Title

A Large and Phylogenetically Diverse Class of Type 1 Opsins Lacking a Canonical Retinal Binding Site

Authors

Becker, Erin A
Yao, Andrew I
Seitzer, Phillip M
et al.

A Large and Phylogenetically Diverse Class of Type 1 Opsins Lacking a Canonical Retinal Binding Site

Erin A. Becker[1,2], Andrew I. Yao[1,3], Phillip M. Seitzer[1,3,4], Tobias Kind[1], Ting Wang[1], Rich Eigenheer[5], Katie S. Y. Shao[6], Vladimir Yarov-Yarovoy[7], Marc T. Facciotti[1,2,3]*

1 Genome Center, One Shields Ave., University of California Davis, Davis, CA, 95616, United States of America, 2 Microbiology Graduate Group, One Shields Ave., University of California Davis, Davis, CA, 95616, United States of America, 3 Department of Biomedical Engineering, One Shields Ave., University of California Davis, Davis, CA, 95616, United States of America, 4 Proteome Software, 1340 SW Bertha Blvd., Portland, Oregon, United States of America, 5 California Department of Food and Agriculture, 1220 N St., Sacramento, CA, 95814, United States of America, 6 William's College, 880 Main St., Williamstown, MA, 01267, United States of America, 7 Department of Physiology and Membrane Biology, One Shields Ave., University of California Davis, Davis, CA, 95616, United States of America

* mtfacciotti@ucdavis.edu

Abstract

Opsins are photosensitive proteins catalyzing light-dependent processes across the tree of life. For both microbial (type 1) and metazoan (type 2) opsins, photosensing depends upon covalent interaction between a retinal chromophore and a conserved lysine residue. Despite recent discoveries of potential opsin homologs lacking this residue, phylogenetic dispersal and functional significance of these abnormal sequences have not yet been investigated. We report discovery of a large group of putatively non-retinal binding opsins, present in a number of fungal and microbial genomes and comprising nearly 30% of opsins in the *Halobacteriacea*, a model clade for opsin photobiology. We report phylogenetic analyses, structural modeling, genomic context analysis and biochemistry, to describe the evolutionary relationship of these recently described proteins with other opsins, show that they are expressed and do not bind retinal in a canonical manner. Given these data, we propose a hypothesis that these abnormal opsin homologs may represent a novel family of sensory opsins which may be involved in taxis response to one or more non-light stimuli. If true, this finding would challenge our current understanding of microbial opsins as a light-specific sensory family, and provides a potential analogy with the highly diverse signaling capabilities of the eukaryotic G-protein coupled receptors (GPCRs), of which metazoan type 2 opsins are a light-specific sub-clade.

Introduction

Opsin proteins catalyze light-dependent processes in all three domains of life, including vision and circadian cycling in animals [1], as well as chlorophyll-independent phototrophy,

osmoregulation and phototaxis in bacteria, archaea, microbial eukaryotes, and multi-cellular fungi [2]. The opsin proteins have been classified into two main categories, the type 1 and type 2 opsins. Type 1 opsins are found in diverse species and in all three domains of life. They have been shown to function as light-driven ion transporters and phototaxis receptors. Type 2 opsins by contrast are found in metazoan species and serve primarily as light dependent photoreceptors in animal eyes and various other tissues of higher eukaryotes [3]. The evolutionary relationships between these two classes of opsins remains unresolved [4–6], however, in both groups, photosensing depends upon a covalent interaction between a conserved lysine residue in the seventh transmembrane helix (bovine visual rhodopsin K296 / bacteriorhodopsin K216) and a retinal chromophore [7,8]. Recent studies have reported discovery of genes encoding opsin homologs lacking this residue in fungal, haloarchaeal and placozoan genomes [3,9,10]. However, these have been treated as isolated instances and the phylogenetic dispersal and functional significance of these abnormal sequences have not yet been investigated.

Here we report discovery of a large group of putatively non-retinal binding opsins comprising nearly 30% of opsin homologs in the archaeal family *Halobacteriacea*, a historically important model clade for study of opsin photobiology. This family of extremely halophilic archaea possesses a diverse range of opsins, which have classically been divided into four groups: the ion pumps halorhodopsin (HR) and bacteriorhodopsin (BR), which respectively regulate cytoplasmic osmolarity and create electrochemical gradients used in ATP production; and two classes of sensory rhodopsins (SRI and SRII), which serve as histidine kinase response regulators for phototactic and photophobic behaviors [11]. Recent studies have expanded our view of haloarchaeal opsin diversity by revealing a third sensory rhodopsin (SR3), a second group of bacteriorhodopsins (BR2), and a proposed intermediate between bacteriorhodopsin and the sensory rhodopsins (MR) [11–14]. Studies of these diverse haloarchaeal opsins have led to major advances in our understanding of the kinetics and structural intermediates of opsin photocycles [15], spectral tuning [16], and signal transduction pathways [17]. Haloarchaeal opsins have also served as models for protein crystallization [17], membrane protein folding [18] and development of optogenetic toolkits [19].

Based on genomic context, phylogenetic analyses, structural modeling and biochemistry, we propose that these abnormal opsin homologs, which are also present in some fungal, cyanobacterial, and chlorophytal genomes, may represent a novel family of sensory opsins potentially involved in taxis response to one or more non-light stimuli. If true these finding challenge current understanding of microbial type 1 opsins as a light-specific sensory family, and provides a potential analogy with the highly diverse signaling capabilities of the eukaryotic G-protein coupled receptors (GPCRs), of which type 2 opsins are a light-specific sub-clade. These results call for more work on this novel protein family and a renewed perspective on the roles of type 1 opsins in microbial physiological responses to diverse environmental inputs.

Results and Discussion

A large-scale survey of 80 complete and high-quality draft haloarchaeal genomes [20] revealed a novel, large opsin class, consisting of 48 homologs lacking the normally conserved lysine residue (K216) required for binding retinal out of 170 total haloarchaeal opsins (Fig 1, sequences available at the Dryad Digital Repository (http://dx.doi.org/10.5061/dryad.963hr)). Given their close evolutionary relationship with opsin proteins we term these opsin-related proteins (ORPs) thus comprises nearly 30% of all known haloarchaeal opsins. We deliberately use the term opsin to describe the ORPs, even though the term is traditionally used to describe the apoprotein of the retinal-bound rhodopsin, due to their proposed constitutively retinal-free nature and phylogenetically close relationship to sensory opsins. The ORPs are broadly distributed
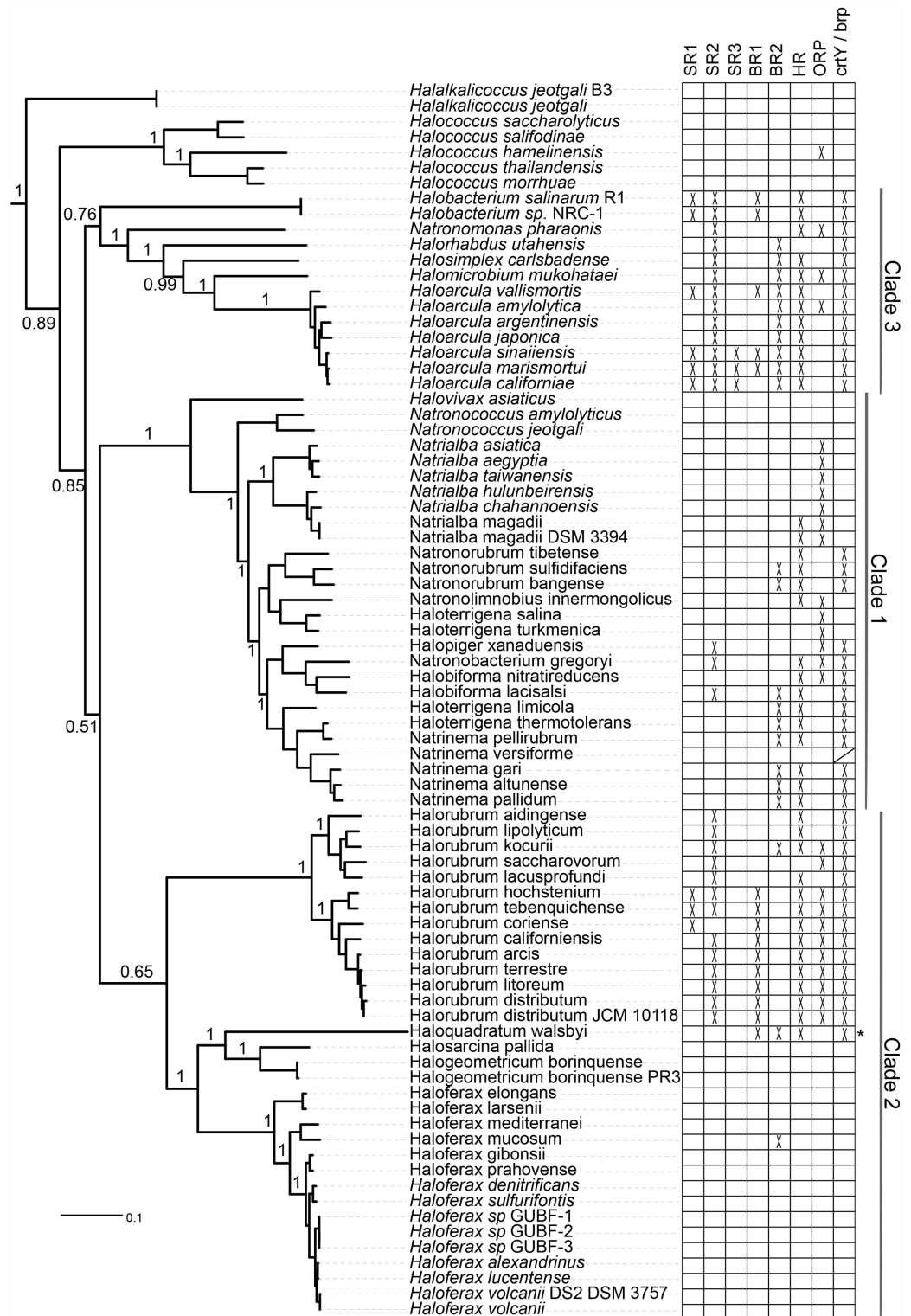
Fig 1. Phylogenetic distribution of opsin classes across the Haloarchaea. Distribution of six previously characterized haloarchaeal opsin families, the putative retinal-free opsins, and the retinal biosynthesis genes *crtY* and *brp* are shown superimposed on a multi-marker phylogenetic tree of 80 sequenced haloarchaea (tree published in Becker *et al*, 2014). SR = sensory rhodopsin, BR = bacteriorhodopsin, HR = halorhodopsin, ORP = retinal-free opsin. Asterisk indicates the presence of middle rhodopsin (MR) [14]. Phylogenetic distribution of *crtY* and *brp* are identical except for one species (*Natrinema versiforme*) which has *brp* but no

