Actinorhodopsins: proteorhodopsin-like gene sequences found predominantly in non-marine environments

Adrian K. Sharma,¹* Olga Zhaxybayeva,¹ R. Thane Papke² and W. Ford Doolittle¹
¹Department of Biochemistry and Molecular Biology, Dalhousie University, 5850 College St., Halifax, Nova Scotia, Canada, B3H 1X5.
²Department of Molecular and Cell Biology, University of Connecticut, 91 North Eagleville Rd., Storrs, CT 06269-3125, USA.

Summary

Proteorhodopsins are light-energy-harvesting transmembrane proteins encoded by genes recently discovered in the surface waters of the world's oceans. Metagenomic data from the Global Ocean Sampling expedition (GOS) recovered 2674 proteorhodopsin-related sequences from 51 aquatic samples. Four of these samples were from non-marine environments, specifically, Lake Gatun within the Panama Canal, Delaware Bay and Chesapeake Bay and the Punta Cormorant Lagoon in Ecuador. Rhodopsins related to but phylogenetically distinct from most sequences designated proteorhodopsins were present at all four of these non-marine sites and comprised three different clades that were almost completely absent from marine samples. Phylogenomic analyses of genes adjacent to those encoding these novel rhodopsins suggest affiliation to the Actinobacteria, and hence we propose to name these divergent, non-marine rhodopsins ‘actinorhodopsins’. Actinorhodopsins conserve the acidic amino acid residues critical for proton pumping and their genes lack genomic association with those encoding photo-sensory transducer proteins, thus supporting a putative ion pumping function. The ratio of recA and radA to rhodopsin genes in the different environment types sampled within the GOS indicates that rhodopsins of one type or another are abundant in microbial communities in freshwater, estuarine and lagoon ecosystems, supporting an important role for these photosystems in all aquatic environments influenced by sunlight.

Introduction

Metagenomic data have already provided tremendous insight into the diversity, evolution and ecology of the uncultivated majority of microorganisms (Handelsman et al., 1998; Hallam et al., 2004; Tyson et al., 2004; Legault et al., 2006). The Global Ocean Sampling expedition (GOS), undertaken by the J. Craig Venter Institute with support from the Gordon and Betty Moore Foundation, increases by 6.3 billion base pairs the breadth and depth of such data, and can be bioinformatically mined for knowledge about the geographic distribution of genes and organisms in marine environments (Rusch et al., 2007; Yoospeh et al., 2007). One family of genes of particular interest, and the focus of the work described here, are the light-energy-harvesting microbial rhodopsins (Beja et al., 2000; Gomez-Consarnau et al., 2007).

All rhodopsins are photo-reactive seven-pass transmembrane proteins covalently linked to the chromophore retinal by a lysine residue in the seventh helix (Spudich et al., 2000). The microbial rhodopsins are classified as type 1 according to conserved residues in the retinal binding pocket that distinguish them from the (possibly non-homologous) type 2 rhodopsins present in non-microbial eukaryotes, for instance, in the vertebrate retina (Spudich et al., 2000). There are two general kinds of functions served by microbial rhodopsins: light-driven ion pumping, and light-sensing activities such as phototaxis or photoadaptation (Spudich and Jung, 2005). Characterized microbial rhodopsins also appear to comprise two types phylogenetically (Sharma et al., 2006): the archaeal proteins long studied by biophysicists (Oesterhelt and Stoeckenius, 1973; Matsuno-Yagi and Mukohata, 1977; Bogomolni and Spudich, 1982; Schober and Lanyi, 1982; Spudich and Spudich, 1982) and the ‘proteorhodopsins’. The first of these latter was detected by Beja and colleagues (2000), as an open reading frame in a marine metagenomic BAC clone, assignable through its encoded 16S rRNA gene to the γ-proteobacteria. As the initial discovery of proteorhodopsin, this sub-family name has expanded to include many bacterial
rhodopsins of marine origins (Gomez-Consarnau et al., 2007; McCarren and DeLong, 2007). For example, genes that by their sequences belong to the proteorhodopsin clade have since been found in marine archaea (Frigaard et al., 2006). Conversely genes of the archaeal-type rhodopsins are present in some non-marine bacteria (Sharma et al., 2006). However, it is still the case that proteorhodopsins recovered from large insert marine metagenomic libraries (where taxonomic affiliations can be provisionally assigned) or sequenced genomes are usually found in marine γ- and α-proteobacteria, and occasionally in other bacterial phyla such as the Bacteroidetes/Chlorobi group and the Planctomycetes (McCarren and DeLong, 2007). More divergent proteorhodopsin-related sequences originating from non-marine organisms like Gloeobacter violaceus (Nakamura et al., 2003) and Roseiflexus sp. RS-1 (Hanada et al., 2002), bacteria isolated from a calcareous rock face and a Japanese hot spring, respectively, might be considered a distinct subfamily of microbial rhodopsins.

Rusch and colleagues (2007) find 2674 proteorhodopsin-related genes among sequences from the 51 sites surveyed in the GOS, reporting a correlation between environment type and spectral tuning as the Beja et al., 2001) of this light-driven proton pump. Of the 2674 full-length and partial sequences, 1874 contain the amino acid position responsible for spectral tuning to absorb a specific wavelength of light (Rusch et al., 2007). Green-tuned spectral variants (those that encode a leucine at the critical position) were more prevalent in high-productivity coastal and 'non-marine' environments where phytoplankton absorb primarily in the blue and red spectra. In open ocean environments, where the decreased amount of nutrients results in less phytoplankton biomass, the blue-tuned variants (encoding glutamine) dominated. A recent study by Sabehi and colleagues (2007) on the ecology of spectrally tuned proteorhodopsin variants across different seasons and oceanic depths confirmed a similar observation for rhodopsins from the surface water of the Sargasso Sea.

These GOS data allow us to ask whether other variables such as salinity or even simple geographical distance can be correlated with predominant locale-specific proteorhodopsin-related sequence types. Finding that certain sequence types predominate at specific locations could mean that rhodopsins of that type are especially adapted to conditions there. However, it could be other characteristics of the lineage in which these genes are most often found that are responsible for such localization.

The majority of the GOS samples are from marine sites: in a search for novel correlations we thus focused on the three non-marine hyposaline GOS sites, Lake Gatun (referred to in the GOS database as sample GS020), and the estuaries Delaware Bay (GS011) and Chesapeake Bay (GS012), and on the relatively hypersaline Punta Cormorant Lagoon (GS033). Our analyses identified three abundant unique groups of rhodopsin sequences from these four GOS sites, clustering outside the major, highly supported proteorhodopsin clade. All three novel sequence clusters were almost exclusively linked to genes that displayed phylogenetic relation to the Actinobacteria, and we have chosen to call them ‘actinorhodopsins’. Actinorhodopsins encode conserved amino acid residues required for proton pumping hinting at a possible ion transport function. In contrast to haloarchaeal rhodopsins (Sharma et al., 2007) and many proteorhodopsins (Sabehi et al., 2005), actinorhodopsins were not linked to genes involved in the production of retinal nor to genes encoding transducer proteins commonly associated with photo-sensory rhodopsins (Martinez et al., 2007). The presence of rhodopsins, abundant in microbial communities in freshwater, estuarine and lagoon ecosystems, supports an important role for these photosystems in all aquatic environments influenced by sunlight.

