Web-site only available at ICM-CSIC

Final Grade Project

Creation and visualization of a database regarding diversity and distribution of proteorhodopsin-containing bacteria across the whole ocean

Mark Martori, Isabel Ferrera, Silvia G. Acinas and Pablo Sánchez

ESCI-UPF, Passeig de Pujades, 1, 08003 Barcelona, Spain
ICM-CSIC, Passeig Marítim de la Barceloneta, 37, 08003 Barcelona, Spain

ARTICLE INFO

Article history: Not publicly published
Received 6 June
Received in revised form 17 June

Supplementary information: All scripts are available at GitLab link: https://gitlab.com/markmartori/PR_DB

ABSTRACT

Phototrophs are organisms that carry out photon capture to acquire energy to carry out metabolic processes. In prokaryotes, light energy can directly drive proton expulsion from cells through proteins called rhodopsins. Among them, proteorhodopsins represent the simplest type of energy-harvesting photoproteins, consisting of a membrane protein bound to a light-sensitive pigment, the retinal. Proteorhodopsins were discovered in year 2000 in a genomic fragment of an uncultured marine bacterium and, ever since, thousands of proteorhodopsins (PR) have been identified from marine prokaryotes expanding their known phylogenetic range, environmental distribution and sequence diversity. In fact, now we know that up to 80% of marine prokaryotes in oceanic surface waters can harbor this light-driven proton pump. The aim of this project is to generate a database containing PR sequences and their associated environmental data linked to an easy and visual tool that will help scientists explore their diversity and distribution patterns in the oceans as well as the influence of environmental parameters such as temperature, salinity and chlorophyll a. The creation of a curated Proteorhodopsin Database (PR_DB) consisting of around 13.000 PR sequences worldwide has been the base of this project. Within these, 500 PR sequences were publicly available in literature whereas most of the sequences come from samples collected during the Tara Oceans (4000 sequences) and Malaspina Expeditions (7000 sequences). All the collected data has been organized by over 100 parameters to perform the comparisons as much detailed as possible and improve the quality of our database. During this project, the domain-specific language that has been used for managing the data is mySQL (Workbench). PYTHON has been the programming language used to create programs able to interconnect and create the Database and HTML the language used to print the Database into a better visualization way. The use of a Flask APP has been extremely necessary to interconnect both SQL and PYTHON with HMTL. The website offers a wide range of options to visualize and compare the data in several different ways. Additionally, a BLAST search against this new Database was implemented too, allowing comparisons with outsider sequences. In summary, we offer the first integrated Marine Proteorhodopsins Database, a valuable resource for the scientific community with interest to explore the diversity and distribution of PR at global scales as well as to uncover potential novel proteorhosopsins.

* *Corresponding author.* Tel.: +34 93 676 01 34
E-mail address: mark.martori@alum.esci.upf.edu

1. Introduction

Borrowed from Latin "Lūx", light has been known for noticeable years to be used for obtaining visual information but at the same time for having huge impact on bodily functions. It is the most prevalent phenomenon therefore life has to cope with it. In humans, the non-visual information via light falls onto our eyes and is conveyed via a nerve connection to the suprachiasmatic nucleus **(01)**. In prokaryotes light can be a threat or an asset so they have developed a variety of molecules to deal with it. Carotenoids (which first emerged in archaea as lipids reinforcing cell membranes) were some of the earliest light-absorbing molecules appearing in evolution. Among proteins, some molecules such as sensory rhodopsins [SRs] or phototropin [LOV] act as light sensors having an important role on the behavior and life history of organisms. Regarding light-driven energy-generating mechanisms we can find chlorophyll-based (very complex systems) or retinal-based organisms (e.g., rhodopsins) which consist of only one protein opsin (integral membrane protein) and one chromopore, retinal (Vitamin A aldehyde produced in one metabolic step from the widely distributed carotenoid beta-carotene) which binds opsins **(02)**.