detected *crtY* homolog (marked with /). Haloarchaeal clade designations as in Becker *et al*, 2014. Bootstrap support values for lower-level clades removed for clarity. Tree file can be accessed at the Dryad Digital Repository (http://dx.doi.org/10.5061/dryad.963hr).

across 11 genera in all three major haloarchaeal clades [20], in species with and without canonical opsin homologs (S1 Table). All eight species whose genomes encode ORPs and lack canonical opsin homologs were also found to lack *crtY* and *brp*, genes encoding enzymes which catalyze the terminal steps in retinal biosynthesis [21]. Together with additional biochemical and structural modeling, these data suggest the hypothesis that ORP genes encode type 1 opsins that have a non-retinal dependent function, providing a potential functional analogy with eukaryotic GPCRs.

Phylogenetic analysis revealed that the ORPs form a monophyletic clade most closely related to sensory opsins (Fig 2A) and themselves sub-divide into two distinct groups, one consisting of sequences primarily from *Halorubrum* species (group A) and the other of sequences from *Natrialba* species and other Clade 1 haloarchaea (group B) (Fig 2B). In group A ORPs, the Schiff base lysine was replaced with arginine, while in group B ORPs this position contained a leucine or other hydrophobic residue. All but one of the 16 group A ORPs were located adjacent to a predicted methyl-accepting chemotaxis (HAMP/MCP) signal transducer, providing evidence for the hypothesis that the ORPs represent a novel form of sensory opsins, which are often co-operonic with their cognate signal transducers [22] (S1A Fig). Many (17/32) group B ORPs were similarly linked to HAMP/MCP family signal transducers, with nine also located in the proximity of chemotaxis and flagellar biosynthesis operons (S1B Fig). This close genomic association suggests a functional role for at least a sub-set of ORPs in modulating the flagellar apparatus in response to an as yet un-identified signal(s). The remaining 15 group B ORPs, not located adjacent to signal transducer genes, belonged to 10 species, three of which have been described as non-motile [23,24]. We therefore propose that a number of group B ORPs may have signal-response functions unrelated to motility, as is the case for a large number of GPCR homologs [25] and has been suggested for *Anabaena* sensory rhodopsins (ASRs) [26]. Several of the ORPs not linked to signal transducers were located near genes implicated in various stress responses, including heat shock proteins, metal chaperones, and carbon starvation proteins—proposing future lines of research for deciphering the functions of these group B ORPs.

Both haloarchaeal bacteriorhodopsin and bovine visual rhodopsin have been shown to function at reduced efficiencies in the absence of K216/K296, when aminated retinylidene compounds are provided in lieu of retinal [27–29]. To investigate the possibility that the ORPs may function in a canonical light-sensitive manner by retaining interaction with retinal, or a retinal-like chromophore, but not Schiff base formation at K216, we conducted residue conservation analysis and structural modeling. Of three residues experimentally characterized as providing a hydrophobic cavity for the retinal ring in *Natronomonas pharaonis* SRII (V108, F127, W178) [30], only one is conserved as a hydrophobic residue of similar size in group A ORPs (M108) and none in group B ORPs (Fig 3A). W178, which is universally conserved as an aromatic residue in canonical haloarchaeal opsins, has been converted to the much smaller alanine and glycine residues in group A ORPs and group B ORPs, respectively. Similarly, the highly conserved aromatic residue at position 127 has been converted to a polar amino acid (T/S) in most ORPs. In addition, group A ORPs are missing two (W76 → D, Y174 → L) and group B ORPs one (Y174 → L) of the aromatic residues involved in steric constraint of the retinal polyene chain [30] (Fig 3A). These results strongly suggest that, in addition to having lost the Schiff base lysine for covalent binding of retinal, ORPs lack the canonical binding pocket to
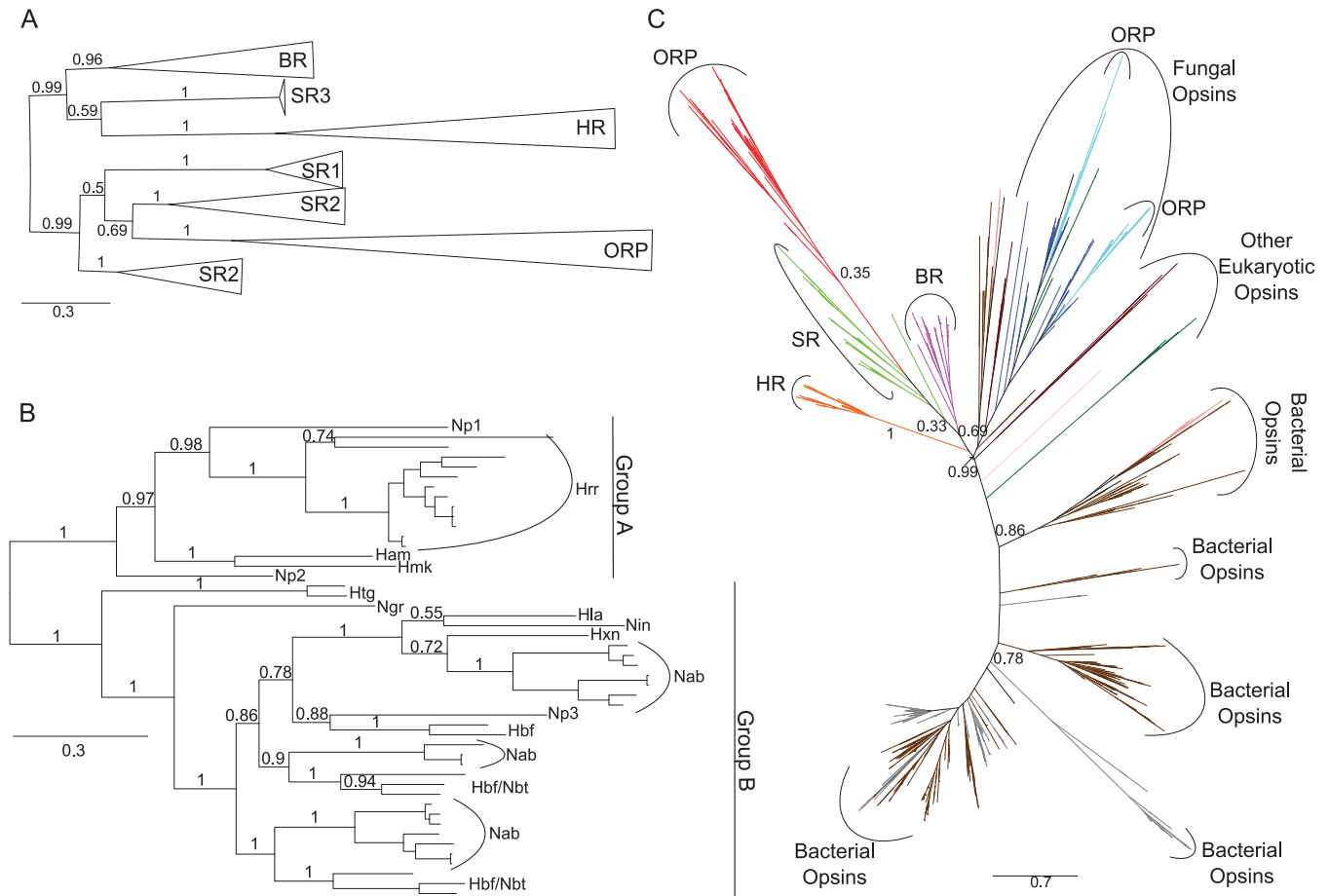
Fig 2. Phylogenies of haloarchaeal and microbial (type 1) opsins. A. Phylogenetic tree of 170 haloarchaeal opsin proteins constructed using Bayesian inference with MrBayes. Individual members of each opsin family were collapsed to indicate relationship among classes. Triangle length is proportional to sequence diversity within clade, triangle width is not significant. Abbreviations as in Fig 1. For fully expanded tree see S5 Fig. Tree file can be accessed at the Dryad Digital Repository (http://dx.doi.org/10.5061/dryad.963hr). **B.** Bayesian inference phylogenetic tree of novel class of putatively retinal-free sensory opsins extracted from tree in Fig 2A. Np = *Natronomonas pharaonis*, Hrr = *Halorubrum*, Ham = *Hararcula amylolytica*, Hmk = *Halomicrobium mukohataei*, Htg = *Haloterrigena*, Ngr = *Natrialba gregoryi*, Hla = *Halobiforma lacisalsi*, Nin = *Natronolimnobius innermongolicus*, Hxn = *Halopiger xanaduensis*, Nab = *Natrialba*, Hbf = *Halobiforma*, Nbt = *Natronobacterium*. **C.** Maximum-likelihood phylogeny of all microbial (type 1) opsins obtained by BLASTp search of NCBI's nr and env_nr databases. Tree inferred using FastTree [71] and ComparetoBootstrap.pl [74] using 500 bootstrap replicates generated with SeqBoot [73]. Branches are colored by phylogenetic affiliation and bootstrap support values above 0.30 are shown for major clades. Colors: grey = unannotated/unassigned, brown = bacterial, salmon = dinoflagellates, dark green = viridiplantae, purple = haloarchaeal bacteriorhodopsin (BR), light green = haloarchaeal sensory rhodopsin (SR), orange = haloarchaeal halorhodopsin (HR), red = haloarchaeal opsin-related protein (ORP), light blue = fungal ORP, dark blue = other fungal opsin. Abbreviations as in Fig 1. For fully expanded tree see S3 Fig. Tree file can be accessed at the Dryad Digital Repository (http://dx.doi.org/10.5061/dryad.963hr).