Results

Phylogenetic analysis of rhodopsins from the Lake Gatun, Delaware Bay, Chesapeake Bay and the Punta Cormorant Lagoon

We performed phylogenetic analyses on rhodopsins to determine whether sequences from Lake Gatun, Delaware Bay, Chesapeake Bay and the Punta Cormorant Lagoon formed unique clades within the current diversity of ‘proteorhodopsins’ in the non-redundant (nr) database. Rhodopsin sequences from non-marine GOS samples were extracted and included in the phylogenetic tree in Fig. 1, which includes also a selection of non-GOS sequences from the nr database representing marine proteorhodopsins and sequences from bacterial isolates cultivated from a variety of environments. To generate useful phylogenetic trees, only rhodopsin-encoding fragments of ≥ 150 amino acids (aa) were used (see Experimental procedures for more details). The phylogenetic tree presented in Fig. 1 represents 315 sequences clustered into 102 operational taxonomic units (OTUs) that were grouped based on amino acid sequence divergence of a distance ≤ 0.1 (see Experimental procedures).

Figure 1 is well supported by bootstrap analysis and shares features with previously published phylogenetic trees (Sharma et al., 2006; McCarren and DeLong, 2007). For example, the rhodopsin sequence from the marine flavobacterium Polaribacter igrensis forms a well-supported clade with a large group of sequences representing marine α- and γ-proteobacteria (such as Pelagibacter ubique of the SAR11 lineage and the uncultured SAR86 lineage), uncultured marine group II
euryarchaeotes and a collection of environmental sequences from unknown organisms. Overall, this large group of proteorhodopsins represents many sequences and forms a highly supported group to the exclusion of other more divergent microbial rhodopsins, such as those from *Salinibacter ruber*, *G. violaceus* and the marine dinoflagellate *Pyrocystis lunula*. These latter sequences appear to form their own clade intermingled with other divergent rhodopsins from both cultivated organisms and environmental samples, including those from the non-marine sites.

Our analyses of rhodopsin sequences from the metagenomic data from four non-marine GOS sample sites revealed several novel clusters of rhodopsin sequences. Within Fig. 1 highly supported sequence clusters LG1 and LG2 appear unique in the context of the known diversity of rhodopsins and consist of a large number of sequences (95) that originate from GOS sites Lake Gatun, the marine estuaries Delaware Bay, Chesapeake Bay and the Punta Cormorant Lagoon. Only four sequences among the 79 within LG1 originate from marine sample sites in the GOS, while no homologs were found in the nr database. LG1 and LG2 appear highly divergent from other sequences in this part of the phylogenetic tree, where the relationships with their nearest neighbours, *S. ruber*, *G. violaceus* and *Rosellinella* sp. RS-1, are poorly supported by bootstrap analyses. Our analyses also revealed an additional novel cluster of sequences from the Punta Cormorant Lagoon (referred to as PCL1 in Fig. 1), which contains 37 nearly identical sequences. BLASTP searches for PCL1-type sequences within the 51 GOS sampling sites revealed three other sequences that grouped within this clade (Fig. 1 and Table 1).

Outside of the LG1, LG2 and PCL1 sequence clusters, other interesting groupings were also observed for sequences from each of the four sample sites (Fig. 1). For example, 20 sequences that originate from both Lake Gatun and the marine estuary sites form a distinct clade closely related to *P. ubique* and the SAR11-type rhodopsins, which we call ‘SAR11-like-1’. Additionally, another 14 sequences that we called ‘SAR11-like-2’ were present in the estuary sites only, and group much more closely with *P. ubique*. The SAR11-like-2 group was absent from both Lake Gatun and the Punta Cormorant Lagoon samples. PCL2, a cluster of 37 nearly identical sequences from the Punta Cormorant Lagoon, grouped among sequences from marine α-proteobacteria also, but did not display any specific relation to SAR11 types. Also noteworthy were those OTU clusters spread throughout the tree that comprised sequences from GOS samples from different environmental types as well as the nr database.

**Examining the prevalence of LG1 and LG2 sequences in the GOS**

Because sites other than Lake Gatun, Delaware and Chesapeake Bays and the Punta Cormorant Lagoon were only sparsely and selectively sampled for data in preparing Fig. 1, we examined the GOS data much more thoroughly and systematically for the presence of LG1 and LG2 sequences, as shown in Fig. 2. We constructed a phylogenetic tree using all sequences retrieved from BLAST searches of all 51 sample sites using LG-type rhodopsins as queries (Fig. 2 and Fig. S2; see *Experimental procedures* for selection criteria). The phylogenetic tree in Fig. 2 also shows individually all the LG1 and LG2 sequences indicated by the yellow wedges in Fig. 1. In total our BLAST analyses retrieved 146 rhodopsin sequences (including these previously identified LG1 and LG2 examples), 135 of which were alignable by our criteria (see *Experimental procedures*). The phylogenetic tree of these 135 sequences in addition to other related non-GOS rhodopsins is presented in Fig. 2. Some sequences that grouped outside LG clades were also retrieved using our derived bit-score cut-offs and are represented in the phylogenetic tree.

By far, most of the sequences that group within LG1 and LG2 originate from the four non-marine sites on which we had initially focused (Fig. 2). Two sequences within LG1 originate from a warm seep in Roca Redonda, Ecuador (GS030), one from the coast of Florida (GS015) and another from sample GS047, an open ocean environment 201 miles from French Polynesia. A summary of the number of sequences from LG1 or LG2 retrieved from metagenomic samples is presented in Table 1. None of the sequences represented by LG2 were present in samples of marine origin.

Although phylogenetic clustering of sequences based on ecological differences might be expected, both LG clades do contain sequences that derive from a broad

---

**Table 1. BLASTP searches of the GOS metagenomic data to retrieve rhodopsin sequences that group with LG1, LG2 and PCL1**

<table>
<thead>
<tr>
<th>Phylogroup</th>
<th>Sum of sequences retrieved from all 51 sites</th>
<th>Lake Gatun (516 Mbp)</th>
<th>Estuaries (314 Mbp)</th>
<th>Punta Cormorant Lagoon (1200 Mbp)</th>
<th>Coastal or open ocean (9266 Mbp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LG1</td>
<td>93</td>
<td>36</td>
<td>26</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>LG2</td>
<td>23</td>
<td>17</td>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>PCL1</td>
<td>40</td>
<td>0</td>
<td>0</td>
<td>37</td>
<td>3</td>
</tr>
</tbody>
</table>

a. Amount of sequence data in the sample.
range of salinities. For example, the LG1 and LG2 sequences from the freshwater Lake Gatun and the hyposaline estuaries show considerable phylogenetic overlap, but LG1 rhodopsins from both of these environment types are interspersed with those from Punta Cormorant (which is in fact more saline than seawater) and as well include the four marine sequences. This intermingling of sequences in Fig. 2 is contrasted by the presence of well-supported ‘microdiverse’ sequence clusters within either clade that appear to have evolved and radiated according to environment type, although this effect could be negated by more thorough metagenomic sampling.

Given the overlap of rhodopsin sequence types from Lake Gatun and estuarine environments it may seem...
surprising that no similar sequences were retrieved from the Canadian estuary site, the Bay of Fundy (GS006). Rusch and colleagues (2007) also noted the lack of genetic similarity between the American and Canadian estuaries, where the Bay of Fundy sample appeared more similar to those of coastal waters in the North Atlantic. The Bay of Fundy is characterized by the highest tides in the world (Garrett, 1972) and contains mostly marine waters at high tide, the only periods at which a large boat might safely sample the waters. Perhaps samples collected at low tides would resemble the other estuaries more closely.

The overlapping complement of rhodopsin sequence types from open ocean metagenomes reinforces the unique composition of non-marine samples

Figure 3 summarizes an independent quantitative analysis confirming not only that rhodopsin types from four non-marine sites are largely absent from marine samples but that types common at the marine sites, although found at many ocean sites, are under-represented or absent at the non-marine sites. Approximately 900 rhodopsin sequences originating from 10 open ocean and the four non-marine metagenomic samples were clustered as
groups in which all pair-wise sequence distances were less than the maximum observed between members of LG1, as defined by Fig. 2.