By their amino acid sequences, opsin proteins are classified into two 'a priori' very different groups: Type I which are found in Bacteria, Archaea and some Eukaryotic microbes and Type II which functions as photoacivated G-protein coupled receptors in animal vision **(03)** even though they seem more likely to share an ancestor. In 1970 the first microbial rhodopsin named "bacteriorhodopsin" or BR was isolated from the cell membrane of *Halobacterium salinarum* by Stoeckenius and Oesterhelt's study **(04)**. The protein has a molecular mass of 26.8 kDa and forms two-dimensional crystalline patches. BR are considered light-driven proton pumps because a proton is transported out of the cell during the protein's photocycle **(05)**. Nowadays, the discovery and research of microbial rhodopsin is a flourishing field in which new understandings of rhodopsin diversity, function and evolution is expanding our knowledge of prokaryotes.

The amount of rhodopsin-like proteins research started to increase significantly around 19 years ago when a new type of rhodopsin derived from bacteria was discovered through genomic analyses of naturally occurring marine bacterioplankton. It was encoded in the genome of an uncultured gamma-proteobacterium (SAR86) so they named it 'Proteorhodopsin' and it was functionally expressed in *Escherichia coli* and forming an active light-driven proton pump **(07)**. Due to this impulse microbial rhodopsins started to get recognized to be not just an exception but probably the rule, existing in more than half of the heterotrophic bacteria living in the surface ocean. Proteorhodopsins are now known to be the most abundant rhodopsins in our planet **(06)** which raise new questions about the importance of light to microbial communities.

1.1. Proteorhodopsin

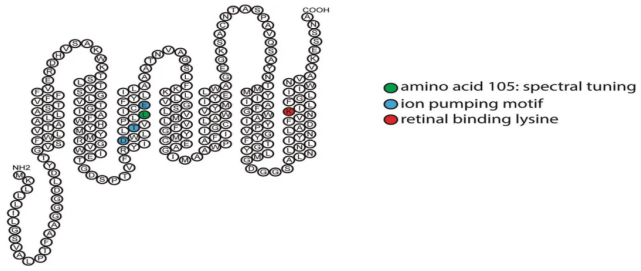
Proteorhodopsin (PR) are light-driven proton pumps **(13)** that belong to the archaeal/bacterial/fungal opsin family and at least two subgroups have been identified in the Pacific and Southern Oceans. One of them seems to be transformed at low pH to a containing 9-cis retinal species under uninterrupted illumination at lambda greater than 530nm absorbing around 430nm and therefore returning blue-light illumination. On the other hand, some proteorhodopsins discovered in Monterey Bay, California, seemed to return a green-light absorbing illumination under lambda close to 490nm. What's more, the main mechanism behind this huge difference seems to be related to the amino acid substitutions at residue 105 **(33)**. This indicates that protonation of acidic groups alters the photoreaction pathway that leads normally to the switch from all-trans retinal to 13-cis retinal.

In the North Pacific subtropical, PR distribution was shown to be stratified with depth, being the green-light absorbing subgroup more probable at the surface and the blue-light absorbing dominant at depth. This different proteorhodopsins however share more than 78% amino acid sequence identity **(08)**. However, despite the elevated number of studies with focus on diversity and distribution of proteorhospsin there is still not available a global PR dataset as community resource integrating all PR available to date to further explore ecological question regarding the diversity and distribution of different light absorbing PR across the global oceans and depths. For that reason, we created the Marine PR-Database (PR_DB), taking advantage of two global expeditions *Tara* Oecans and *Malaspina* expedition in which a large scale metagenomic has been done and therefore data containing proteorhodopsin genes were generated along with associated environmental data such as depth and other physico-chemical parameters that could be relevant for proteorhodopsin distribution.