doi:10.1371/journal.pone.0156543.g002

accommodate retinal-like chromophores. Conversely, residues involved in sensory signaling are conserved in one or both ORP clades. Y51 and R72, which together form a water-mediated hydrogen-bond complex important in propagation of signal to the linked signal transducer [31], are highly conserved in group B, but not group A ORPs. Y199, which forms a hydrogen bond with the signal transducer in *Natronomonas pharaonis* SRII [32], is universally conserved across the ORP clade. Similarly, D189, which in *Nmn. pharaonis* SRII also hydrogen bonds with the cognate signal transducer [32], is highly conserved in both group A and group B ORPs, however, this residue is poorly conserved in other canonical SRII homologs. The high level of conservation of residues involved in signal transduction, combined with lack of
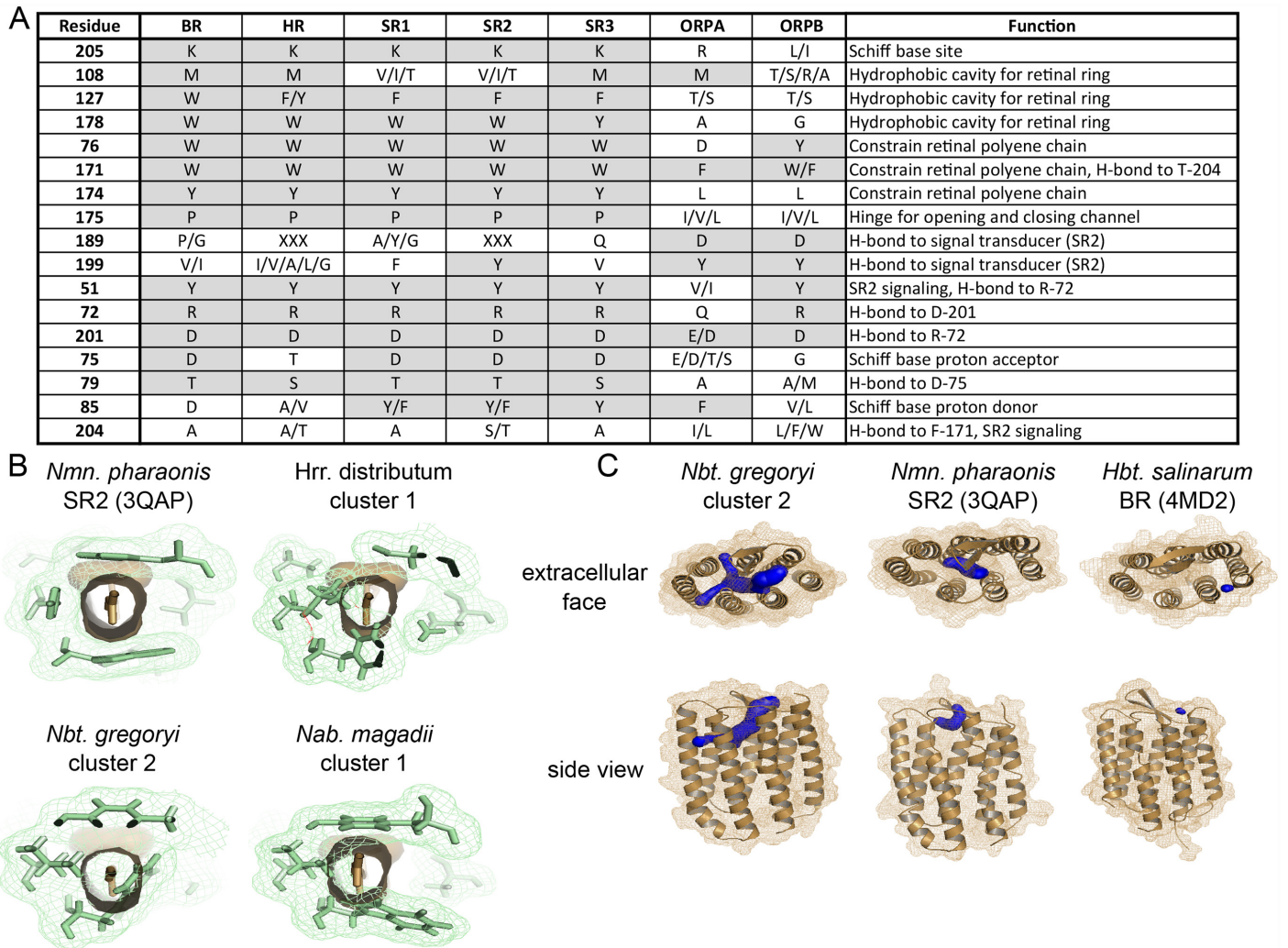
Residue	BR	HR	SR1	SR2	SR3	ORPA	ORPB	Function
205	K	K	K	K	K	R	L/I	Schiff base site
108	M	M	V/I/T	V/I/T	M	M	T/S/R/A	Hydrophobic cavity for retinal ring
127	W	F/Y	F	F	F	T/S	T/S	Hydrophobic cavity for retinal ring
178	W	W	W	W	Y	A	G	Hydrophobic cavity for retinal ring
76	W	W	W	W	W	D	Y	Constrain retinal polyene chain
171	W	W	W	W	W	F	W/F	Constrain retinal polyene chain, H-bond to T-204
174	Y	Y	Y	Y	Y	L	L	Constrain retinal polyene chain
175	P	P	P	P	P	I/V/L	I/V/L	Hinge for opening and closing channel
189	P/G	XXX	A/Y/G	XXX	Q	D	D	H-bond to signal transducer (SR2)
199	V/I	I/V/A/L/G	F	Y	V	Y	Y	H-bond to signal transducer (SR2)
51	Y	Y	Y	Y	Y	V/I	Y	SR2 signaling, H-bond to R-72
72	R	R	R	R	R	Q	R	H-bond to D-201
201	D	D	D	D	D	E/D	D	H-bond to R-72
75	D	T	D	D	D	E/D/T/S	G	Schiff base proton acceptor
79	T	S	T	T	S	A	A/M	H-bond to D-75
85	D	A/V	Y/F	Y/F	Y	F	V/L	Schiff base proton donor
204	A	A/T	A	S/T	A	I/L	L/F/W	H-bond to F-171, SR2 signaling