Two trends are visible in Fig. 3. First, the open ocean samples exhibit a large amount of sequence overlap between them, as based on the values highlighted in yellow (sequences within a cluster that represent >10% of the overall sequence composition of an individual site). For instance, a large proportion of sequences from all open ocean samples belong to one of the three sequence clusters, and each of these clusters contains rhodopsins from geographically distant locations. Second, the highlighted blue values for non-marine samples clearly show a different clustering pattern than open ocean data. Although a large proportion of estuarine sequences (GS011 and GS012) fall within the first cluster, which is mostly comprised of rhodopsins from marine environments (Fig. 3), these sites contain rhodopsin diversity within LG1 and LG2 that is absent from the surface ocean waters. Some Lake Gatun (GS020) sequences also fall within the first cluster, however this freshwater environment clearly harbours a decreased amount of this sequence type in relation to open ocean samples. The majority of GS020 sequences fall within LG1 and LG2. The Punta Cormorant Lagoon (GS033) contains no sequences from the first cluster, differentiating this sample from oceanic ones, and while 27% of this sample's sequence types group into the second cluster (which overlaps with many open ocean rhodopsins), ~60% of GS033 rhodopsins are spread among LG1 and PCL1.

Organismal origins of rhodopsins from phylotypes LG1, LG2 and PCL1

The presence of LG1 and LG2 sequences in non-marine environments and their near absence in oceanic settings might reflect differential adaptation of the proteins they encode (thus bearing on rhodopsin function), or the differential occurrence of these genes in the genomes of related organisms that are for other reasons adapted to these sites (thus reflecting the phylogeny of the genome in which these sequences are found). We sought to identify the organisms carrying LG1 and LG2 phylotypes through a phylogenomic analysis of adjacent sequences (the gene directly upstream and/or downstream of the rhodopsin-encoding sequence of interest), with the recognition that because of lateral gene transfer, phylogenetic assignments must be regarded as provisional.

Some of the metagenomic assemblies that encoded LG1 or LG2 rhodopsins harboured a (sometimes partial) linked gene (Table 2) that encoded a protein sequence of sufficient length for phylogenetic analyses. Adjacent genes were selected on the criteria that they were detected in multiple assemblies such that several copies were available for phylogenetic analyses. In a few instances, examination across sample sites revealed the same homologues adjacent to rhodopsin genes, e.g. a gene encoding a similar function was found on assemblies from Lake Gatun, Punta Cormorant and estuarine sites. A list of these eight multiply-recovered rhodopsin-linked genes, the number of copies of each gene present in rhodopsin-encoding assemblies and a summary of the phylogenetic affiliation can be found in Table 2. Surprisingly, none of the linked genes appeared to code for proteins that were directly related to rhodopsin function.

Seven out of eight phylogenetic trees identify the phylum *Actinobacteria* as the potential origin of these rhodopsin-encoding genomic fragments from Lake Gatun, Punta Cormorant Lagoon, and the Delaware and Chesapeake Bays. The genes encoding formate dehydrogenase accessory protein (FdhB) are linked to LG2-encoding sequences in five separate instances and represent the
Table 2. Summary of phylogenetic analyses of genes adjacent to rhodopsins from LG1 and LG2

<table>
<thead>
<tr>
<th>Gene</th>
<th>Functional category (according to COG db)</th>
<th>Number of assemblies (total reads)</th>
<th>Homologues found with E-value &lt; 10&lt;sup&gt;-10&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glutamine kinase</td>
<td>M (Cell wall/membrane/envelope biogenesis)</td>
<td>15 (33)</td>
<td>10 (10)</td>
</tr>
<tr>
<td>ATP-dependent Clp protease</td>
<td>S (Signal transduction)</td>
<td>12 (19)</td>
<td>10 (10)</td>
</tr>
<tr>
<td>O-acetylhomoserine racemase</td>
<td>E (Amino acid transport and metabolism)</td>
<td>5 (12)</td>
<td>10 (10)</td>
</tr>
<tr>
<td>UbiA prenyltransferase</td>
<td>H (Coenzyme transport and metabolism)</td>
<td>5 (7)</td>
<td>10 (10)</td>
</tr>
<tr>
<td>Competence-damaged protein</td>
<td>R (General function prediction only)</td>
<td>3 (17)</td>
<td>10 (10)</td>
</tr>
<tr>
<td>Death-on-curing protein</td>
<td>R (General function prediction only)</td>
<td>3 (17)</td>
<td>10 (10)</td>
</tr>
</tbody>
</table>

In two larger assemblies these two genes co-occur as upstream and downstream of the rhodopsin gene.

A high proportion of the rhodopsin-encoding sequences recovered from both LG1 (77%) and LG2 (65%) were assembled adjacent to those genes listed in Table 2. The taxonomic association of other sequences linked to LG1 or LG2 rhodopsin genes that were present at a lower frequency than those in Table 2 were determined by the affiliation of the top scoring BLAST hit. The total number of LG-type rhodopsin genes that are linked to sequences of actinobacterial origin as determined by both phylogenetics and top scoring BLAST hits is summarized in Table 3. Table 3 also summarizes other qualitative characteristics regarding metagenomic assemblies encoding LG-type rhodopsins. Very few of these rhodopsins genes were adjacent to sequences that originate from outside the Actinobacteria.

We also assessed the phylogenetic signals from rhodopsin-linked genes outside the LG1 and LG2 phylogroups to ensure that rhodopsin-encoding sequences recovered from the GOS were generally not associated with actinobacteria-like genes. In this control exercise, all rhodopsin sequence OTU representatives present on the phylogenetic tree in Fig. 1 that originated from the GOS were assessed for the presence of adjacent genes and assigned affiliation to known taxonomic groupings based on the top scoring BLAST hit (Fig. S1). Surprisingly, our BLAST analyses revealed that sequences encoding the PCL1 phylotype were also linked to genes that appeared actinobacterial in origin. Genes commonly linked to PCL1 sequences encoded zinc proteases or polyribonucleotide nucleotidyl transferases. The taxonomic affiliation of genes linked to those encoding LG1, LG2 and PCL1 provides an overwhelming signal to the Actinobacteria, whereas only a small fraction of the associated sequences from any of these phylotypes displays affinity for other bacterial phyla (Table 3).

In contrast to LG1, LG2 and PCL1, the remainder of the rhodopsin sequences spread among many different clades.
The majority of GOS sites were categorized under the following environment types: open ocean (23 samples), coastal (22), estuarine (GS011 and GS012), freshwater (GS012) and hypersaline lagoon (GS033). BLAST searches identified bacterial recA or archaeal radA homologues within specific GOS sample metagenomes that we then grouped into their respective environmental category (see Experimental procedures). The ratio of recA and radA to rhodopsins is high across all GOS environmental categories (Table 4). In a single copy scenario, in which we assume one rhodopsin gene per genome, the abundance measurements appear similar across open ocean (63%), coastal (67%) and estuarine environments (62%), whereas environments like Lake Gatun (35%) and the Punta Cormorant Lagoon (36%) display lower, but nonetheless substantial values (Table 4).

Relative abundance of rhodopsin sequence types for individual metagenomic samples

To gain assurance that the unique sequence groups indeed represented a substantial fraction of the rhodopsin sequences from each of the non-marine GOS sample sites, we sought to capture the diversity represented by the numerous smaller partial rhodopsin sequences created by the shotgun sequence method that were not used in our phylogenetic analyses. In silico rhodopsin libraries of variable sequence lengths \( \geq 70 \) amino acids (aa) were extracted from each metagenomic sample, and

**Table 3.** Summary of the assigned taxonomic affiliation of genes directly upstream or downstream of rhodopsin phylotypes found in non-marine GOS samples.