1.2. Structure

Bacterial rhodopsins are retinal-binding proteins that provide light-dependent ion transport and sensory functions to a family of halophilic and other bacteria. They are integral membrane proteins with seven transmembrane helices, the last of which contains the attachment point for retinal (a conserved lysine) around amino acid 230. The expected length is around 245 amino acids. [Figure 1]

Figure 1 – Retinal conformation: Topology of a representative proteorhodopsin highlighting key residues in retinal binding, spectral tuning and ion pumping. (18).



2. Materials & Methods

2.1. Sampling and sequence extraction

A total of 4759 proteorhodopsin-containing sequences were obtained from samples collected between September 2009 and 2012 from 179 different located across the world sampling stations during the *Tara* Oceans Expedition**[S1]**. The sampling done during the expedition combined traditional and novel methods **(09)** and the sequence were provided by the Acinas Lab at the Marine Sciences Institute (ICM-CSIC) in Barcelona. A total of 7107 nucleotide and amino acid proteorhodopsin sequences were obtained from metagenomica data generated from 116 different sampling stations at very diverse depths **[S2](10)** from the Malaspina 2010 Expedition, likewise provided by the ICM-CSIC.

Additionally, an amount of 400 proteorhdopsin-containing gene sequences were collected from PFAM, NCBI and Uniprot databases by the extraction of accession numbers searched in different already published articles **(13-32)**, some were obtained through keyword searches as proteorhodopsin, rhodopsin or Bacteriorhodopsin **[S1]**. Summing up, a total of 12.266 sequences have been used to create the PR database.

2.2. Accession numbers

The PR gene nucleotide sequences and its Accession IDs from NCBI can be accessed through the web-page.

2.3. Curation of sequences

All publicly available sequences were firstly compared to already correctly annotated PR sequences in order to avoid non-reviewed or misinformation by using BLAST **(P1)** above a 75% identity regarding nucleotides. Then, PSI-BLAST **(P2)** was the tool selected for further and better comparisons. Regarding non-publicly available PR sequences, alignments using MUSCLE **(P3)** have been provided in order to understand which amino acid was responsible for tuning color, in position 105.

2.4. Database content and organization

The Proteorhodopsin Database contains most of the currently available sequences regarding proteorhodopsins encoded in NCBI, Pfam and Uniprot as well as the very recent found ones in the previous mentioned expeditions (unpublished). An average of 119 different parameters for each single sequence have been implemented, for instance, Chlorophyll *a* concentration, collection depth, latitude and longitude and other crucial physico-chemical parameters. The full information content in the database can be achieved in the Supplementary Materials **[S3,S4,S5,S7].** The structure of the database can be divided in 3 entities. The genes that have been used and for which each one has an accession number and its personal information such as Taxonomy. The publications selected for information retrieval and the sampling stations of the expeditions with its metadata. Using a workbench SQL database, we have linked all the records in such a way that a client from an exterior via like a query on a web-page can interact and obtain with the whole data. To program and organize the structure, we have used the latest version of Python3.7.2 which offers some flexibility with openpyxl and collection packages to manipulate big data with different formats **(P4)**.

2.5. Web server design and implementation

To share the work done and the information collected in those circumnavigation expeditions, we have created a web-site in order to facilitate the search and understanding of the interpretation of the results so far by scientists working at the ICM-CSIC interested on it and with the goal to make it publicly available worldwide. HTML has plenty of libraries dedicated to visualization and easy access to data. Cascading Style Sheet files allow the formatting content on web pages to get displayed as the developer desire, therefore facilitating clients understanding and interpretation and Javascript provides dynamism to the server. The PR Database brings a user-friendly web interface for searching, querying, filtering and blasting. The first 3 parts were implemented using Python as the

main language, with a Flask App as the microframework based on Werkzeug (a toolkit for Web Server Gateway Interface Applications) and Jinja2, querying and searching is done inside the database which is managed with MySQL and Workbench environment and blasting using a local blast free tool 'Sequenceserver'**(11)**.