Fig 3. Residues important for signal propagation, but not retinal binding, are conserved across haloarchaeal ORP homologs. A. Conservation patterns in positions of functional importance across five canonical opsin and two ORP subclasses show that residues involved in signaling, but not those involved in retinal binding, are conserved in ORPs. XXX = highly variable. Positions with conserved or biochemically similar amino acid residue across multiple opsin classes are shaded. Residue numbering is according to *Natronomonas pharaonis* SRII (YP_331142). **B.** Residues forming the retinal binding cavity in canonical opsins show steric clash with canonical retinal binding pocket in predicted ORP structural models. Residues corresponding to *Natronomonas pharaonis* SRII V108, F127, W178, W76, W171, and Y174 are shown in green with lattice representing atomic surface. Retinal is shown in tan, looking down polyene chain, with solid tan representing atomic surface. **C.** Representative output visualizations from CAVER [58] showing predicted novel extracellularly accessible binding pocket and internal tunnel network for *Nbt. gregoryi* WP_005575895. Similar search parameters revealed much smaller (SRII, BR) or no (HR) cavities for canonical opsins. For all CAVER predictions, see S2 Fig.

doi:10.1371/journal.pone.0156543.g003

conservation of the retinal binding pocket, strongly suggest that ORPs are a non-retinal utilizing family of sensory opsins.

Structural models of three ORP homologs, using *Natronomonas pharaonis* SRII (PDB ID: 3QAP) as a template, were consistent with residue analysis in showing lack of a canonical retinal binding pocket. Residues homologous to those lining the canonical binding pocket were shown to overlap the space normally occupied by retinal (Fig 3B). Additionally, a novel extracellularly accessible binding pocket was identified which was missing or highly diminished in canonical haloarchaeal opsins (Fig 3C, S2 Fig). To identify candidate ligands for these novel binding pockets, we used the program iDock [33] to screen 229,358 natural product ligands

against the three structurally modeled ORP homologs. An overlap analysis of the results showed that, despite their structural similarity, the three opsins likely have affinities for very different compounds. Nevertheless, most were sourced from the same ligand classes with similar scaffolds including oxygen and nitrogen heterocycles and sesquiterpene (See SI Discussion). Many examples of extracellularly accessible ligand binding pockets exist for Class A (rhodopsin-like) GPCRs [34–37], further suggesting a non-light related signal response function for the ORPs. Although both structural modeling and residue analysis suggest that the ORPs lack retinal binding capabilities, the possibility that they may utilize an alternate chromophore remains to be tested.

As type 1 opsin homologs lacking the Schiff base lysine have previously been reported for fungi [3], we were interested in determining whether haloarchaeal and fungal ORPs are monophyletic or represent independent losses of retinal-binding ability. We collected type 1 opsin homologs from the NBCI's nr and env_nr databases and performed maximum-likelihood phylogenetic analysis. A total of 1,077 opsin sequences were included. Interrogation of alignments for sequences lacking the Schiff base lysine revealed 45 non-haloarchaeal ORPs, none of which branched within the haloarchaeal ORP clade (Fig 2C). Thus, the haloarchaeal ORPs are evolutionarily distinct from other potentially non-light sensitive microbial opsins. The 45 non-haloarchaeal ORPs included two groups of fungal homologs, one containing 31 sequences from 10 Ascomycota genera and the other eight sequences from seven genera spanning the Ascomycota and Basidomycota. The remaining six sequences were singletons scattered across the phylogeny (S3 Fig). Thus, although loss of the Schiff base lysine, and therefore probable loss of retinal-binding ability, has occurred at least nine times in the evolutionary history of type 1 microbial opsins, the expansion and diversification of both haloarchaeal and fungal ORPs make these exciting targets for learning novel signal transduction strategies used by type 1 opsins.

We also performed a targeted screen for haloarchaeal-type ORPs in the NCBI nr and env_nr databases, as well as the CAMERA metagenomic databases. We detected only one new non-haloarchaeal ORP belonging to the basidiomycete *Trametes versicolor* (EIW60452), which also possesses a fungal-type ORP (EIW51701). The haloarchaeal-type *Trametes versicolor* ORP branched with haloarchaeal ion pumps (BR/HR) rather than the haloarchaeal ORPs and SRs (S4 Fig). We therefore propose that this sequence represents horizontal gene transfer of either BR or HR from the Haloarchaea, followed by loss-of-function, rather than fungal acquisition of a haloarchaeal ORP homolog.

As all previous mentions of K216/K296-lacking opsins in the literature have been based on genomic data [3,9,10], and these sequences have not yet been shown to be expressed, here we verified transcription of several ORP homologs in four haloarchaeal species (*Halorubrum litoreum* JCM 13561, *Halorubrum distributum* JCM 9910, *Natrialba magadii* DSM 3394 and *Natronobacterium gregoryi* SP2). Eight of the nine investigated ORP homologs were shown to be transcribed under standard laboratory conditions (Fig 4A). For one species (*Hrr. distributum*), transcription was also interrogated under conditions of maximal salt tolerance (5.0 M NaCl), reduced dissolved oxygen (shaking at 50 rpm vs 350 rpm), and high cell density (stationary phase). ORP transcript was detected under all conditions. Despite robust transcriptional response, we were unable to detect ORP protein in native hosts by LC-MS/MS, suggesting ORP expression may be very low or under post-transcriptional control (see SI Methods). Additional work will be required to determine biologically relevant expression conditions for ORP proteins. To experimentally verify the proposed inability of ORP proteins to bind retinal, we heterologously expressed His-tagged ORPs from two species (*Hrr. distributum* and *Nab. magadii*) (Fig 4B), and incubated purified ORPs with free all-trans retinal. Neither ORP showed
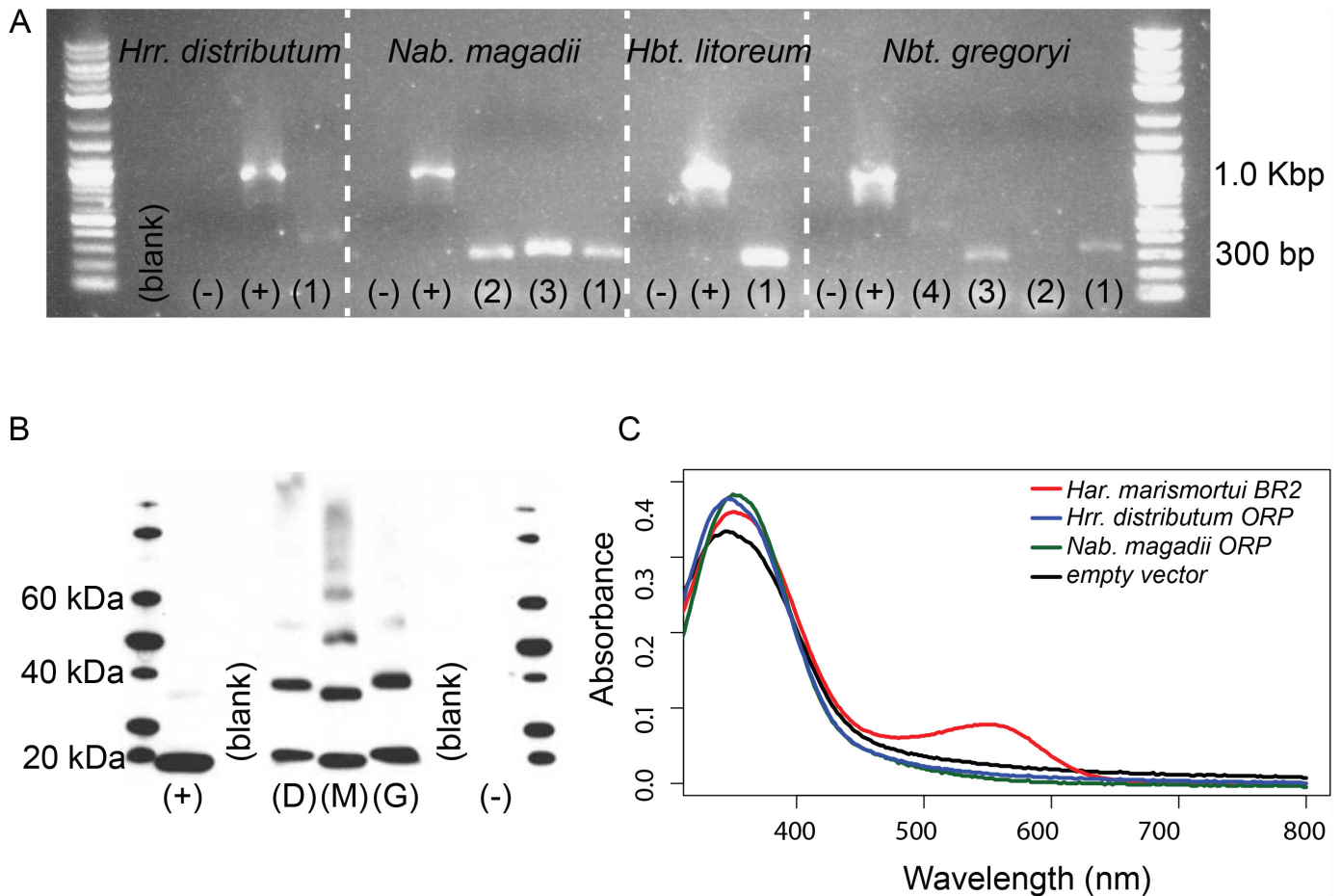
Fig 4. Heterologously expressed ORP homologs do not display retinal-dependent light absorption. A. Transcription of ORPs under standard laboratory conditions was confirmed in four native hosts from both ORP clades. Eight of nine tested ORPs were transcribed, many at low levels. Positive controls are 16S rRNA gene products. *Halorubrum distributum* JCM 9100 (1) = ELZ45759. *Natrialba magadii* DSM 3394 (1) = WP_004215682, (2) = WP_004267173, (3) = WP_004267171. *Halorubrum. litoreum* (1) = WP_008366300. *Natronobacterium gregoryi* (1) = WP_005581268, (2) = WP_005575895, (3) = WP_015233632, (4) = WP_005579638. **B.** Heterologous expression of ORP proteins was confirmed by Western blot. D = *Halorubrum distributum* JCM 9100 ELZ45759, M = *Natrialba magadii* DSM 3394 WP_004267173, G = *Natronobacterium gregoryi* DSM 3393 WP_005575895. Positive control = *Haloarcula. marismortui* BR2 YP_137573. **C.** Visible light spectra of heterologously expressed bacteriorhodopsin and opsin-related proteins cultivated and purified in the presence of excess retinal. The positive control *Haloarcula marismortui* BR2 = YP_137573 shows a canonical absorption peak while the two ORP homologs *Halorubrum distributum* ORP = ELZ45759. *Natrialba magadii* ORP = WP_004267173 shows no peak in the canonical region (480–580 nm) suggesting that retinal has not bound the apoprotein. Protein expression is verified by absorbance at 280 nm and by Western blot (panel B).

