

Environ. Microbiol. Accepted

Novel Proteorhodopsin Variants from the Mediterranean and Red Seas

**Gazalah Sabehi¹, Ramon Massana², Joseph P. Bielawski³, Mira
Rosenberg¹, Edward F. Delong⁴ & Oded Béjà^{1*}**

¹Department of Biology, Technion- Israel Institute of Technology, Haifa
32000, Israel

²Departament de Biologia Marina i Oceanografia, Institut de Ciències del Mar,
CSIC, E-08003 Barcelona, Spain

³Department of Biology, University College London, Darwin Building, Gower
Street, London WC1E 6BT, United Kingdom

⁴Monterey Bay Aquarium Research Institute, Moss Landing, California 95039-
0628, USA

Running title: Proteorhodopsins from the Mediterranean and Red Seas

Keywords: Diversity; SAR86; Rhodopsin, Bacterioplankton

*For correspondence. E-mail beja@tx.technion.ac.il; Tel. (+972) 4829 3961;
Fax (+972) 4822 5153.

Abstract

Proteorhodopsins, ubiquitous retinylidene photoactive proton pumps, were recently found in the widespread uncultured SAR86 bacterial group in oceanic surface waters. To survey proteorhodopsin diversity, new degenerate sets of proteorhodopsin primers were designed based on a genomic proteorhodopsin gene sequence originating from an Antarctic fosmid library. New proteorhodopsin variants were identified in Red Sea samples that were most similar to the original green-light absorbing proteorhodopsins found in Monterey Bay California. Unlike green-absorbing proteorhodopsins however, these new variants contained a glutamine residue at position 105, the same site recently shown to control spectral tuning in naturally occurring proteorhodopsins. Different proteorhodopsin variants were also found in the Mediterranean Sea. These proteorhodopsins formed new and distinctive proteorhodopsin groups. Phylogenetic analyses show that some of the new variants were very different from previously characterized proteorhodopsins, and formed the deepest branching groups found so far among marine proteorhodopsins. The existence of these varied proteorhodopsin sequences suggests that this class of proteins has undergone substantial evolution. These variants could represent functionally divergent paralogous genes, derived from the

same or similar species, or orthologous proteorhodopsins that are distributed amongst divergent planktonic microbial taxa.

Introduction

Proteorhodopsins (PRs) are light-driven proton pumps (Béjà *et al.*, 2000a; Dioumaev *et al.*, 2002) belonging to the microbial-rhodopsin super family (Spudich *et al.*, 2000) that were recently identified based on environmental genomics techniques. The first PR identified was shown to be associated (Béjà *et al.*, 2000a) with the uncultured marine gamma proteobacterial group, SAR86 (Mullins *et al.*, 1995; Béjà *et al.*, 2000b; Eilers *et al.*, 2000; Gonzalez *et al.*, 2000; Rappé *et al.*, 2000; Kelly and Chistoserdov, 2001; Suzuki *et al.*, 2001; Zubkov *et al.*, 2001; Pernthaler *et al.*, 2002). Using biophysical techniques, PR proteins were later shown to be physically present in substantial quantities in bacterioplankton membranes isolated from the Pacific Ocean (Béjà *et al.*, 2001).

Two related PR subgroups, identified in the Pacific and the Southern Oceans, were found that absorb light with different absorption maxima, λ_{\max} 525 nm (green) and λ_{\max} 490 nm (blue). In the North Pacific subtropical gyre, PR distribution was shown to be stratified with depth, with green-absorbing pigments more prevalent at the surface and blue-absorbing dominant at depth, consistent with the ambient light availability (Béjà *et al.*, 2001). These PR subgroups are highly similar sharing more than 78% amino acid sequence

identity (200 out of 247 identical amino acid residues). The mechanism underlying the different absorption spectra of these highly similar PRs was recently pinpointed to one amino acid residue change (Man *et al.*, 2003). Amino acid substitutions at residue 105 were shown to be responsible for changes in absorption maxima, and naturally occurring variation at that position was found in Red Sea PR variants (Man *et al.*, 2003).