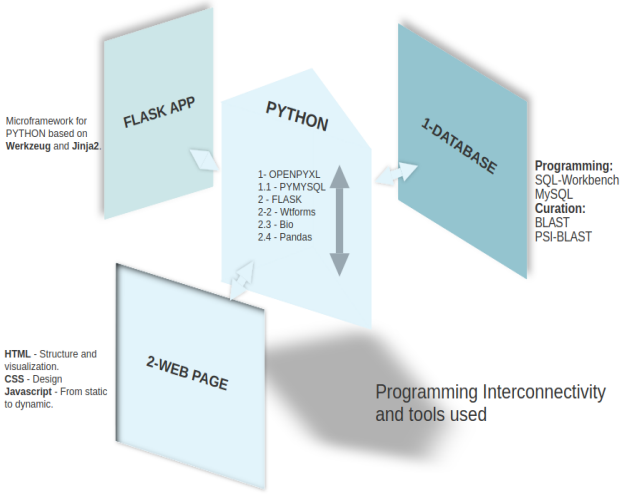


Fig. 2 – Inter-connectivity: This graph illustrates the programming languages used, the main libraries and how are the database and the web page implemented and connected. Flask app works as the Python micro-framework .

3. Results and discussion

3.1.The Marine PR-Database

In the database we included a total of 12.662 proteorhodopsin-containing sequences extracted worldwide, containing associated information on 119 physico-chemical parameters from 302 different sampling stations. We followed the Longhurst province marine distribution **[S6]** and added general geography location. From the data, approximately half of the amount of sequences have been given a detailed taxonomy while the rest were related to uncultured bacteria. Regarding publicly available sequences, temperature was annotated in less than 50% but concerning the rest, this parameter has been recorded for every single gene sequence as well as main metadata such as collection depth, oxygen, carbon total and chlorophyll *a* concentrations. The data was also divided regarding the amino acid in position 105, related to light tuning, in 3 main categories, sequences containing a Met or a Leu at this location were saved as green light-absorbing sequences, instead when Gln was found, blue color tuning was recorded; the rest are annotated as 'ND' for Not Determined. Sequences were not starting from the same position, therefore we based this categorization by searching for an amino acid pattern (TVPL) which indicated that the following amino acid was responsible for color tuning.

Taking into account the complementary circumnavigation tracking of the two global *Tara* Oceans and Malaspina expeditions, the main oceanic regions has been recorded in our study, which will allow us to draw more general conclusions. On the other hand, taking advantage of the time it took to perform the expeditions (Malaspina – 1 year; *Tara* Oceans – 3 years), we can take into account seasonality in our search within our PR-DB. Abundance for each gene within a huge subset of this PRs has been recorded, outputting same gene sequence in many different sampling stations giving a vast variety of metadata for each entry. To handle this, the database relationship must be consistent. We used the Gene ID as primary keys as well as the combination of Gene IDs and the sampling station where it was found, in another entity. This way facilitates all kind of querys. Foreign keys were needed to interconnect tables when searching for data. The web-site was created in order to share the work done and the information collected in those expeditions as well as to facilitate the search and understanding of the interpretation of the results by scientist belonging the ICM-CSIC.

3.2.Web-server

The web-site aggregates 4 main paths. The first interface provides unique information regarding a single gene. It allows the user to entry as input an Accession ID, once this data has been properly validated, the database is queried and the result is output in the computer screen. Therefore, providing to the user several physico-chemical metadata specific for that proteorhodopsin such as its taxonomy, depth, temperature, etc…Every time a purchaser entries data, a unique session is created in order to save the input for further use.

The Pubmed ID interface follows same procedure as the previous mentioned. In this case, the input will be an article ID regarding proteorhodopsin and the output is based on the information found on the scientific paper. Both kinds of results are displayed in a data frame format. Scientists will no longer need to search by keywords in those articles, just input the ID through this interface and retrieve the genes studied or used by the authors. Furthermore, we implemented the possibility to download the proteorhodopsin-containing genes sequences in fasta format, both in amino acids and nucleotide sequence by asking the session and querying the database again.