<table>
<thead>
<tr>
<th>Phylotypes</th>
<th>LG1</th>
<th>LG2</th>
<th>PCL1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of rhodopsin-encoding sequence reads</td>
<td>93</td>
<td>23</td>
<td>40</td>
</tr>
<tr>
<td>Number of metagenomic assemblies with rhodopsin-encoding sequences(^a)</td>
<td>53</td>
<td>15</td>
<td>24</td>
</tr>
<tr>
<td>Number of rhodopsin-encoding assemblies linked with genes of actinobacterial origin(^b)</td>
<td>44  (83%)</td>
<td>10  (67%)</td>
<td>18  (75%)</td>
</tr>
<tr>
<td>Number of assemblies linked with genes that originate from outside actinobacteria</td>
<td>3   (6%)</td>
<td>1   (7%)</td>
<td>0</td>
</tr>
<tr>
<td>Number of assemblies that do not contain any linked genes(^c)</td>
<td>8   (15%)</td>
<td>4   (26%)</td>
<td>6    (25%)</td>
</tr>
<tr>
<td>Total number of individual rhodopsin-encoding sequence reads linked with genes of actinobacterial origin</td>
<td>81  (87%)</td>
<td>16  (70%)</td>
<td>32   (80%)</td>
</tr>
</tbody>
</table>

\(^a\) As multiple rhodopsin-encoding sequence reads were assembled into the same metagenomic scaffold.
\(^b\) As determined by phylogenetic analyses or top scoring BLAST hit.
\(^c\) Assemblies only contained rhodopsin sequence as significant BLASTP hit.

Exposing the prevalence of rhodopsin sequences in different environmental categories

To estimate the proportion of genomes that contain a rhodopsin gene within a particular metagenomic sample...
used to determine the abundance of rhodopsins belonging to the phylogenetic clusters identified in Fig. 1.

Figure 4A shows that 70% of the sequences within the Lake Gatun partial library belong to LG1 and LG2 clades, whereas 14% belong to the SAR11-like-1 clade. The remaining 16% of sequences are scattered throughout the tree (Fig. S3). Data retrieved from the two estuary sites were merged together and considered a single sample due to their overlapping composition of closely related rhodopsin sequence types (Fig. S4). In contrast to the freshwater rhodopsin library, LG1 and LG2 phylotypes account for less than 50% of the overall abundance of rhodopsins in the estuarine library (Fig. 4B). Additionally, the estuarine library also contained 16 sequences in common with the Lake Gatun sample that grouped into SAR11-like-1, while a group of 22 SAR11-like-2 sequences that were present in the estuaries were absent in the freshwater library. Differences in environmental parameters between Lake Gatun and the estuaries may explain their contrasting rhodopsin sequence phylotypes. However, these contrasts could also be the result of multiple factors, such as sampling depth and seasonal variation.

Eighty-six per cent of Punta Cormorant Lagoon sequences fall into three major groupings, LG1 (29%), LG2 (27%), and SAR11-like-2 (14%).

**Table 4.** Estimate of the proportion of genomes that contain a rhodopsin gene in different GOS environmental categories as estimated by the ratio of rhodopsins to single-copy recA + radA BLAST hits in the GOS.

<table>
<thead>
<tr>
<th>Environment type</th>
<th>Open ocean (23 samples, 4901 Mbp)</th>
<th>Coastal (22 samples, 4365.5 Mbp)</th>
<th>Estuaries (GS011 and GS012, 314 Mbp)</th>
<th>Lake Gatun (GS020, 516 Mbp)</th>
<th>Punta Cormorant (GS033, 1200 Mbp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(No. of rhodopsin hits) / (No. of recA hits + radA hits)</td>
<td>63% / (1484/2364)</td>
<td>67% / (1814/2718)</td>
<td>62% / (129/208)</td>
<td>35% / (135/383)</td>
<td>36% / (176/492)</td>
</tr>
</tbody>
</table>

a. Amount of sequence data in the sample(s).

**Fig. 4.** Phylogenetic affiliation of rhodopsin sequences ≥ 70 aa from the (A) Lake Gatun (93 sequences), (B) estuarine (104 sequences) and (C) Punta Cormorant Lagoon (138 sequences) metagenomic samples. Partial sequence data (see Results and Experimental procedures) were used to determine the phylogenetic affiliation of rhodopsin sequences from each data set according to the sequence clusters represented in Fig. 1 and described in Results. The different phylogenetic groupings are colour coded according to the key.
PCL1 (30%) and PCL2 (27%) (Fig. 4C). PCL1 and PCL2 appear to represent less sequence diversity than LG1 (Fig. 2 and Fig. S5), suggesting that the former clades may be representative of a high abundance of single organism type within the lagoon, whereas sequences from the latter may represent a more diverse group of closely related microbes.

We estimated the library coverage of cloned rhodopsin sequences for each of the GOS sample sites of interest with rarefaction curves and library coverage calculations to assess the reliability of binning rhodopsin protein sequences into phylogenetic groups to determine population structure. Rarefaction curves revealed a similar trend for each of the three environment types, where increasing the phylogenetic distance at which OTUs were clustered from zero to 0.1 amino acid substitutions per site caused the curve to begin to plateau with less library coverage (Fig. S6). This suggests that rhodopsins from each of the metagenomic samples are relatively well sampled when OTUs are clustered at a distance ≤ 0.1. Clustering OTUs at this level resulted in library coverage estimates of 86% for the Lake Gatun data, 77% for the combined estuary data sets and 93.5% for the Punta Comorant Lagoon sample.

Primary sequence features of LG1, LG2 and PCL1 rhodopsins compared with proteorhodopsin and bacteriorhodopsin

All LG1, LG2 and PCL1 sequences conserve the amino acid residues required for proton pumping (Spudich and Jung, 2005), and in this regard they appear to be more similar to the biochemically characterized Sar86 proteorhodopsin (AAG10475) than to the haloarchaeal proton pump, bacteriorhodopsin (V00474). As in proteorhodopsin, all three novel sequence clusters conserve Glu108, rather than the homologous Asp96 present in bacteriorhodopsin. Acidic amino acids present in this position are responsible for a reprotonation step in the latter half of the rhodopsin photocycle (Spudich and Jung, 2005). A carboxylic acid residue involved in the proton release pathway at the extracellular surface in bacteriorhodopsin (Glu204) is replaced with non-protonatable (mostly hydrophobic) amino acids in proteorhodopsin (Partha et al., 2005), PCL1 and LG2. However, nearly all LG1 sequences encoded arginine at this position.

Within the retinal-binding pocket, all three sequence clusters also show greater sequence similarity to proteorhodopsins, conversely in some amino acid positions they are more bacteriorhodopsin-like (Fig. S7). These novel sequences also conserve leucine at the position responsible for spectral tuning in proteorhodopsin (Beja et al., 2001), perhaps suggesting that these non-marine rhodopsins are tuned to absorb green wavelengths of light. However, the differences at other positions in the retinal-binding pocket in the three sequence clusters versus proteorhodopsin (Fig. S7) may potentially affect the wavelength of light absorbed by retinal.

Discussion

Abundant unique groups of rhodopsin sequences indicates a broad environmental distribution for rhodopsins

Many of the novel divergent rhodopsin proteins included in the phylogenetic analyses (Fig. 1) here would not be amplified by the polymerase chain reaction using primers constructed to target large groups of marine proteorhodopsins, or archaeal bacteriorhodopsins. Nor could a single primer set capture rhodopsin genes from all three of the most abundant sequence groups in the Lake Gatun metagenomic data (Fig. 4A). Similar to the initial finding of proteorhodopsin in open ocean environments (Beja et al., 2000), unique rhodopsin diversity from different environment types will likely be captured through metagenomic studies in the future. Although in some cases fortuitous cultivation of rhodopsin-containing microbes whose genomes are sequenced [e.g. Roseiflexus sp. RS-1 available at the Joint Genome Institute (http://www.jgi.doe.gov), G. violaceus (Nakamura et al., 2003)] has occurred, the divergent properties of rhodopsin proteins present difficulties in capturing whole community rhodopsin diversity from environments that have yet to be explored for such functions.