doi:10.1371/journal.pone.0156543.g004

absorption in the 480–580 nm range characteristic of canonical opsins [11]. This evidence further suggests that ORP proteins do not bind the retinal chromophore (Fig 4C).

Conclusions

In summary, evidence from genomic context, phylogenetic analysis, structural modeling, and biochemistry provide strong support for the existence of a large family of non-retinal binding opsins derived from a haloarchaeal-specific duplication and divergence of canonical sensory opsins, many of which are possibly involved in chemotaxis and/or transcriptional response to environmental stresses. This new family comprises nearly 30% of all known haloarchaeal opsins, providing a rich set of models through which to explore evolutionary diversification of

signaling protein inputs as well as an enriched understanding of the roles played by microbial opsins in integrating diverse environmental inputs into a coordinated physiological response.

Methods and Materials

Haloarchaeal opsin sequence acquisition, alignment, and phylogenetic inference

During automated annotation of 80 haloarchaeal genomes with the Rapid Annotation Using Subsystem Technology (RAST) server [38], five sequences were annotated as opsin homologs which did not show significant sequence similarity to canonical haloarchaeal opsins by BLASTp search (e-value cutoff of 10^{-5}) and which were missing the Schiff base lysine required for binding of the retinal chromophore in all experimentally characterized opsins (K216 in the haloarchaeal model opsin *Halobacterium sp.* NRC-1 bacteriorhodopsin). These sequences were also annotated as opsins in the NCBI genomic database. A BLASTp search (e-value cutoff of 10^{-5}) against the 80 haloarchaeal genomes using these five opsin sequences as queries, recovered 43 additional K216-lacking, putative opsin homologs for a total of 48 homologs in 28 genomes. These were combined with 122 canonical haloarchaeal opsins recovered by BLASTp search (e-value cutoff of 10^{-5}) of the 80 haloarchaeal genomes using all *Haloarcula marismortui* ATCC 43049 opsins as query sequences. This set of queries was chosen because *Har. marismortui* possesses homologs for each of the six previously categorized classes of haloarchaeal opsins [11,12]. A total of 170 haloarchaeal opsin homologs were included in subsequent analyses (sequences available at the Dryad Digital Repository (http://dx.doi.org/10.5061/dryad.963hr)).

Multiple sequence alignments were created separately for canonical and non-canonical haloarchaeal opsins using MUSCLE 3.8 [39,40], and checked for accuracy against a previously published haloarchaeal opsin alignment [30]. Miscalled start sites for 56 sequences were manually corrected based upon alignment with sequences of experimentally characterized opsins. Individual alignments were combined using the profile-profile alignment option in MUSCLE 3.8 and manually trimmed in the alignment editor Jalview [41]. A total of 204 positions were used to infer phylogeny using the Bayesian tree-building software MrBayes [42,43] with 1.2 million rounds of Markov chain Monte Carlo iteration. The final potential scale reduction factor was 1.000 and final standard deviation of split frequencies was 0.010270. The tree was visualized using FigTree [44]. The tree file and alignment used for tree inference are available from the Dryad Digital Repository (http://dx.doi.org/10.5061/dryad.963hr). For a fully expanded tree, see S5 Fig.

Distribution of opsin classes and retinal biosynthesis genes in the Haloarchaea

The presence/absence pattern of haloarchaeal opsin subclasses and the retinal biosynthesis genes *crtY*, *crtE*, *crtB*, *crtI*, and *brp* were superimposed on a previously published multi-marker phylogeny of the haloarchaea [20] using iTOL [45,46]. Subclass membership for each opsin homolog was determined based on clade affiliation in haloarchaeal opsins tree. Sequences for *crtY*, *crtE*, *crtB*, *crtI*, and *brp* homologs were retrieved via BLASTp searches against the local haloarchaeal database, using query sequences from six haloarchaeal species (sequences available at the Dryad Digital Repository (http://dx.doi.org/10.5061/dryad.963hr)) and an e-value cutoff of 10^{-20}. As all species with *crtY* and *brp* also had a full complement of *crtE*, *crtB*, and *crtI*, presence or absence of *crtY* and *brp* was used to represent retinal biosynthesis ability.

Genome context of ORPs

Genomic context of haloarchaeal ORPs was investigated using JContextExplorer [47]. Haloarchaeal genomes can be loaded using the "Retrieve Popular Genome Set" function. Annotations associated with this genome set were done using the RAST annotation service [38] (see [20]). To view ORP contexts, search by Cluster Number for "4077;15722" (group A) or "2453;13537" (group B). Cluster numbers represent homology families as defined in [20].

Structural modeling of ORP proteins

We used x-ray structure of *Natronomonas pharaonis* SRII (PDB ID: 3QAP) [48] as a template for generating structural models for three ORP proteins (*Nab. magadii* WP_004267173, *Hrr. distributum* ELZ45759, and *Nbt. gregoryi* WP_005575895) using the Rosetta-Membrane method [49–51]. We used *Nmn. pharaonis* SRII because it was identified by the HHpred server [52,53] as the closest structural homolog to the ORPs (21–25% sequence identity). Due to a two residue deletion in the loop between TM6 and TM7 in the *Hrr. distributum* ORP compared with *Nmn. pharaonis* SRII, we predicted the structure of this region *de novo* using Rosetta cyclic coordinate descent (CCD) and kinematic (KIC) closure loop modeling, developed to model loop structures with sub-angstrom accuracy [54,55]. Several rounds of CCD and KIC loop modeling were performmed with at least 10,000 models generated during each round. Models were ranked based on total Rosetta energy after each loop modeling round [49,51,56]. Ten percent of the lowest energy models were clustered [57] using a root mean square deviation (RMSD) threshold that placed 1–2% of all models in at least one of the largest clusters. Models representing centers of the top 20 clusters (early rounds) and/or the best 10 models by total energy (later rounds) were used as input for the next round of loop modeling. Rosetta's full atom relaxation protocol [51,56] was used to explore potential differences in backbone and side chain conformations of ORPs compared with the SRII template. Selection of the best ORP models was guided by clustering of the lowest energy models to generate the most frequently sampled conformations.