The Metadata interface allows the client to manually specify the parameters for a more concrete search. In this case, depth range varies concerning ocean layers (from 0 to 25m, 26 to 200m, 201 to 1000m and from 1001 to 4000m), upper photic, lower photic, mesopelagic and

bathypelagic, respectively. The user is also able to filter the search by the fraction collected size, both lower (from 0.2µ to 0.8µ) and upper (0.6µ-5µ). Temperature and season of the year are also a client-select choice. Regarding geography, this interface offers a search based on Longhurst province marine distribution such as SATL, GUIA, ARCT and the rest. Last displayed option is Color Tuning which is based on the different amino acid in position 105.

The BLAST interface provides an easy homology search between aligned sequences as well as conserved domains in the case of proteins such as proteorhodopsins. The big amount of biological data in the form of both nucleotide and amino acid, are extremely complex and difficult to understand at a first glance. This database will help scientists identifying sequence similarity across genes, to extract right primers for research work, to understand mutations, to build up taxonomy relationships but mainly this tool will allow analysis and comparisons related to such important and recently discovered organisms. This BLAST server detects BLAST software -if absent the correct version is automatically downloaded, it identifies existing BLAST databases as many as there are in the system, it allows to paste query sequences or dropping a FASTA file to search. It uses advanced parameters that the client can display in the command line such as the *e* value, maximum number of targets or score limitation. It provides the option of swapping from nucleotides to amino acids as TBLASTX would do. It outputs a number of best hits and its respective comparison at nucleotide or amino acid level.

Once we have finished building the database, we can not only compare all the genes to each other, either by depth, temperature or location, but we can try to solve questions that remain unknown. One of the questions related to our hypothesis consists of the following. We know that rhodopsins are not chlorophyll-based proteins and we also know that chlorophyll *a* is found in every single photosynthesizing organism, from plants to algae and cyanobacteria. These other chlorophyll types still absorb sunlight and assist in photosynthesis. Relying on our database, scientist will be able to contrast the abundance of chlorophyll *a* and the quantity of proterohodopsin-containing genes found at several positions in the ocean, or what's more, their color tuning. Given this tool scientists are closer to understand the correlation between this compound concentration and the appearance of this type of non-photosynthetic based proteins. From another point of view, yet using the same Database, scientists will be able to infer the optimal conditions under which the diversity of PR genes is favored due to the amount of variety related to coordinates and parameters that this tool offers. Comparisons between different places on the earth or in the same location but at different depths. This database is a great tool to put into use the observations given during years of PR studies and to establish relationships with available environmental parameters to further understand the ecology and function of microorganism containing the Proterhodopsin. The Public Available genes retrieved for this study have a main purpose. Over the past years, many computational methods have been developed to predict function through identifying sequence similarity between a protein of unknown function and one or more proteins with experimentally characterized or computationally predicted functions. However, it is widely recognized that functional annotations should be transferred with caution. Using our database it is not possible to reveal the taxonomy and function of unknown genes significantly but we envisioned the possibility to uncover novel PR genes that could be experimentally validated. As we have explained, some proteorhodopsin functionality falls into improving survival in starvation conditions. Many of these genes are integrated inside the Database, if we find a significant sequence similarity with the genes collected in any of the two expeditions, we can create experiments to provide experimental evidences about their functions.