Our analyses of the surface water metagenomic data from Lake Gatun, the estuaries Delaware and Chesapeake Bays, and the Punta Cormorant Lagoon within the context of all GOS sites (Fig. 1 and Table 1), revealed that these environments harboured novel groups of rhodopsins (LG1, LG2 and PCL1) that represented a large proportion of sequences in each respective data set (Fig. 4). The discovery of unique and abundant sequence types from freshwater, estuary and lagoon sites, subject to frequent interactions with marine waters, indicate that rhodopsins from these environments are not simply the result of colonization by marine-type proteorhodopsins. The lack of LG1, LG2 and PCL1 sequence types in 9266 Mbp of metagenomic data from open ocean and coastal surface water (Fig. 2 and Table 1) suggests that these sites harbour rhodopsins that either have evolved to the selective pressures represented by different ecologies, and/or originate from organisms native to these habitats but not to marine surface waters. Notably, LG1 types predominate in the most and the least saline of the GOS sites: salinity per se is probably not the key environmental determinant in this distribution.
Potential organismal origins of LG1-, LG2- and PCL1-type rhodopsins

Phylogenetic analyses of protein-coding genes linked to sequences encoding LG1, LG2 and PCL1 rhodopsins (Tables 2 and 3) identify members of the phylum Actinobacteria as the potential carriers of these sequence types in the metagenomic data from all four GOS sites (Fig. S1). Many previous publications describe the assignment of metagenomic fragments to taxonomic groupings based on the phylogenetic signal of the genes present (Nesbo et al., 2005; McCarren and DeLong, 2007). For example, when Beja and colleagues (2000) discovered rhodopsin phototrophy in the ocean they found a metagenomic fragment encoding the photo-protein linked to a γ-protobacterial 16S rRNA gene, an observation that led to the suggestion that the SAR86 lineage harboured ‘proteorhodopsins’ phylogenetically divergent from haloarchaeal and fungal-type opsins. Based on the strong association of LG-type and PCL1 sequences with actinobacterial related genes (Table 3), we propose these rhodopsins be called actinorhodopsins. In combination with a strong actinobacterial signal, actinorhodopsins are predominately found in the non-marine hypo- and hyper-saline GOS samples (Fig. 3), where they are more abundant than proteorhodopsin sequence types (Fig. 4). Actinorhodopsins, distributed among three phylgroups, are nearly absent from marine surface water samples (Table 1). In contrast, proteorhodopsins, which are most often present in proteobacterial genomes (McCarren and DeLong, 2007 and Fig. S1), are predominately found in marine samples (Fig. 3) although their distribution, like actinorhodopsins, is not restricted to a single environment type. Primary sequence analysis of the retinal-binding pocket in actinorhodopsins indicates a unique pattern where similarities to proteorhodopsin are predominant, although individual features of the haloarchaeal proton pump bacteriorhodopsin are also present (Fig. S7).

Genomic and metagenomic studies reveal microbial rhodopsins in an increasing variety of phylogenetic, genomic and ecological contexts and of varying sequence, regulatory interactions and physiological functions. No single naming system is likely to be consistent with all these parameters, and we do not claim that ‘actinorhodopsins’ will be unambiguously identifiable. The terms ‘proteorhodopsin’ and ‘archaeal rhodopsin’ are similarly imprecise, and yet we feel all have utility in emphasizing the structural and functional variety within this important family of proteins while offering a preliminary and rough binning in terms of likely origins and roles. For us, ‘actinorhodopsins’ are proteins more closely related to proteorhodopsins than to classic haloarchaeal proteins, bearing membership in one of the three clades identified here by sequence phylogeny, and found preferentially in the non-marine aquatic samples, more specifically in Actinobacteria.

A potential actinobacterial origin for actinorhodopsins is interesting from many different perspectives. Rusch and colleagues (2007) described the abundance of dominant ribotypes for GOS sites and found that similar actinobacterial 16S rRNA sequences were present in Lake Gatun, estuarine and Punta Cormorant samples but did not contribute to dominant types from marine surface water metagenomes. Coincidently, the distribution of these actinobacterial ribotypes corresponds to that of actinorhodopsins (Fig. 4). Furthermore, the analyses of rhodopsin-linked genes from sequences outside of LG1, LG2 and PCL1 indicates that the strong actinobacterial association observed here is restricted to rhodopsins from these novel clusters (Fig. S1).

Interestingly, Elifantz and Kirshman (Elifantz et al., 2005) observed that actinobacteria in the Delaware Bay were abundant in the freshwater region near the mouth of the Delaware River but diminished as the salinity gradient increased towards the entrance of the bay where marine waters flowed in. Each of the GOS metagenomic samples from Delaware and Chesapeake Bays were taken where salinity levels were reduced (Rusch et al., 2007; K. Wong, pers. comm.) due to the high input of freshwaters from river systems. The presence of actinobacterial lineages at higher abundances in the freshwater regions of estuaries is likely linked to the global distribution of these organisms in lakes and rivers, where actinobacteria often belong to the dominant fraction of heterotrophic bacterioplankton (Kirchman et al., 2005; Allgaier and Grossart, 2006).

Previously, the presence of rhodopsin sequences from freshwater environments was restricted to a single sequence from the cyanobacterium Anabaena sp. PCC7120 (Jung et al., 2003). The Anabaena sequence is not only absent from the Lake Gatun metagenomic sample, but represents a highly divergent haloarchaeal-type rhodopsin that does not display any specific affiliation to any bacterial rhodopsins (Sharma et al., 2006) and therefore was not included in the phylogenetic analysis presented here. Prior to the GOS, no data set had yet revealed the widespread existence of rhodopsins in a freshwater ecosystem (Rusch et al., 2007). If indeed rhodopsins do originate from actinobacteria in freshwater environments perhaps they are linked to the success of this group previously documented for many rivers and lakes (Pernthaler et al., 1998; Warnecke et al., 2005). While rhodopsins are believed in general to be supplementary energy producers (Bryant and Frigaard, 2006; Gomez-Consarnau et al., 2007), the presence of rhodopsin-encoding genes in dominant members of the microbial community is a trend emerging from both marine and hyperhalophilic ecosystems, where organisms such
as *P. ubique* (Morris et al., 2002) and *Haloquadratum walsbyi* (Benlloch et al., 2002) dominate.

Rhodopsins are encoded in the genomes of *Rubrobacter xylanophilus* and *Kineococcus radiotolerans*, both of which belong to the phylum *Actinobacteria* (genome sequences available at http://www.jgi.doe.gov/). However neither rhodopsin sequence groups with LG-type or PCL1 sequences. Previous work showed that *R. xylanophilus* encodes a haloarchaeal-type rhodopsin homologue (Sharma et al., 2006) and Fig. 1 shows that the *K. radiotolerans* sequence does not display any specific affiliation within the bacteria. The lack of sequence similarity between LG-type and PCL1 sequences and the two other actinobacterial rhodopsins could be interpreted as evidence against an actinobacterial origin for these novel phylotypes. However, the suspected haloarchaeal origin of the *R. xylanophilus* homologue combined with the knowledge of the high frequency of rhodopsin transfer events (Frigaard et al., 2006; Sharma et al., 2006; McCarren and DeLong, 2007) could alternatively be interpreted as evidence for the propensity of actinobacteria to utilize rhodopsin photosystems. The observation of vastly divergent rhodopsin sequences within a single phylum was made for the cyanobacteria *G. violaceus* and *Ana- baena* sp. PCC7120 (Sharma et al., 2006). Furthermore, divergent rhodopsin sequences have also been observed within the *Proteobacteria* (McCarren and DeLong, 2007) (also see Fig. S1).