Intramolecular pathway and ligand-binding pocket predictions

Comparative structural models described above were used for prediction of ligand binding sites, internal cavities and tunnels using CAVER 3.0 [58]. Analysis was performed individually for each of the top five models generated by Rosetta for each ORP, which had a mean pair-wise RMSD across all models ranging from 0.610 to 0.694 Å. CAVER settings used were minimum probe radius of 0.9 Å, shell depth of 4 Å, shell radius of 3 Å, clustering threshold of 3.5 Å, 12 approximating balls, and maximum distance and desired radius for starting point optimization of 3 Å and 5 Å, respectively. Starting point coordinates were optimized to enable use of homologous starting positions for the majority of models. The starting point for all but two of the 15 ORP models was G193 (*Nab. magadii*) / A185 (*Hrr. distributum*) / G196 (*Nbt. gregoryi*). For two *Hrr. distributum* models, this starting position resulted in no predicted cavities, however, cavities similar to those predicted in other models were discovered with a starting position of Q175-I176. This starting position also resulted in no predicted cavities for SRII, as expected. For comparative visualization purposes, the starting position of P183 was used for SRII, resulting in prediction of a small, likely artifactual ligand binding pocket. For comparison, cavities were also predicted for bacteriorhodopsin (PMID: 4MD2) [59] and halorhodopsin (PMID: 1E12) [60] homologs. Starting points for these predictions were P4 and N3, respectively. Binding pockets and tunnels were visualized with MacPyMOL [61]. For comparison of cavity predictions between ORPs and canonical opsins, see Fig 3C. For comparison of all ORP models, see S2 Fig.

Virtual ligand screening methods

Autodock VINA [62] (1.1.2 May 11, 2011), downloaded from http://vina.scripps.edu/, was used for fast protein-ligand screening. For fast multi-threaded ligand screening the program idock [33] (v2.1.3) was downloaded from https://github.com/HongjianLi/idock. For additional docking experiments we used Autodock (v4.2.6) and Autodock tools [63], the PyRx [64] virtual screening software (v0.8) from http://pyrx.sourceforge.net/ as well as UCSF Chimera [65] for ligand visualization and protein modifications. For SDF file and metadata handling we used the freely available OSIRIS DataWarrior [66] www.openmolecules.org/datawarrior/. The statistical analysis of docking results was performed with StatSoft Statistica (v12) and computational compound clustering of results including 3D mapping was performed with CheS-Mapper [67]. Venn diagrams for compound overlap based on compound ID were calculated on the web page http://www.bioinformatics.lu/venn.php.

Protein data included three canonical opsins (PDB ID: 1E12, 3QAP, and 4MD2) and three ORPs (*Nab. magadii* WP_004267173, *Hrr. distributum* ELZ45759, and *Nbt. gregoryi* WP_005575895). Receptor preparation included removal of water and unwanted ligands as well as pdbqt conversion with AutoDock tools and PyRx software.

A total of 229,358 ligands were sourced from the Universal Natural Products Database (UNPD, pkuxxj.pku.edu.cn/UNPD/) [68]. The 3D-conformers were created using the ChemAxon molconvert software (ChemAxon molconvert; For 3D conformer generation; Molecule File Converter, version 6.0.5, 2013 ChemAxon Ltd. ChemAxon), utilizing the MMFF94 force field optimization with the lowest energy conformer option and explicit hydrogens. The SDF file was then transformed with OpenBabel [69] (v2.3.2) into a ligand pdbqt file and subsequently split into multiple single files with VINA_split.exe. Idock configurations files were created for each receptor with a box size of 15 Å on each axis, including the following docking parameters, (x:16.560 y:40.852 z:-4.072) for *Nab. magadii* WP_004267173, (x:14.244 y:40.502 z:-1.779) for *Hrr. distributum* ELZ45759 and (x:14.527 y:41.251 z:-2.290) for *Nbt. gregoryi* WP_005575895. All 229,358 ligands were screened for each protein. Because the scoring functions of VINA and idock are non-deterministic we repeated the scoring for the best candidates in order to screen for outliers. For the creation of a small decoy library [70], known ligands from the canonical opsins were submitted to http://dude.docking.org/generate. The decoy library with 750 compounds was then used to generate energy cutoff values for the idock-based screening.

The hardware for virtual screening included a Dual Xeon High Performance Workstation with two Intel E5-2687W processors (16 cores, 32 threads), 8 Tb RAID10 disks, 2 Tb SSDs and 196 Gb RAM under Windows 7 64-bit. All modeling was performed on an 80 Gb SoftPerfect RamDisk allowing sequential read/write speeds of up to 4 Gb/second. For additional discussion of ligand screening, see S1 File.

Type 1 opsin sequence acquisition, alignment, and phylogenetic inference

Sequences for bacterial, archaeal, and microbial eukaryotic type 1 opsins were obtained by BLASTp search of the NCBI nr and env_nr databases (as of October 30th, 2012) with an e-value cutoff of 10^{-5} and maximum target sequences set to 100,000 using all *Har. marismortui* opsins as query sequences as described above. A total of 1,430 sequences were recovered. After removing mutants, synthetic constructs, highly fragmentary sequences and sequences lacking a predicted Bac_rhodopsin domain (Pfam clan CL0192), 907 opsin homologs remained. Sequences acquired from each database were independently aligned using MUSCLE 3.8.31 [39,40], manually trimmed to remove poorly aligned regions and to account for the

fragmentary nature of metagenomic sequencing, and alignments combined using the profile-profile alignment option. After adding the 170 haloarchaeal sequences from our dataset, a total of 147 positions for 1,077 opsin sequences were used in tree inference. For these sequences, an initial guide tree was constructed using FastTree [71,72], then 1000 re-sampled alignments were generated using the Phylip package SEQBOOT [73] and used to determine bootstrap support values for clades in the initial tree with CompareToBootstrap.pl [74]. The resulting tree was visualized with FigTree [44]. The tree file is available from the Dryad Digital Repository (http://dx.doi.org/10.5061/dryad.963hr). For fully expanded tree, see S3 Fig. (red leaf labels represent sequences lacking the Schiff base lysine).

Phylogenetic distribution of haloarchaeal-type ORPs

All protein and peptide databases in CAMERA and the NCBI nr and env_nr databases (as of December 18th, 2012) were interrogated for non-canonical opsin homologs using BLASTp searches with all 48 haloarchaeal ORPs as queries. BLASTp parameters and database details are listed in S2 Table. Redundant sequences were removed using Jalview [41] and unique sequences searched against the Pfam database [75] for presence of a Bac_rhodopsin domain (CL0192). Sequences with a Bac_rhodopsin domain were aligned using MUSCLE 3.8 [39,40] and presence or absence of K216 was recorded. To assess the evolutionary origin of the only non-haloarchaeal ORP recovered from this search (*Trametes versicolor* EIW60452), a phylogeny was inferred for all 170 haloarchaeal opsins and the *T. versicolor* ORP. Sequences were aligned using Muscle 3.8.31 [39,40] and alignment manually trimmed in Jalview [41]. After trimming, 220 positions were used in phylogenetic inference using FastTree [71,72] and the resulting tree was visualized in FigTree [44]. For tree see S4 Fig.

Confirmation of native transcription of ORPs

Liquid cultures of *Hrr. litoreum* JCM 13561, *Hrr. distributum* JCM 9910, *Nab. magadii* DSM 3394 and *Nbt. gregoryi* SP2 were grown to mid-log phase in JCM 168 (*Hrr. litoreum* and *distributum*) or DSM 371 (*Nab. magadii* and *Nbt. gregoryi*) media. Cell pellets were collected from 2 mL of mid-log phase cultures and RNA harvested using 1.0 mL TRIzol® Reagent with standard extraction protocol followed by a DNase digestion step and a second TRIzol® extraction with 0.5 mL Trizol® Reagent. DNase digestion was with NEB DNase I (M0303S) using standard protocol. Following second Trizol® extraction, RNA was tested for gDNA contamination using PCR with haloarchaeal 16S primers (S3 Table) prior to reverse-transcription. Reverse transcription was carried out with SuperScript® III Reverse Transcriptase from Life Technologies™ using standard protocols, and random hexamers (Qiagen, 79236). Presence of transcript was confirmed via PCR with primers listed in S3 Table. Cross-reactivity of primers for species with multiple ORP homologs was tested and primer sets found to be gene-specific. For *Hrr. distributum*, additional cultures were grown under the following conditions, and transcript presence verified using methods described above: a) 350 rpm, 3.42 M NaCl, b) 350 rpm, 5.0 M NaCl, c) 50 rpm, 3.42 M NaCl. Conditions (a) and (b) were incubated using a G-53 gyratory tier shaker (New Brunswick Scientific, M1074), condition (c) in an Innova 44R incubator (New Brunswick Scientific, M1282). For each condition, samples were collected at mid-log and stationary phase.