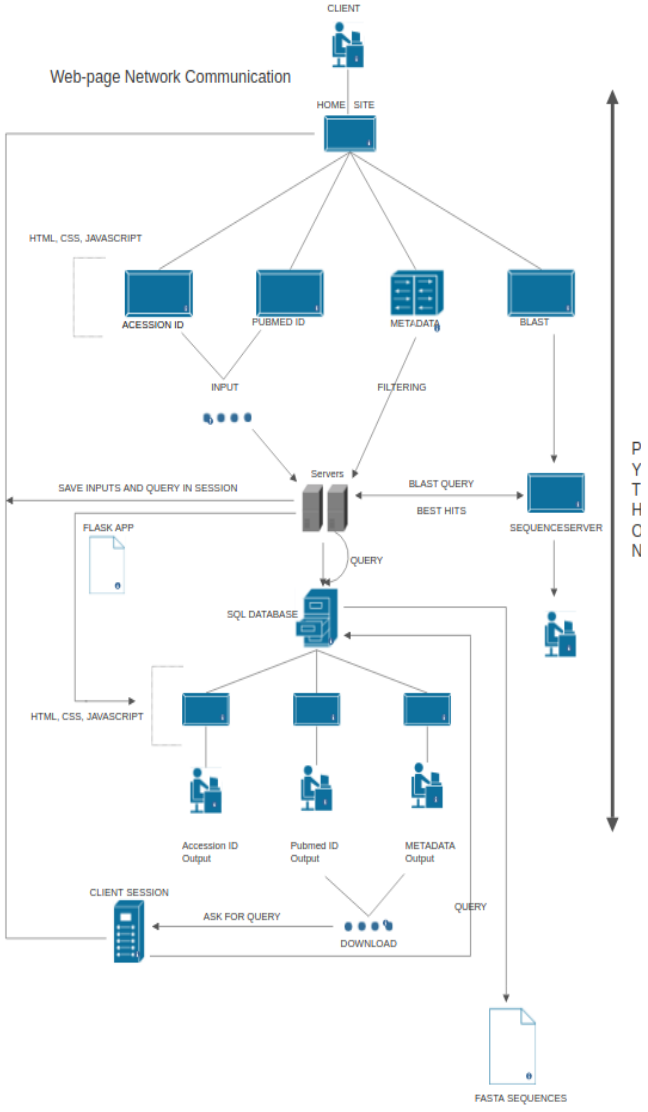


Figure 3 – Example of Use of the Web-server and how does it work.

4. Conclusions

This study represent a first attend to built a very-well curated Marine PR- database to explore in a really quick manner such as a few 'clicks', relevant ecological questions. The inquiries that we can answer deal with correlations between the presence or distribution of specific PR sequences with physico-chemical and/or biological parameters such as Salinity, Carbon total concentration, O_2, CO_2, et. Such mentioned relationships if significant would help to understand the ecological niches of specific PR-containing bacteria or archaea. We used the exact position (latitude, longitude and depth) in which the genes were sampled and obtained in order to establish a distribution across the whole ocean of the proteorhodopsin protein. It's also very surprising the fact that very distant (in important environmental compounds such as temperature) gene sequences like those from Arctic sampling stations and those from the Mediterranean have much more similarities than we expected. Overall, the built dataset will allow researchers to easily access results from PR sequence analyses without doing time-intensive literature searches, enabling to answer some of the ecological questions listed above as well as developing new hypothesis in relation to PR-containing organisms.

Acknowledgments

We thank the R/V Hesperides crew, the chief scientists in Malaspina legs, and all project participants for their help in making this project possible. Malaspina 2010 Expedition was funded by the Spanish Ministry of Economy and Competitiveness (MINECO) through the Consolider-Ingenio program (ref. CSD2008-00077) to Carlos M. Duarte. Additionally, this work has been supported by the grant Malaspinomics CTM2011-15461-E. We also thank our fellow scientists, the crew and chief scientists of the different cruise legs involved in *Tara* Oceans for collecting the samples used in this study and to Tara ocean consortium for the public available metagenomics dataset as well as all the authors of the articles used. This work was supported by the ICM-CSIC and ESCI-UPF.

5. References

1. Hegemann P. - (2008). Algal sensory photoreceptors. Annu Rev Plant Biol 59:167–189. http://dx.doi.org/10.1146/annurev.arplant.59.032607.092847.

2. Mackin KA, Roy RA, Theobald DL. - (2014). An empirical test of conver- gent evolution in rhodopsins. Mol Biol Evol 31:83–95.