A third PR group, based on DNA comparisons, was recently identified in the Mediterranean Sea (see (Man *et al.*, 2003) and group III in figure 3A). The Mediterranean variants shared 88% DNA sequence identity with Monterey Bay and shallow water North Pacific gyre PRs (group I in figure 3A), and 69% with Antarctic and deep water North Pacific gyre varieties (group II in figure 3A). However, nearly all DNA differences between Monterey Bay PR cluster and the Mediterranean cluster were synonymous substitutions, showing only few differences in their predicted protein sequences (Man *et al.*, 2003). This striking feature might point to the evolution mechanisms of these two PR groups, evolution by genetic drift via geographic isolation of the two different water bodies. The few differences in their predicted protein sequences lead also to the speculation that this Mediterranean PR group was encoded by SAR86-like bacteria (Man *et al.*, 2003).

To better understand the extent of naturally occurring PR variability, we surveyed their sequence variation in the Mediterranean and Red Seas using PCR primers based on the full sequence of a blue-absorbing PR gene

discovered in an Antarctic fosmid library (5). We present here results on the identification and diversity of novel deep branching PR proteins based on newly designed PR PCR primers. The results add to the previously reported variation in naturally occurring PRs and further imply that these widespread pigments are important compounds in the marine environment.

Results & discussion

In our search for PR diversity, an Antarctic fosmid library was screened (Béjà *et al.*, 2002) for the presence of PRs using a PCR primer set used to clone the original 31A08 PR (Béjà *et al.*, 2000a). A fosmid-encoded blue-absorbing PR (fosmid 32C12) was identified and fully sequenced (de la Torre & DeLong, in preparation). When the region spanning the sense primer was compared to the original 31A08 sequence, two mismatches were identified (Fig. 1A). A new degenerate set of PR primers was designed in order to amplify more diverse PRs from the environment (Fig. 1B). Beside four mismatches change, the original primer was also shortened in its 3' end by one nucleotide, again in order to gain more diversity.

Samples from the Red and Mediterranean Seas were screened using the new primers in combination with primer PRrev. New PR sequences were identified in the Red Sea samples encoding new PR proteins (RED_29, REDs3a7 and REDr7a1a15 in Fig. 2 and in group I in Fig. 3). Interestingly, these new proteins were highly similar to the original green-absorbing PRs,

but like blue-absorbing rhodopsins contained a glutamine residue at position 105 (Fig. 2). Position 105 has recently been shown to control spectral tuning (shifts in absorption maxima) in environmental PRs (Man *et al.*, 2003).

The new PR variants recovered from the Mediterranean Sea comprised a number of new highly distinct PR groups when phylogenetic tree topologies were compared at the DNA level (Fig. 3A). The new groups shared between 87% (group IV) down to 66% (group XI) identity with previously characterized PRs (groups I-III). The assignment of new DNA established groups was based on identity percentage and no new group was assigned unless it had less than 90% identity to other groups. As more PR sequences get determined, it is possible that some of those DNA based groups will dissolve into a continuum.

When compared at the amino acid level, the new sequences form new PR protein groups (protein groups 4-10 in Fig. 3B) with some having as low as 60% identity over the entire protein length to previously reported sequences (compare to almost 80% identity between the original 31A08 clone and the Antarctic palE6 rhodopsin). The changes were not restricted to the loop regions only and were spread over the entire protein including the transmembrane helices (Fig. 2). In some cases, medA15_r8b9 in PR protein group 6 for example (DNA group VI in Fig. 2), there was an addition of 6 amino acids to the loop connecting helices B and C.

When phylogenetic trees of the entire microbial-rhodopsin super family proteins were constructed, some of the novel PR sequences created deep branching nodes (medA17r8_11, medA17r8_15 and medA15r8ex4 in Fig. 5). These new PR variants were positioned between the previously identified PR sequences and the rest of the microbial-rhodopsin super family and represent the most divergent PR sequences identified to date. In fact, the average distance along the tree between the deep branching PR sequences and the remaining PR sequences was ~ 0.53 amino acid substitutions per site, a value comparable to the divergence of *Neurospora crassa* and the dinoflagellate *Leptosphaeria maculans* rhodopsins (~ 0.51).