**Importance of rhodopsin photosystems to microbial communities of different environment types**

Analyses of abundant ribotypes across GOS sites also found SAR11 sequences in both Lake Gatun and the estuaries (Rusch et al., 2007), which correlates with the presence of rhodopsins from SAR11-like clades 1 and 2 (Figs 1 and 4). An extended environmental distribution of rhodopsins, combined with their potential links to abundant ribotypes, suggests that these proteins may be significant components in the microbial ecosystems under study here. In agreement with Rusch and colleagues (2007) our estimate of the abundance of rhodopsin genes indicates they are highly abundant relative to *recA* and *radA* genes across individual GOS samples (data not shown). When samples are grouped according to their environmental category (assigned by Rusch et al., 2007) interesting trends are revealed (Table 4).

We calculate that proteorhodopsin-encoding genes are present in similar ratios between coastal and open ocean environments (Table 4), as estimated by their ratio to the sum of occurrence of *recA* and *radA* genes. The similarities between the values observed here indicate that even in more nutrient-rich coastal waters rhodopsins play a substantial role in adaptation to light at the surface, and appear to be an equally important component of microbial communities in both environmental categories. In the less sampled estuary metagenomes, rhodopsin genes also are present in similar ratios as in open ocean and coastal surface water samples (Table 4). Similarly high ratios in estuaries support an important role for these photosystems in an environment type differentiated from coastal and open ocean ecosystems according to environmental parameters, geographical location and differences in the rhodopsin phylotypes present (Fig. 3).

Abundance estimates from Lake Gatun and the Punta Cormorant Lagoon samples are reduced compared with the other environmental categories (Table 4); however, the range of values represented by our measurements still indicates a substantial fraction of the microbial community harbours rhodopsin genes in freshwater and shallow marine-associated lagoons. Our knowledge of the number of rhodopsin paralogues per genome is restricted in marine and freshwater environments due to a limited number of sequenced genomes encoding rhodopsins from isolates cultivated from either environment type. The relatively high proportion of rhodopsin sequences from all environmental categories investigated here may be the amplified effect of microbial genomes encoding numerous rhodopsin paralogues per genome, similar to the haloarchaea which can encode zero to six per genome (Baliga et al., 2004; Sharma et al., 2007).

While the functional role of actinorhodopsins from new environments and indeed many proteorhodopsins is unknown, the extended environmental distribution of these photo-proteins is a strong indicator of the positive selection inferred by the ability to exploit light for energy or direction. Interesting features of actinorhodopsins are the conservation of the acidic amino acid residues (Asp97 and Glu108) associated with proton pumping in proteorhodopsin (Spudich and Jung, 2005). The conservation of such amino acids may indicate a putative proton pumping function; however, exceptions to this observation exist. A type 1 rhodopsin homologue encoded in the genome of the fungus *Neurospora crassa* also conserves these features but lacks the ability to transport protons (Furutani et al., 2004). Another clue that supports the potential that actinorhodopsins are proton pumps is the lack of genomic association between the genes encoding these rhodopsins and sensory rhodopsin transducers (Martinez et al., 2007). Considering that the proton pump bacteriorhodopsin in the halolarchaeon *Halobacterium* sp. NRC-1 can support periods of phototrophic growth during low oxygen tension (Hartmann et al., 1980), the ability to generate an electrochemical gradient from light may be an important adaptation in any environment where an organisms terminal electron acceptor is periodically depleted or absent. Lakes, estuaries and lagoons are all subject to anoxic conditions where energy derived from rhodopsin phototro-
phy may aid in compensating for decreases in aerobic respiration. Furthermore, proteorhodopsin phototrophy was shown by Gomez-Consarnau and colleagues (2007) to stimulate growth of marine flavobacterial isolates under both natural seawater and artificial low nutrient conditions in culture, but did not enhance growth levels in media with higher levels of dissolved organic carbon. Perhaps energy derived from rhodopsin photosystems may enhance growth rates in any scenario that involves competition for scarce amounts of an organism’s energy sources.

**Experimental procedures**

**Extraction of rhodopsin homologues from the GOS**

Shotgun sequence reads from 51 sample sites in the GOS (Rusch et al., 2007) were downloaded from the CAMERA database (http://camera.calit2.net/). Sequence reads from an individual site were translated into all six reading frames to generate protein sequences from the available nucleotide metagenomic data. The translated sequences were used as databases for BLASTP similarity searches using different query sequences that represent the broad range of rhodopsin diversity. The selected query sequences were Marinobacter sp. ELB17 (ZP_01737880), marine euryarchaeal proteorhodopsin HF70_19B12 (ABBB2977), Sar86 marine \( \gamma \)-proteobacteria (AAK30175) and sensory proteorhodopsin (EAG13698). Preliminary tblastn searches of the Lake Gatun data (sample GS020) on the CAMERA website (http://camera.calit2.net/) indicated unique sequence types JCVI_READ_1091143055461, JCVI_READ_1095333018364 (SAR11-like), JCVI_READ_1095313000154 (LG1), which were also used as queries.

Each of the seven query sequences was used in a BLASTP search to extract shogun sequence reads from each individual sample site that displayed similarity to rhodopsin sequences with an E-value cut-off of \( 10^{-10} \). The results of the BLASTP searches were merged and any identical shotgun sequence reads that were extracted by more than one query sequence were removed from the analyses to eliminate redundancy. The portion of the translated shotgun sequence read that contained the rhodopsin protein was defined as a translated read region of at least \( N \) amino acids between two stop codons, between a beginning of the read and a stop codon or between a stop codon and the end of the read, residing on the same frame where the BLASTP hit to rhodopsin was observed. Rhodopsin sequences were extracted at two different size lengths, \( N \approx 150 \) aa (to have fragments of sufficient length for phylogenetic analyses) and \( N \approx 70 \) aa (to perform analyses of the diversity sampled by the shotgun sequencing method). Some translated shotgun sequence reads contained more than one region qualified for the definition above, only one of which was the partial rhodopsin protein sequence. In such rare cases, the fragment with rhodopsin proteins was identified manually.

**Data notation**

Throughout the manuscript we used sample numbers assigned by the CAMERA database (http://camera.calit2.net): GS000 through GS051. For convenience, sequence read identifiers were renamed to reflect sample origin of the sequence. A conversion table of these identifiers to CAMERA sequence read identifiers is available as Table S1).

**Assessment of rhodopsin abundance**

BLASTP searches to identify the total number of bacterial RecA or archaeal RadA homologues within specific GOS sample metagenomes were performed in the same way as identification of rhodopsin homologues (see above). A relaxed bit-score cut-off of 55 was derived from BLASTP searches against the nr database with query RecA sequences such that it retrieved both full-length and partial RecA sequences from the database. These query sequences were then used in BLASTP searches of the GOS samples to estimate the number of RecA homologues present in each site. The RecA query sequences were: Acetobacter polyoxygens (Q08327), Chlorobium chlorochromatii CaD3 (YP_378560.1), Prochlorococcus marinus str. NATL2A (YP_292326.1), Frankia sp. CcI3 (YP_482607.1), Thermus aquaticus str. HB8 (JX0292), Morella thermoacetica ATCC 39073 (YP_429173.1).

A more stringent bit-score cut-off was derived for archaeal RadA sequences due to paralogous RadB sequences present in many archaeal genomes. Using RadA query sequences in BLASTP analyses of the nr database a bit-score cut-off of 120 was derived. This score recovered full-length and partial RadA sequences but in most cases excluded RadB paralogues. The query sequences used to estimate the number of RadA genes present in different GOS samples were: Methanospirillum hungatense DSS 3081 (YP_448501.1), Haloarcula marismortui ATCC 43049 (YP_137357.1), Nanoarchaeum equitans Kin4-M (NP_963710.1), Staphylococcus marinus F1 (YP_00140983.1), Crenarchaeum symbiosium (O93749), Methanopyrus kandleri AV19 (NP_614725.1).