3. Grote M, O'Malley MA. - (2011). Enlightening the life sciences: the history of halobacterial and microbial rhodopsin research. FEMS Microbiol Rev 35:1082–1099. http://dx.doi.org/10.1111/j.1574-6976.2011.00281.x.

4. Norgård S., Aasen A. J., Liaaen-Jensen S. - (1970) . Bacterial carotenoids. 32. C50-carotenoids, Carotenoids from Corynebacterium poinsettiae including four new C50-diols.

5. Finkel OM, Béjà O, Belkin S. - (2013). Global abundance of microbialrhodopsins. ISME J 7:448 – 451. http://dx.doi.org/10.1038/ismej.2012.112.

6. Hideki Kandori. - (2015). Ion-pumping microbial rhodopsins.

7. Gazalah Sabehi, Ramon Massana, Joseph P. Bielawski, Mira Rosenberg, Edward F. Delong and Oded Béjà. - (2003). Novel Proteorhodopsin variants from the Mediterranean and Red Seas.

8. John S. G, Mendez C. B, Deng L, Poulos B, Kauffman A. K. M, et al. (2010) A simple and efficient method for concentration of ocean viruses by chemical flocculation. Environ Microbiol Rep 3: 195–202.

9. Achermann, P., Borbély, A.A. (1994): Simulation of daytime vigilance by the additive interaction of a homeostatic and a circadian process. Biol. Cybern. Vol. 71, Nr. 2, S. 115-121.

10. Emilie Villar, Gregory K. Farrant, Michael Follows, Laurence Garczarek – (2015). Environmental characteristics of Agulhas rings affect interocean plankton transport.

11. Anurag Priyam, Ben J. Woodcroft, Vivek Rai, Alekhya Munagala, Ismail Moghul, Filip Ter, Mark Anthony Gibbins, HongKee Moon, Guy Leonard, Wolfgang Rumpf – (2015). Sequenceserver: a modern graphical user interface for custom BLAST databases.

12. Lionel Guidi, Samuel Chaffron, Lucie Bittner, Damien Eveillard, Abdelhalim Larhlimi, Simon Roux, Youssef Darzi. - (2016). Plankton networks driving carbon export in the oligotrophic ocean.

13. Oded Béjà, L. Aravind, Eugene V. Koonin, Marcelino T. Suzuki, Andrew Hadd, Gates, Robert A. Feldman, John L. Spudich, Elena N. Spudich, Edward F. DeLong. -(2000). Bacterial Rhodopsin: Evidence for a New Type of Phototrophy in the Sea.

14. Oded Beja , Elena N. Spudich, John L. Spudich, Marion Leclerc & Edward F. DeLong. - (2001). Proteorhodopsin phototrophy in the ocean.

15. Bo WEI. - (2012). Diversity and distribution of proteorhodopsin-containing microorganisms in marine environments.

16. Yohei Kumagai ,Susumu Y. Kogure ,Edward F. DeLong Wataru Iwasaki. - (2018). Solar-panel and parasol strategies shape the proteorhodopsin distribution pattern in marine Flavobacteriia

17. Dominique Boeuf, Raphaël Lami ,Emelyne Cunnington and Christian Jeanthon. Summer. - (2016). Abundance and Distribution of Proteorhodopsin Genes in the Western Arctic Ocean.

18. Daniel K. Olson 1 Susumu Yoshizawa, Dominique Boeuf, Wataru Iwasaki, Edward F. DeLong - (2018). Proteorhodopsin variability and distribution in the North Pacific Subtropical Gyre.

19. Marta Royo-Llonch, Isabel Ferrera, Francisco M. Cornejo-Castillo, Pablo Sánchez, Guillem Salazar, Ramunas Stepanauskas, José M. González, Michael E. Sieracki, Sabrina Speich, Lars Stemmann , Carlos Pedrós-Alió and Silvia G. Acinas. - (2017). Exploring Microdiversity in Novel Kordia sp. (Bacteroidetes) with Proteorhodopsin from the Tropical Indian Ocean via Single Amplified Genomes.