Some bacterial cytoplasmic membrane proteins (like the major coat protein of bacteriophage M13) require the presence of a cleavable signal peptide in their N-terminus, in order to be correctly targeted to the membrane. Such signal peptides have a common structure: a short, positively charged amino-terminal region; a central hydrophobic region; and a more polar carboxy-terminal region containing the signal peptidase cleavage site. In the case of the original 31A08 PR (Béjà *et al.*, 2000a), a region in the N-terminus was identified as a potential signal peptide. This region was found also in PR members from group II (paIE6 in Fig. 4). In the region of the protein signal peptide, divergence from the original 31A08 PR sequence is seen in proteins medA17r8_15 and medA15r8ex4 (Fig. 2). However, as shown in their hydropathy plots (Marck, 1988) their N-termini exhibit a pattern suggesting

the presence of signal peptides in these new PRs as well (Fig. 4). It is important to note that the presence or absence of a signal peptide in any potential bacterial rhodopsin does not necessarily affect the predicted topology, because in most bacterial membrane proteins the topological information is spread throughout the entire primary and secondary sequence (Broome-Smith *et al.*, 1994).

Phylogenetic analysis of the PR proteins revealed eleven distinct evolutionary lineages, indicated as “Groups” in figure 3A. Such genetically divergent lineages are expected to correspond to a level of ecological divergence sufficient for evolution under independent selection pressures (Cohan, 1994a, b; Palys *et al.*, 1997; Cohan, 2001). Due to the low rate of recombination in bacteria, novel mutants with adaptive value will “sweep” through a population and purge genetic variation at all the other loci (Cohan, 1994a, b). When lineages of bacteria are ecologically divergent (*i.e.*, ecotypes), this type of selection purges only the variation within the ecotype; hence, each selective sweep promotes further genetic divergence between ecotypes at all loci (Cohan, 1994a, b), which leads to genetically divergent lineages such as those found in the PR protein phylogeny. Neutral evolution in geographic isolation also could lead to genetically divergent lineages, even if they are members of the same ecotype (Cohan, 2002); however, this model is an unlikely explanation for the evolution of the major groups of PR proteins. Divergent members of the same ecotype would come into direct competition if

they subsequently colonized the same geographic region, leading to the extinction of the less-adapted lineages. Hence, highly divergent groups of PR proteins present in the same geographic region are more likely to represent different planktonic ecotypes.

It is possible that evolution by gene duplication also has contributed to the diversity of PR proteins. The PR variants we report here could reflect duplication and functional diversification within the PR gene family, producing functionally different paralogues that occur within the same or closely related species. This mode of evolution is not unprecedented for rhodopsins, as duplication of Archaeal rhodopsin may have resulted in the evolution of bacteriorhodopsin, halorhodopsin, and several sensory rhodopsins, all of which can be found within a single haloarchaeal species (Ihara et al., 1999; Mukohata et al., 1999).

In contrast to the pattern of divergence between PR groups, divergence within groups could be explained by geographical isolation. For instance, identical sequences were never found in different geographic locations, and a pattern of geographic endemism was apparent within protein Group 1. Specifically, Group I is divided between a largely Monterey Bay cluster and a largely Red Sea cluster (Figure 3A and 3B). As most of the differences between these two clusters are synonymous, it is tempting to speculate that this case represents largely neutral divergence in isolation, with the geographic outliers representing recent dispersals. However, we

cannot rule out the possibility that these synonymous changes represent neutral variants that were linked to a series of selective sweeps for adaptive amino acid mutations at other loci. Moreover, it is possible that the distributions of the relevant microbial taxa are global and that their relative abundances change with depth and time of year at a given geographic location. More detailed environmental sampling is required to resolve some of these issues.

The ecological distributions of the different variants we observed cannot fully resolve the above alternatives, since a number of different primer sets were used, and the relative efficiency of each in recovering different PR variants is not well quantified. Environmental genomic approaches have the potential to resolve some of the above uncertainties, and should shed more light on ecology and evolution within the PR gene family.

The new PR types reported here are derived from only one station in the Mediterranean, using a few simple modifications of a common PCR priming site. Our results likely reflect only a small sampling of the true diversity existing amongst the bacterial PR family. Further studies using a variety of molecular, biochemical and ecological approaches, should help better define the functional diversity within PRs, as well as their ecological and organismal distributions.

Regardless of their evolutionary origin and exact organismal dispersal, the enormous PR diversity reported here suggests that PR derived energy

mechanisms have a significant role in energy transduction to the world oceans

Material & Methods

Environmental DNA collection. Red Sea samples were collected from depths of 60 m and 150 m (fraction $>0.45 \mu\text{m}$) at a station in the northern Red Sea near the entrance to the Gulf of Suez (27.17°N , 34.22°E) on February 1999 (Lindell and Post, 2001). Mediterranean samples were from the Alborán Sea (36.0°N , 4.25°W) collected on May 1998. The Mediterranean samples were derived from filtered size fractions, as follows: sample A15- 5 m, fraction 0.2 to 2 μm ; sample A17- 50 m, fraction 0.2 to 2 μm ; sample A19- 100 m, fraction 0.2 to 5 μm . DNA was extracted from the samples according to (Massana *et al.*, 1997).