To characterize the abundance of rhodopsin genes we grouped metagenomic samples according to their environmental category in the GOS. The majority of GOS sample sites are categorized under the following environmental types: open ocean (23 samples), coastal (22 samples), estuarine (Delaware Bay sample GS011 and Chesapeake Bay sample GS012), freshwater (Lake Gatun sample GS020) and lagoon (Punta Cormorant sample GS033). Note that other environmental categories are not considered here because they do not clearly group into the categories above (e.g., coral reef environments). Also the Bay of Fundy sample GS006 is considered here as a coastal environment due to its lack of genetic similarity to estuary sites GS011 and GS012 (see Results), but enhanced similarity to North Atlantic coastal samples (Rusch et al., 2007). From each of these environmental categories we calculated rhodopsin abundance by dividing the sum of all rhodopsin BLASTP hits from sites within that category by the sum of all the RecA and RadA hits.

**Identification of LG1 and LG2 sequences within the GOS**

Rhodopsin sequences within LG1 or LG2 were pair-wise aligned using ‘align two sequences using BLAST’ (bl2seq) to derive a bit-score cut-off for further BLASTP searches against
all 51 GOS sample sites. To ensure that LG-type sequences belonging to either of these clades were retrieved we relaxed the bit-score cut-off to capture a greater diversity of sequences by deriving a cut-off value from BLASTP of LG1 or LG2 sequences against each other and against other sequences outside these phylotypes. Sequences GS020–26 (LG1) and GS020–62 (LG2) (see Fig. S2) were used as queries in BLASTP searches against all 51 sample sites and only the sequences with the bit-score cut-off above 160 for LG1 and 130 for LG2 were retrieved. Extraction of partial rhodopsin homologues was performed as described in Extraction of rhodopsin homologues from the GOS.

Rhodopsin alignments

All rhodopsin alignments were initially generated using the MUSCLE program (Edgar, 2004) and then manually edited similar to techniques used for previously published alignments (Ihara et al., 1999; Spudich and Jung, 2005), i.e. poorly aligned and hypervariable loop regions were excluded from the alignment. In alignments of rhodopsins ≥ 150 aa nearly all sequences were trimmed to the same length, to minimize number of alignment sites with missing data. In addition, sequences from the ≥ 150 aa data sets were required to span a sufficient length of the rhodopsin-coding protein in order to be included in the alignment. For example, sequences that only overlapped at the beginning or end of the alignment were excluded. In some cases the total length of the alignment was greater than some of the metagenomic sequences included in the analyses, resulting in missing alignment positions at either end of the protein. In these cases the absence of certain positions was treated as missing data for phylogenetic reconstruction.

Halorhodospina rhodopsin sequences were not included in any of the phylogenetic trees presented here because none of the Lake Gatun rhodopsin sequences, or sequences from other environments we analysed, were retrieved by BLASTP using halorhodospina rhodopsins as queries. In addition, halorhodospina rhodopsins are extremely divergent from bacterial-type rhodopsins and alignment of sequences from these two distantly related groups would result in fewer amino acid positions available for phylogenetic analyses.

The alignment used to generate the tree in Fig. 1 initially contained a collection 315 sequences that was used to generate OTU clusters grouped at aa phylogenetic distances of ≤ 0.1 in DOTUR (http://www.plantpath.wisc.edu/fac/joh/dotur.html) (Schloss and Handelsman, 2005). To make this data set more manageable and to improve alignment quality of the large numbers divergent rhodopsin sequences, an alignment was first generated with the MUSCLE program, edited manually and then refined using multiple iterations in the HMMER program (Eddy, 1998). HMMER trained a Hidden Markov Model based on the initial MUSCLE alignment, which was then used to generate a new alignment in HMMER. This process was repeated for 19 iterations until the only changes observed in the automated alignment were minor alterations made to the beginning and end of the protein. The final alignment generated from this process was manually edited and used to generate a distance matrix in TREE-PUZZLE (Schmidt et al., 2002). From ~300 sequences in the initial alignment 102 OTUs were generated as indicated above. One sequence from each of these 102 OTUs was chosen as a representative of the entire OTU cluster and included in the alignment used to generate Fig. 1. The composition of each OTU cluster is represented in tab-delimited text format in Table S2.

Alignments were also generated for partial sequence libraries of rhodopsin proteins ≥ 70 aa to be used to estimate the diversity represented by the numerous partial rhodopsin sequences sampled by the shotgun sequence method. In these alignments we retained both sequences that overlapped at either end of the protein and those that spanned the central portion with the criteria that all sequences present in the final edited alignment represented at least 70 aa positions. In some cases this resulted in sequences that did not overlap at all or only overlapped by a few positions (which was accounted for in the distance matrices generated for population analyses, see below).

Rhodopsin phylogenetic analyses

Unless noted otherwise, phylogenetic trees and their bootstrap support values presented in this analyses were calculated in the PHYL program version 2.4.4 (Guindon and Gascuel, 2003; Guindon et al., 2005) under the WAG substitution matrix, with a discrete I distribution divided into eight rate categories, proportion of invariable sites estimated, and with 100 bootstrap replicates.

Identification of rhodopsin sequence types to determine their relative abundance in each metagenomic sample

Distance matrices were generated for both the ≥ 150 aa and ≥ 70 aa rhodopsin data sets for each of the four metagenomic sites (GS011, GS012, GS020, GS033) using TREE-PUZZLE version 5.2 (Schmidt et al., 2002) under the WAG substitution matrix, with a discrete I distribution divided into eight rate categories, and proportion of invariable sites estimated. Note that unlike population-level analyses for Lake Gatun and estuary sites that included sequences ≥ 70 aa, only sequences extracted that were ≥ 150 aa were considered from the Punta Cormorant Lagoon due to the high level of sequence coverage for this site (Rusch et al., 2007).

However, unlike sequence selection for phylogenetic analyses, Punta Cormorant sequences ≥ 150 aa that only overlapped at the beginning or end of the rhodopsin protein were included with the criteria that all sequences were greater than 70 aa following the exclusion of unaligned positions from the alignment. The matrices were used as input files for the DOTUR program (Schloss and Handelsman, 2005) to generate rarefaction curves and to cluster rhodopsin sequences into OTUs at various distances.

Rarefaction curves were generated for OTUs grouped at four different levels of phylogenetic distance: zero, ≤ 0.03, ≤ 0.05 and ≤ 0.1. When a rarefaction curve reaches a plateau no new OTU groups will be uncovered with increasing sequencing depth. Therefore the shape of a rarefaction curve can be used as an estimator of library coverage. To more thoroughly describe the levels of library coverage observed for rhodopsin sequences recovered from various sample sites within the GOS, library coverage was also calculated. Library coverage is defined as the percentage of sequences in a library that is not comprised of singletons and
was calculated using the equation Coverage = \([1 - (n_i/N)] \times 100\), where \(n_i\) is the number of singletons within a library and \(N\) is the total number of sequences sampled (Good, 1953). The number of singletons in a sequence library was calculated with DOTUR (Schloss and Handelsman, 2005).

To measure the overall abundance of rhodopsin sequence types from an individual metagenomic library, we took into account the diversity represented by the partial rhodopsin sequences sampled by the shotgun sequence method. In order to place rhodopsin sequences from the two \(\geq 70\) aa data sets [Lake Gatun (GS020), estuarine (GS011 and GS012)] and one \(\geq 150\) aa data set [Punta Cormorant Lagoon (GS033)] into broad phylogenetic groups, a three-step process was derived. First, partial sequences from a single sample were clustered in DOTUR at a phylogenetic distance \(\leq 0.1\). This level of distance was chosen because each of the rhodopsin sequence libraries appeared fairly well sampled at this coarse phylogenetic cut-off in the rarefaction curves. Therefore, for the purposes of estimating the abundance of sequence types in unique clades, the partial sequence library should be roughly representative of the actual abundance of sequence types within each of the metagenomic samples. Next, OTUs within the partial sequence libraries were mapped onto the robust phylogenetic trees generated from the \(\geq 150\) aa data sets for each site, respectively, assuming that one of the OTUs within a cluster had a phylogenetic anchor on the tree. This method accounted for the mapping of 79 of the 93 sequences in the partial sequence library for Lake Gatun, 93 of 111 partial sequences from the marine estuaries and all of the sequences from the Punta Cormorant Lagoon data set. Lastly, in the few remaining cases, OTU groups did not have a sequence anchor in the robust phylogenies, and in these instances clusters were assigned positions based on a maximum likelihood phylogenetic tree of the partial sequence data alignment (data not shown). Surprisingly, perhaps due to the large amount of divergence between rhodopsin groups, sequences were placed into these unique clades with high bootstrap support.

