-activated metabolic processes, are driving PR function, evolution, maintenance, and distribution. These observations are entirely consistent with PR's postulated photophysiological role in the bioenergetics of oceanic bacterioplankton worldwide.

We thank the captain and crew of the R/V *Point Lobos*, the R/V *Ka'imikai-o-Kanaloa*, and the Hawaii Ocean Time support staff for expert assistance at sea. This work was supported by the David and Lucile Packard Foundation and National Science Foundation Microbial Observatory Grant MCB-0084211 (to J.H. and E.F.D.). The HOT program is supported by National Science Foundation Grant OCE-9617409 (to D.M.K.).

1. Spudich, J. L., Yang, C. S., Jung, K. H. & Spudich, E. N. (2000) *Annu. Rev. Cell Dev. Biol.* **16,** 365–392.
2. Béjà, O., Aravind, L., Koonin, E. V., Suzuki, M. T., Hadd, A., Nguyen, L. P., Jovanovich, S. B., Gates, C. M., Feldman, R. A., Spudich, J. L., *et al.* (2000) *Science* **289,** 1902–1906.
3. Béjà, O., Spudich, E. N., Spudich, J. L., Leclerc, M. & DeLong, E. F. (2001) *Nature* **411,** 786–789.
4. Sabehi, G., Massana, R., Bielawski, J. P., Rosenberg, M., DeLong, E. D. & Béjà, O. (2003) *Environ Microbiol.* **5,** 842–849.
5. Man, D., Wang, W., Sabehi, G., Aravind, L., Post, A. F., Massana, R., Spudich, E. N., Spudich, J. L. & Béjà, O. (2003) *EMBO J.* **22,** 1725–1731.
6. Béjà, O., Koonin, E. V., Aravind, L., Taylor, L. T., Seitz, H., Stein, J. L., Bensen, D. C., Feldman, R. A., Swanson, R. V. & DeLong, E. F. (2002) *Appl. Environ. Microbiol.* **68,** 335–345.
7. Karl, D. M. & Lukas, R. (1996) *Deep-Sea Res.* **43,** 129–156.
8. Béjà, O., Suzuki, M. T., Koonin, E. V., Aravind, L., Hadd, A., Nguyen, L. P., Villacorta, R., Amjadi, M., Garrigues, C., Jovanovich, S. B., *et al.* (2000) *Environ. Microbiol.* **2,** 516–529.
9. Stein, J. L., Marsh, T. L., Wu, K. Y., Shizuya, H. & DeLong, E. F. (1996) *J. Bacteriol.* **178,** 591–599.
10. Sambrook, J., Fritsch & Maniatis, T. (1990) *Molecular Cloning: A Laboratory Guide* (Cold Spring Harbor Lab. Press, Plainview, NY).
11. Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25,** 3389–3402.
12. Delcher, A. L., Harmon, D., Kasif, S., White, O. & Salzberg, S. L. (1999) *Nucleic Acids Res.* **27,** 4636–4641.
13. Salzberg, S. L., Delcher, A. L., Kasif, S. & White, O. (1998) *Nucleic Acids Res.* **26,** 544–548.
14. Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M. A. & Barrell, B. (2000) *Bioinformatics* **16,** 944–945.
15. Swofford, D. L. (2001) PAUP*, Phylogenetic Analysis Using Parsimony (*and Other Methods) (Sinauer, Sunderland, MA), Version 4.0b10.
16. Broome-Smith, J. K., Gnaneshan, S., Hunt, L. A., Mehraein-Ghomi, F., Hashemzadeh-Bonehi, L., Tadayyon, M. & Hennessey, E. S. (1994) *Mol. Membr. Biol.* **11,** 3–8.
17. Subramaniam, S., Greenhalgh, D. A., Rath, P., Rothschild, K. J. & Khorana, H. G. (1991) *Proc. Natl. Acad. Sci. USA* **88,** 6873–6877.
18. Giovannoni, S. J., Britschgi, T. B., Moyer, C. L. & Field, K. G. (1990) *Nature* **345,** 60–63.
19. Rappé, M. S., Connon, S. A., Vergin, K. L. & Giovannoni, S. J. (2002) *Nature* **418,** 630–633.