Environmental PR PCR amplification. PRs were amplified by PCR from DNA extracts obtained from environmental samples using modified PR forward primers (see Fig. 1) and the original reverse PR primer (Béjà *et al.*, 2000a) using independently two different high fidelity proof reading polymerase mixes (BIO-X-ACT™ from Bioline and TaKaRa Ex Taq™ from Takara Shuzo Co.). PCR amplification was carried out in a total volume of 20 μl containing 10 ng of template DNA, 200 μM dNTPs, 1.5 mM MgCl_2 , 0.2 μM primers, and 2.5 U of BIO-X-ACT DNA polymerase or TaKaRa Ex Taq™ polymerase. The amplification conditions comprised steps at 92°C for 4 min, and 35 cycles at 92°C for 1 min, 52°C for 1 min, and 68°C (BIO-X-ACT polymerase) or 72°C (TaKaRa Ex Taq™ polymerase) for 1 min. PCR products were cloned using the QIAGEN® PCR

cloning kit and unique *EcoRI* and *RsaI* RFLP groups were sequenced. All PCR products identical to samples handled or amplified previously in the lab were omitted from the analyses to avoid influence from possible contamination.

PR phylogenetic inference. Sequence alignments were performed using the Clustal X (1.81) program (Thompson *et al.*, 1997). Neighbour-Joining (NJ) and maximum likelihood (ML) analyses were conducted on PR amino acid and nucleotide datasets by using test version 4.0b10 of PAUP* (Swofford, 2002). Default parameters were used in NJ analyses. ML analysis for nucleotides was conducted under the HKY85 substitution matrix (Hasegawa *et al.*, 1985) combined with a discrete gamma model of among sites rate variation (Yang, 1994). Relative support for internal nodes was assessed by non-parametric bootstrapping (Felsenstein, 1985). Bootstrap resampling (100 pseudoreplications) of the MP, NJ, and ML trees was performed in all analyses to provide confidence estimates for the inferred topologies.

Amino acid distance values estimated by maximum likelihood were under the WAG substitution matrix (Whelan and Goldman, 2001) combined with empirical estimates of amino acid frequencies and a gamma model of among sites rate variation (Yang, 1994). Relative support for internal nodes was assessed by non-parametric bootstrapping (Felsenstein, 1985).

Note that the CODEML program of the PAML package (Yang, 1997) was used to obtain maximum likelihood estimates of the amino acid distances under the model described above. The CODEML program was also used to obtain the amino acid distances for each of the 100 pseudoreplicated datasets. Distance matrices from the pseudoreplicated data were summarized by using PAUP* (Swofford, 2002), which produced a bootstrap consensus tree.

Data deposition. Sequences reported in this paper have been submitted to GenBank under the following accession numbers: AY250714-AY250741.

Acknowledgments

This research was supported by the Human Frontiers Science Program P38/2002 (O.B.), Israel Science Foundation grant 434/02 (O.B.), J. S. Frankford, E. & J. Bishop, J. & A. Taub and the New-York Metropolitan research funds (O.B.), EU project PICODIV EVK3-CT1999-00021 (R.M.), grant G14969 from the Biotechnology and Biological Sciences Research Council (United Kingdom) (J.P.B), NSF grants OCE0001619, MCB-0084211 and the David and Lucile Packard Foundation (E.F.D.), and by support provided to MBARI by the David and Lucile Packard Foundation.

Figure Legends

Figure 1. Alignment of PR forward primer region. A. Comparison of the forward primer region of Antarctic fosmid 32C12 with the original PR clone (eBAC31A08). Nucleotide differences between the different variants are marked in white on black. B. Primers designed and used in this study. *Nco*I restriction enzyme site is underlined.

Figure 2. Analysis of PR sequences from the Red and Mediterranean Seas. Multiple protein alignment of PRs from the different DNA based groups. Residue differences between the protein variants are marked in white on red. Predicted transmembrane helices A, B, C, D, E, F & G are marked in grey. Different DNA group affiliations are marked on the right. Position 105 is marked with red number.