Growth of *Halobacterium salinarum* clones

Unless otherwise specified, *Halobacterium salinarum* NRC-1 and derivatives were grown in CM medium (250 g/L NaCl, 20 g/L MgSO$_4$ • 7H$_2$O, 2 g/L KCl, 3 g/L Na$_3$Citrate, 10 g/L Oxoid

Neutralized Peptone (Oxoid, LP0037)) at 37°C in an Innova 44R incubator (New Brunswick Scientific, M1282) shaken at 175 rpm.

Cloning

In order to heterologously express His-tagged ORP proteins, genes encoding *Nab. magadii* WP_004267173 and *Hrr. distributum* ELZ45759 were PCR amplified from genomic DNA using primers listed in S3 Table in 50 μL PCR reactions: 5 μL 10X buffer, 10 μL Q solution, 2.5 μL 10 μM forward primer, 2.5 μL 10 μM reverse primer, 1.25 μL 10 mM dNTPs, 0.3 μL TAQ polymerase (Qiagen, Q201203), 0.1 μL Pfu Polymerase (Stratagene, 600153), 28.35 μL MilliQ H_2O using the following PCR cycle: 98°C 10 min, (98°C 30 sec, 55°C 1 min, 72°C 1 min 35 sec) x 25 cycles, 72°C 5 min in a Dyad Peltier Thermocycler (BioRad). Additionally, genes encoding a canonical BR (*Har. marismortui* YP_137573) and a colorless transmembrane protein (*Hbt. sp*. NRC-1 *pstC2* VNG0455G), were amplified as controls from genomic DNA as described above. Amplified genes were then ligated into two expression vectors, one under the control of the NRC-1 *bop* promoter (pDJLCHIS) and the other under the control of the NRC-1 ferredoxin promoter (pMTFCHIS2) using the restriction enzymes NdeI (NEB, R0111S) and BamHI (NEB, R0136S for pMTFCHIS2) or HindIII (NEB, R0104S for pDJLCHIS). Plasmid sequences available from the Dryad Digital Repository (http://dx.doi.org/10.5061/dryad.963hr). Clones were transformed into chemically competent DH5α cells and plated on LB agar plates containing 100 μg/mL carbenicillin (Fisher Scientific, BP26485). Success of cloning was verified through Sanger sequencing using plasmid specific sequencing primers (S3 Table). Constructed pDJLCHIS and pMTFCHIS2 expression vectors (S6 Fig) were transformed into *Hbt. sp*. NRC-1 SD23 and *Hbt. sp*. NRC-1 SD20 strains, respectively.

Similarly, for heterologous expression in *Escherichia coli*, genes encoding *Nab. magadii* WP_004267173, *Hrr. distributum* ELZ45759, and *Har. marismortui* BR2 YP_137573 were PCR amplified from genomic DNA and cloned into pET29b+ (Novagen) using *NdeI* and *HindIII* restriction sites as described above. All cloned constructs, including an empty pET29b+ vector, were transformed into *E. coli* BLR(DE3) cells (Merck Millipore) for expression.

Protein-level expression of ORPs in heterologous hosts

To express ORP proteins in *E. coli*, all clones were grown in Luria Broth (LB) under 50 μg/mL kanamycin resistance (kan) in an Innova 44R Shaking Incubator (New Brunswick) with shaking at 175 rpm. The pET29b+/ORP, pET29b+/BR2 and pET29b+/empty BLR(DE3) clones were revived from freezer stock overnight in 4 mL of LB + kan at 37°C. Revived cells were re-cultured in 25 mL of LB + kan with a starting OD_{600} of 0.1 and grown at 37°C to mid-log (OD_{600} = 0.4). The mid-log subcultures were then used to inoculate 1 L LB + kan with a starting OD_{600} of 0.01 and placed in the 37°C shaking incubator. Once the cultures returned to mid-log (OD_{600} = 0.4), all cultures were induced with 0.5 mM IPTG (Fisher Scientific, BP1755), supplemented with 10 μM all-trans retinal (Sigma-Aldrich, R2500), and incubated at 18°C for 18 hours with shaking. Cells were harvested by centrifuging at 8000 rpm for 10 min and stored at -80°C until use.

To purify, cell pellets were thawed and re-suspended in 15 mL of solubilization buffer (100 mM Na/K phosphate pH 7.4, 2% Triton-X100, 10 μM all-trans retinal) with one cOmplete® Mini EDTA free protease inhibitor cocktail tablet (Roche, 04693159001). Samples were sonicated three times with a Model 120 Sonic Dismembrator (Fisher Scientific, FB120) fitted with a Model CL-18 probe at 65% power alternatively for 2 seconds ON and 2 seconds OFF; for a total of 2 minutes ON. Sonicated samples were incubated on a rotisserie for 3 hours at 4°C. Cell debris was spun down by centrifugation at 8000 rpm for 10 minutes, and the supernatant was

mixed with an equal volume of equilibration/wash buffer (50 mM Na$^+$ phosphate, 300 mM NaCl, 10 mM imidazole; pH 7.4, 0.5% Triton-X100, 10 μM all-trans retinal). 250 μL of HisPur™ Cobalt Resin (Thermo Scientific, 89964) was washed with 500 μL of equilibration/wash buffer and mixed with equilibrated lysate for 45 minutes in a rotisserie at 4°C. Conjugated resin was washed three times with 500 μL equilibration/wash buffer and eluted twice with 250 μL of elution buffer (50 mM Na$^+$ phosphate, 300 mM NaCl, 150 mM imidazole; pH 7.4, 0.5% Triton-X100, 10 μM all-trans retinal). 250 μL of the first eluent was concentrated to ~100 μL using a Microcon 30 kDa MWCO centrifugal filter column (Millipore, 42410) and aliquoted into a Greiner Half Area UV-Star® microplate (Greiner Bio-One, 675801). Absorbance was measured from 250–800 nm in 2 nm intervals with an Infinite M200 plate reader (Tecan, 30016056). Background (elution buffer) was subtracted from all spectra and spectra were normalized to the 280 nm absorbance value of the *Har. marismortui* BR2 positive control. For spectra see Fig 4C.

To verify expression of ORP protein, we performed a Western blot on HisPur™ purified lysate from cloned ORPs in *E. coli* BLR(DE3) background using the following protocol: SDS-PAGE was run as described in SI Methods and the gel was soaked in 15 mL of transfer buffer (25 mM Tris base, 192 mM glycine, 10% methanol, pH 8.4) for 15 min. A 0.2 μm pore diameter nitrocellulose membrane (Biorad, 162–0112), foam pads, and 3 mm Whatman filters were soaked in transfer buffer before assembling the sandwich for transfer. Proteins were transferred from the PAGE gel to the nitrocellulose membrane at 30V for 1 hr on ice. His-tagged proteins were probed using the SuperSignal® West HisProbe™ Kit (Thermo Scientific™, 15168) and signal was detected using Amersham ECL Prime Western Blotting Detection Reagent (GE Healthcare Life Sciences, RPN2232). For Western blot, see Fig 4B.

Protein-level expression of ORPs in native hosts

Several methods were used to investigate protein-level expression of ORP homologs in both native and heterologous hosts. First, we attempted identification of ORPs by LC-MS/MS in the native host *Nbt. gregoryi* SP2, which has four unique ORP genes. Cell pellets collected from mid-log phase cultures were lysed with 200 μL lysis buffer containing 100 mg SDS, 10 mL diH$_2$O, 60 μL DNase I (GoldBio D-300-1), 0.25 mg RNase A (Roche, 10109169001), and 1 cOmplete® Mini EDTA free protease inhibitor cocktail tablet (Roche, 04693159001) per 15 mL buffer. Lysate was sonicated for 20 minutes (30 s on/off) on a Bioruptor® UCD-200 (Diagenode). Cellular debris was removed by centrifugation, and supernatant denatured for 10 min at 100°C with 6x SDS buffer. Approximately 40 μg protein was loaded into pre-poured 4–20% SDS-PAGE gel (Bionexus Inc., 2BNPC420) and run approximately 1 cm into gel. Total protein band was excised and subjected to in-gel trypsin digest according to the following protocol: gel was cut into 1 mm^3 pieces, washed with 50 mM Ammonium Bicarbonate (AmBic), shrunk with acetonitrile (ACN), reduced with 10 mM DTT/50 mM AmBic, shrunk again with ACN, incubated in 55 mM iodoacetamide/50mM AmBic 20 min in the dark, washed with 50 mM AmBic, shrunk with ACN and partially dried in a vacuum concentrator (Labconco). Overnight digestion was carried out at 37°C with 250 ng of trypsin (Promega, V5117) in 50 mM AmBic (pH 8). The supernatant was sonicated in 60% ACN and 0.1% trifluoroacetic acid for 10 min, then dried in the vacuum concentrator. Digested peptides were analyzed by LC-MS/MS on a Thermo Q-Exactive mass spectrometer with Michrom Paradigm LC and CTC Pal autosampler. Peptides were directly loaded onto an Agilent ZORBAX 300SB C$_{18}$ reversed phase trap cartridge, which, after loading, was switched in-line with a Michrom Magic C$_{18}$ AQ 200 um x 150 mm C$_{18}$ column connected to a Thermo-Finnigan LTQ iontrap mass spectrometer through a Michrom Advance Plug and Play nano-spray source. The nano-LC column (Michrom 3μ 200Å MAGIC C18AQ 200μ x 150 mm) was used with a 90 min-long gradient (1–10% buffer B