20. Meiru Zhao, Feng Chen, Nianzhi Jiao - (2009). Genetic Diversity and Abundance of Flavobacterial Proteorhodopsin in China Seas.

21. Susumu Yoshizawa, Akira Kawanabe, Hiroyasu Ito, Hideki Kandori and Kazuhiro Kogure - (2012). Diversity and functional analysis of proteorhodopsin in marine Flavobacteria.

22. Alina Pushkarev and Oded Béjà - (2016). Functional metagenomic screen reveals new and diverse microbial rhodopsins.
23. Yong Min Kwon, So-Young Kim ,Kwang-Hwan Jung & Sang-Jin Kim - (2015). Diversity and functional analysis of light driven pumping rhodopsins in marine Flavobacteria.
24. Ella T. Sieradzki, Jed A. Fuhrman, Sara Rivero-Calle and Laura Gómez-Consarnau - (2018). Proteorhodopsins dominate the expression of phototrophic mechanisms in seasonal and dynamic marine picoplankton communities.
25. Edward F. DeLong, Oded Béjà - (2010). The Light-Driven Proton Pump Proteorhodopsin Enhances Bacterial Survival during Tough Times.
26. Laura Gómez-Consarnau, José M. González, Montserrat Coll-Lladó, Pontus Gourdon, Torbjörn Pascher, Richard Neutze , Carlos Pedrós-Alió & Jarone Pinhassi - (2007). Light stimulates growth of proteorhodopsin-containing marine Flavobacteria.
27. Laura Gómez-Consarnau , Neelam Akram, Kristoffer Lindell , Anders Pedersen , Richard Neutze,Debra L. Milton, José M. González, Jarone Pinhassi - (2010). Proteorhodopsin Phototrophy Promotes Survival of Marine Bacteria during Starvation.
28. Hideki Kandori - (2015). Ion-pumping microbial rhodopsins.
29. Keiichi Inoue, Yoshitaka Kato , and Hideki Kandori - (2015). Light-driven ion-translocating rhodopsins in marine bacteria.
30. Jarone Pinhassi,Edward F. DeLong, Oded Béjà, José M. González,Carlos Pedrós-Alió - (2015). Marine Bacterial and Archaeal Ion-Pumping Rhodopsins: Genetic Diversity, Physiology, and Ecology.
31. Donald A. Bryant and Niels-Ulrik Frigaard – (2006). Prokaryotic photosynthesis and phototrophy illuminated.
32. Dikla Man, Weiwu Wang, Gazalah Sabehi, L.Aravind, Anton F.Post, Ramon Massana, Elena N.Spudich, John L.Spudich, Oded Béjà – (2013). Diversification and spectral tuning in marine proteorhodopsins.

Supplementary Material

S.1. Tara Oceans Expedition - Information regarding the Expedition route, the amount of time and more specific details.

S.2. Malaspina 2010 Expedition - Information regarding the Expedition route as well as some of the parameters extracted.

S.3. GENE Entity – Database - Gene Ids and its unique data used in this study as taxonomy, gene providence, and color tuning. Due to privacy, fasta sequences will not be provided.

S.4. GENE PUBMED Entity – Database – Each gene with its Pubmed ID and metadata associated.

S.5. GENE SAMPLE Entity – Database – Abundance of each gene.

S.6. Longhurts_Code.

S.7. PUBMEDs – Proteorhodopsin related articles.

P1. - BLAST - https://blast.ncbi.nlm.nih.gov/Blast.cgi
P2. - PSI-BLAST - https://blast.ncbi.nlm.nih.gov/Blast.cgi

P3. - MUSCLE - https://www.ebi.ac.uk/Tools/msa/muscle/P4. - PYTHON - https://www.python.org/