Figure 3. Phylogenetic analysis of DNA and inferred amino-acid sequence of PR variants. A. DNA phylogenetic tree based on distance analysis of 740 nt positions by neighbour-joining (NJ) using the uncorrected ("p") distance option of the PAUP* program. Maximum likelihood (ML) analysis under the HKY86 substitution matrix (Hasegawa et al., 1985) yielded a similar topology and is not shown. B. Neighbour-joining protein distance analysis of 220 positions. Analysis

of ML estimates of genetic distance under the WAG matrix (Whelan and Goldman, 2001) yielded a similar topology and is not shown. Bootstrap values greater than 50% are indicated above the branches in the following order NJ/ML. Scale bar represents number of substitutions per site. Bold names indicate the PRs that were identified in this study.

Figure 4. Hydropathy plots of PRs. Kyte-Doolittle hydropathy plots (Kyte and Doolittle, 1982) of selected PRs were performed using the DNA strider™ 1.3f15 program (Marck, 1988). The predicted signal-peptides are marked with a bold line.

Figure 5. Phylogenetic analysis of PR protein variants. Neighbour-joining phylogenetic analysis of PRs (PR prefix) with archaeal (BR, HR, SRI and SRII prefixes), *Neurospora crassa*, *Nostoc* and *Pyrocystis lunula* rhodopsins. Newly identified variants are marked with red. A similar phylogeny was inferred from ML estimates of genetic distances (Whelan and Goldman, 2001) and is not shown. Selected bootstrap values greater than 50% are indicated in the following order NJ/ML. Scale bar represents number of substitutions per site.

References

- Béjà, O., Spudich, E.N., Spudich, J.L., Leclerc, M., and DeLong, E.F. (2001) Proteorhodopsin phototrophy in the ocean. *Nature* **411**: 786-789.
- Béjà, O., Koonin, E.V., Aravind, L., Taylor, L.T., Seitz, H., Stein, J.L. et al. (2002) Comparative genomic analysis of archaeal genotypic variants in a single population, and in two different oceanic provinces. *Appl Environ Microbiol* **68**: 335-345.
- Béjà, O., Aravind, L., Koonin, E.V., Suzuki, M.T., Hadd, A., Nguyen, L.P. et al. (2000a) Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science* **289**: 1902-1906.
- Béjà, O., Suzuki, M.T., Koonin, E.V., Aravind, L., Hadd, A., Nguyen, L.P. et al. (2000b) Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. *Environ. Microbiol.* **2**: 516-529.
- Broome-Smith, J.K., Gnaneshan, S., Hunt, L.A., Mehraein-Ghomi, F., Hashemzadeh-Bonehi, L., Tadayyon, M., and Hennessey, E.S. (1994) Cleavable signal peptides are rarely found in bacterial cytoplasmic membrane proteins. *Mol Membr Biol* **11**: 3-8.
- Cohan, F.M. (1994a) Genetic Exchange and Evolutionary Divergence in Prokaryotes. *Trends Ecol Evol* **9**: 175-180.
- Cohan, F.M. (1994b) The Effects of Rare but Promiscuous Genetic Exchange on Evolutionary Divergence in Prokaryotes. *Am Nat* **143**: 965-986.
- Cohan, F.M. (2001) Bacterial species and speciation. *Syst Biol* **50**: 513-524.

- Cohan, F.M. (2002) What are bacterial species? *Annu Rev Microbiol* **56**: 457-487.
- Dioumaev, A.K., Brown, L.S., Shih, J., Spudich, E.N., Spudich, J.L., and Lanyi, J.K. (2002) Proton Transfers in the Photochemical Reaction Cycle of Proteorhodopsin. *Biochemistry* **41**: 5348-5358.
- Eilers, H., Pernthaler, J., Glockner, F.O., and Amann, R. (2000) Culturability and *in situ* abundance of pelagic bacteria from the North Sea. *Appl Environ Microbiol* **66**: 3044-3051.
- Felsenstein, J. (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**: 783-791.
- Gonzalez, J.M., Simo, R., Massana, R., Covert, J.S., Casamayor, E.O., Pedrós-Alió, C., and Moran, M.A. (2000) Bacterial community structure associated with a dimethylsulfoniopropionate-producing North Atlantic algal bloom. *Appl. Environ. Microbiol.* **66**: 4237-4246.
- Hasegawa, M., Kishino, H., and Yano, T. (1985) Dating the human-ape splitting by a molecular using clock mitochondrial DNA. *J Mol Evol* **22**: 160-174.
- Ihara, K., Umemura, T., Katagiri, I., Kitajima-Ihara, T., Sugiyama, Y., Kimura, Y., and Mukohata, Y. (1999) Evolution of the archaeal rhodopsins: evolution rate changes by gene duplication and functional differentiation. *J Mol Biol* **285**: 163-174.