As noted above, alignment of partial sequence data in the \(\geq 70\) aa data sets resulted in sequences that only overlapped by a small number of amino acid positions. In a few cases, sequences that only overlapped by \(\sim 10\) aa could form an OTU if this overlap occurred in a conserved part of the rhodopsin protein, resulting in the formation of false OTU clusters. To circumvent the artefacts created by this phenomenon, we discarded any pair-wise distance values that were generated from two sequences that overlapped by \(\geq 30\) aa. Note that no changes in the number of OTU groups were observed upon increasing the stringency of overlap to 60 aa, suggesting that this effect was caused only by sequences that shared minimal overlap. The OTU groupings that resulted from this analyses were cross-referenced against the maximum likelihood phylogenies generated from the partial sequence libraries (data not shown) to verify that sequences were placed in the proper clusters.

**Extraction of metagenomic fragments linked to rhodopsins from non-marine GOS samples**

In order to extract genes immediately adjacent to rhodopsins from non-marine samples, we used assemblies of GOS data available at the CAMERA website (http://camera.calit2.net; for assembly parameters see Rusch et al., 2007). Assemblies were extracted for all rhodopsin sequence OTU representatives present on the phylogenetic tree in Fig. 1 that originated from the GOS, as well as for all rhodopsin sequences present in phytypes LG1, LG2, PCL1 and PCL2. Only assemblies containing reads from the same sample site were used. Scaffold containing rhodopsin sequences were extracted and used as queries for BLASTX searches against nr database with E-value cut-off of 10\(^{-10}\). The BLASTX results were parsed for the functional title of top-scoring hits in scaffold fragments adjacent to the rhodopsin-coding part and taxonomic affiliation was assigned based on that of the top-scoring hit(s). Eight genes linked to rhodopsins from phytypes LG1 and LG2 (see Table 2) determined as adjacent to rhodopsins and occurring in at least two assemblies were selected for further phylogenetic analyses. For each gene, corresponding regions of the assemblies as well as homologues from the BLASTX searches against nr databases were extracted and aligned in CLUSTALW version 1.83 (Jeanmougin et al., 1998). The alignments were manually edited and phylogenetic trees were reconstructed in the PHYLIP program version 2.4.5 (Guindon and Gascuel, 2003; Guindon et al., 2005) under JTT+I model with 100 bootstrap samples.

**Cluster composition analysis**

Rhodopsin homologues retrieved from all open ocean sites and four non-marine sites (GS011, GS012, GS020, GS033) were aligned as in Fig. 1, distance values were calculated using TREE-PUZZLE and OTU clusters obtained in the DOTUR program at 0.82 distance (the distance was set at the divergence of LG1 clade) using the furthest neighbour approach. All rhodopsin sequences present in the alignment overlapped by a length \(\geq 60\) aa. The composition of each cluster with respect to sample origin is summarized in Fig. 3.

**Acknowledgements**

A.K.S. thanks the Nova Scotia Health Research Foundation Student Research Program and the Canadian Institute for Health Research (CIHR) Doctoral Student Program for their support and O.Z. is grateful to the CIHR Postdoctoral Research Fellowship. Research was funded by a CIHR grant (MOP-4467) to W.F.D. The authors also thank Katrin Sommerfeld. We acknowledge that sequence data used in this study were obtained during the Global Ocean Sampling expedition from the waters of Canada, Mexico, Honduras, Costa Rica, Panama, Ecuador and French Polynesia/France. All sequencing data collected from the waters of the above named countries remain part of the genetic patrimony of the country from which they were obtained.

**References**


Baliga, N.S., Bonneau, R., Facciotti, M.T., Pan, M., Glusman,


Supplementary material

**Fig. S1.** Phylogenetic tree of rhodopsin sequences shown in Fig. 1 labelled (in red) with assigned taxonomic affiliations based on genes immediately adjacent to rhodopsin gene in assemblies. For GOS sequences with no affiliation shown, there was no significant BLASTX hit to perform taxonomic assignment.

**Fig. S2.** A. Phylogenetic tree represented in Fig. 2 shown with GOS identifiers. See the legend in Fig. 2 for full details. Bootstrap support values less than 65% are not shown. B. Sub-tree of phylogroup LG1. C. Sub-tree of phylogroup LG2.

**Fig. S3.** Phylogeny of Lake Gatun rhodopsin sequences within the context of known rhodopsin diversity from the nr database. Bootstrap support values less than 50% are not shown. Clades discussed in the text are labelled.

**Fig. S4.** Phylogeny of estuarine (Delaware Bay and Chesapeake Bay) rhodopsin sequences within the context of known rhodopsin diversity from the nr database. Bootstrap support values less than 65% are not shown.

**Fig. S5.** Phylogeny of Punta Cormorant rhodopsin sequences within the context of known rhodopsin diversity from the nr database. Bootstrap support values less than 65% are not shown.

**Fig. S6.** Rarefaction curves for rhodopsin protein sequences ≥ 150 aa and ≥ 70 aa from the Lake Gatun (GS020), marine estuary (GS011 and GS012 combined) and Punta Cormorant Lagoon (GS033) for each sample individually. (A), (C) and (E) display the rarefaction curves for rhodopsin sequences extracted from each sample site using the ≥ 150 aa length cut-off (see *Experimental procedures*). All sequences used in the analyses shown in (A) and (C) were trimmed to the same length in the alignment (no variable length sequence data). (B), (D) and (E) display those curves that were generated using a fragmented sequence library from each sample site that aligned partial rhodopsin sequences of variable sizes that were at least 70 aa in length.

**Fig. S7.** Primary sequence analysis of the amino acid residues present in the retinal-binding pocket of sequence clusters LG1, LG2 and PCL1 versus bacteriorhodopsin (BR) and proteorhodopsin (PR). Positions where all sequence types are identical are shaded grey. Positions where sequence clusters resemble bacteriorhodopsin or proteorhodopsin are coloured orange or blue respectively. Positions that were unique to sequence clusters are indicated in green. Positions of the residues in the retinal-binding pocket were determined using the rhodopsin protein sequence alignment from Spudich and Jung (2005). These positions are presented in the order they appear in their alignment as denoted by the asterisks. Note that the first amino acid listed for a cell in the PR column corresponds to the Sar86 sequence (AAG10475) and the following character states to other proteorhodopsin sequences contained within our alignment. The first amino acid listed for a cell in any of the sequence cluster columns corresponds to the most numerically dominant character state, followed by less frequent ones.
Table S1. Correspondence between sequence identifiers used throughout the manuscript and the CAMERA sequence read identifiers. Frame shows the frame on which rhodopsin sequence was found (assigned frame values 1, 2, 3, 4, 5, 6 denote +1, +2, +3, −1, −2, −3 frames respectively). The table is provided as a tab-delimited ASCII text file.

Table S2. Composition of 102 OTU clusters, representatives of which were used in the phylogenetic tree in Fig. 1. The table is provided as a tab-delimited text file.

This material is available as part of the online article from http://www.blackwell-synergy.com

Please note: Blackwell Publishing are not responsible for the content or functionality of any supplementary materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.