- Kelly, K.M., and Chistoserdov, A.Y. (2001) Phylogenetic analysis of the succession of bacterial communities in the Great South Bay (Long Island). *FEMS Microbiol Ecol* **35**: 85-95.
- Kyte, J., and Doolittle, R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J Mol Biol* **157**: 105-132.
- Lindell, D., and Post, A.F. (2001) Ecological aspects of *ntcA* gene expression and its use as an indicator of the nitrogen status of marine *Synechococcus* spp. *Appl Environ Microbiol* **67**: 3340-3349.
- Man, D., Wang, W., Sabehi, G., Aravind, L., Post, A.F., Massana, R. et al. (2003) Diversification and spectral tuning in marine proteorhodopsins. *EMBO J.* **22**: 1725-1731.
- Marck, C. (1988) 'DNA Strider': a 'C' program for the fast analysis of DNA and protein sequences on the Apple Macintosh family of computers. *Nucleic Acids Res* **16**: 1829-1836.
- Massana, R., Murray, A.E., Preston, C.M., and DeLong, E.D. (1997) Vertical distribution and phylogenetic characterization of marine planktonic *Archaea* in the Santa Barbara channel. *Appl Environ Microbiol* **63**: 50-56.
- Mukohata, Y., Ihara, K., Tamura, T., and Sugiyama, Y. (1999) Halobacterial rhodopsins. *J Biochem* **125**: 649-657.
- Mullins, T.D., Britcshgi, T.B., Krest, R.L., and Giovannoni, S.J. (1995) Genetic comparisons reveal the same unknown bacterial lineages in Atlantic and Pacific bacterioplankton communities. *Limnol. Oceanogr.* **40**: 148-158.

- Palys, T., Nakamura, L.K., and Cohan, F.M. (1997) Discovery and classification of ecological diversity in the bacterial world: the role of DNA sequence data. *Int J Syst Bacteriol* **47**: 1145-1156.
- Pernthaler, A., Pernthaler, J., and Amann, R. (2002) Fluorescence *In Situ* Hybridization and Catalyzed Reporter Deposition for the Identification of Marine Bacteria. *Appl. Environ. Microbiol.* **68**: 3094-3101.
- Rappé, M.S., Vergin, K., and Giovannoni, S.J. (2000) Phylogenetic comparisons of a coastal bacterioplankton community with its counterparts in open ocean and freshwater systems. *FEMS Microbiol Ecol* **33**: 219-232.
- Spudich, J.L., Yang, C.S., Jung, K.H., and Spudich, E.N. (2000) Retinylidene proteins: Structures and Functions from Archaea to Humans. *Annu. Rev. Cell Dev. Biol.* **16**: 365-392.
- Suzuki, M.T., Preston, C.M., Chavez, F.P., and DeLong, E.F. (2001) Quantitative mapping of bacterioplankton populations in seawater: field tests across an upwelling plume in Monterey Bay. *Aquat. Microbiol. Ecol.* **24**: 117-127.
- Swofford, D.L. (2002) PAUP*. Phylogenetic Analysis Using Parsimony. In: Sinauer Associates, Sunderland, Massachusetts.
- Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., and Higgins, D.G. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res* **25**: 4876-4882.

- Whelan, S., and Goldman, N. (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol* **18**: 691-699.
- Yang, Z. (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J Mol Evol* **39**: 306-314.
- Yang, Z. (1997) PAML: a program package for phylogenetic analyses by maximum likelihood. *Cabios* **13**: 555-556.
- Zubkov, M.V., Fuchs, B.M., Burkill, P.H., and Amann, R. (2001) Comparison of cellular and biomass specific activities of dominant bacterioplankton groups in stratified waters of the Celtic Sea. *Appl Environ Microbiol* **67**: 5210-5218.