in 5 min, 10–35% buffer B in 65 min, 35–70% buffer B in 5 min, 70% buffer in 1 min, 1% buffer B in 14 min) at a flow rate of 2 uL min^{-1} for the maximum separation of tryptic peptides. A top 15 method was used with Xcalibur software to collect Q-Exactive data, with a scan range of 300–1600 m/z. Results were searched against a *Nbt. gregoryi* database with cRAP proteins and reversed sequences (5402 proteins total) in X!Tandem [76] with a fragment ion mass tolerance of 20 PPM and a parent ion tolerance of 20 PPM. Carbamidomethyl of cysteine was specified in X!Tandem as a fixed modification. Glu->pyro-Glu of the N-terminus, ammonia-loss of the N-terminus, gln->pyro-Glu of the N-terminus, deamidation of asparagine and glutamine, oxidation of methionine and tryptophan, dioxidation of methionine and tryptophan and acetylation of the N-terminus were specified in X!Tandem as variable modifications. Scaffold v.4.0.7 [77] was used to validate peptide and protein identifications. Peptide identifications were accepted if they could be identified with confidence of greater than or equal to 95% and protein identifications were accepted if they could be identified with confidence of greater than or equal to 95% and contained at least two identified peptides. Protein probabilities were assigned by the Protein Prophet algorithm [78]. Proteins that contained similar peptides and could not be differentiated were grouped. Proteins sharing significant peptide evidence were grouped into clusters. No ORP proteins were identified from *Nbt. gregoryi* SP2 lysate.

After failing to detect ORP protein in a native host, we purified His-tagged cloned proteins from a *Hbt. sp.* NRC-1 SD23 *bop*(-) background using HisPur$^{®}$ Cobalt resin (Thermo Scientific™, 89965). Exudate was loaded onto a 4–20% SDS-PAGE gel as described above, and size separated. Based upon presence of band in the three experimental samples and absence in *Hbt. sp.* NRC-1 SD23 control lacking expression vector, the region of the gel corresponding to 50 kDa MW proteins was excised and prepared for proteomics as described above. No ORP proteins were detected.

We next performed Western blot on lysate from cloned ORPs in *Hbt. sp.* NRC-1 SD23 *bop* (-) background using the following protocol. SDS-PAGE was run as described above and gel soaked in 15 mL of transfer buffer (25 mM Tris base, 192 mM glycine, 10% methanol, pH 8.4) for 15 min. PVDF membrane (Novex, LC2002) was successively soaked in 100% methanol, MilliQ™ H$_2$O, and transfer buffer. Foam pads and 3mm Whatman filters were also soaked in transfer buffer. Western gel was run at 30V for 1 hr. His-tagged proteins were probed using the SuperSignal$^{®}$ West HisProbe™ Kit (Thermo Scientific™, 15168). Based upon presence of two bands at ~15 and ~20 kDa in the three experimental samples, the region of the gel corresponding to 13–25 kDa was extracted and prepared for proteomics as described above. No ORPs were detected, however, the three opsin proteins present in the host (HR, SRI, and SRII) were observed, indicating the region of the gel used was the correct MW range for opsins. Over 900 proteins were identified with a protein false discovery rate of 1.7% and peptide false discovery rate of 0.36%, using a *Hbt. sp.* NRC-1 database with ORPs added.

Supporting Information

S1 Fig. Genomic context of putative retinal-free opsins. Representative genomic contexts of Group A (A) and Group B (B) ORPs. Genome context was visualized in JContextExplorer [47]. HP = hypothetical protein/unknown, Aminotransferase = serine-pyruvate aminotransferase/ archaeal aspartate aminotransferase, ST = signal transducer, ORP = opsin-related protein, ABC = ABC transporter ATP-binding protein, ATH = acyl-CoA thioester hydrolase, UDP-diP = undecaprenyl-diphosphatase. B1/2/4 = Flagellin B1/B2/B4 precursor, FOP = conserved fla operon protein, MACP = methyl-accepting chemotaxis protein, W = CheW, G = FlaG, CP = conserved che operon protein/chemotaxis protein, Soj = sporulation initiation inhibitor protein. (PDF)

S2 Fig. Novel extracellularly accessible binding pocket predicted for ORPs. Output visualizations from CAVER [58] showing predicted novel extracellularly accessible binding pocket and internal tunnel network for top five structural models predicted for *Hrr. distributum* ELZ45759, *Nab. magadii* WP_004267173, *Nbt. gregoryi* WP_005575895.
(PDF)

S3 Fig. Phylogeny of Microbial type 1 opsins (expanded). Unrooted maximum-likelihood phylogeny of all microbial (type 1) opsins obtained by BLASTp searches of NCBI's nr and env_nr databases using canonical haloarchaeal opsins as queries. Tree inferred using FastTree [71] and ComparetoBootstrap.pl [74] using 500 bootstrap replicates generated with SeqBoot [73]. Sequences lacking the Schiff base lysine (K216) are colored red. Tree file can be accessed at the Dryad Digital Repository (http://dx.doi.org/10.5061/dryad.963hr).
(PDF)

S4 Fig. Phylogeny placing *Trametes versicolor* haloarchaeal-type ORP. Unrooted maximum-likelihood phylogeny of sequences used for Fig 2C plus the single recovered haloarchaeal-type ORP from outside of the Haloarchaea. Tree inferred using FastTree [71].
(PDF)

S5 Fig. Phylogeny of haloarchaeal opsins (expanded). Phylogenetic tree of 170 haloarchaeal opsin proteins constructed using Bayesian inference with MrBayes [42]. Abbreviations as in Fig 1. Tree file can be accessed at the Dryad Digital Repository (http://dx.doi.org/10.5061/dryad.963hr).
(PDF)

S6 Fig. Expression vectors. Plasmid maps of heterologous expression vectors pDJLCHIS and pMTFCHIS2. Vectors differ only in promoter driving expression of inserted gene. Pfdx = ferredoxin promoter, Pbop = bop promoter. Plasmid sequences available at the Dryad Digital Repository (http://dx.doi.org/10.5061/dryad.963hr).
(PDF)

S7 Fig. Virtual ligand screening. A Histograms for solvent accessible surface area ranges [Å]2 for compounds with binding energies less than -9 kcal/mol. **B** idock binding energies for 229,358 ligands from natural product space. Compounds with binding energies less than -9 kcal/mol were selected for further analysis. **C** Representative compounds identified during virtual compound screening. These compounds include naphthoquinones, nitrogen-containing heterocycles and a number of sesquiterpenes. The name or compound class is given, as well as the docking energy for each single compound with one of the three modeled ORPs.
(PDF)

S1 File. Supplementary discussion of Virtual Ligand Screening. Additional discussion regarding the results of the virtual ligand screening performed herein are described.
(PDF)

S1 Table. Haloarchaeal species with ORPs. Strain names and number of ORP homologs detected for each haloarchaeal species possessing putative retinal-free opsins.
(PDF)

S2 Table. Databases searched for haloarchaeal-type ORPs. Search statistics and results for search for haloarchaeal-type ORPs. For details on filtering parameters, see Methods and Materials. Briefly, databases were searched using all haloarchaeal ORPs as query, and results were filtered for unique hits containing a Bac_rhodopsin domain (Pfam clan CL0192). Results

matching these criteria were aligned and checked for presence or absence of the Schiff base lysine (K216).
(PDF)

S3 Table. Primer sequences. PCR and cloning primers for ORPs from four species used in experimental confirmation of ORP expression. Expected product lengths and primer set used to screen for gDNA contamination also provided.
(PDF)

Acknowledgments

Author Contributions

Conceived and designed the experiments: EAB MTF. Performed the experiments: EAB AIY PMS TK TW RE VY. Analyzed the data: EAB AIY MTF PMS TK KSYS VY. Contributed reagents/materials/analysis tools: MTF RE TK VY. Wrote the paper: EAB MTF VY TK AIY RE PMS.

References

1. Koyanagi M, Terakita A. Diversity of animal opsin-based pigments and their optogenetic potential. Biochim Biophys Acta. Elsevier B.V.; 2013; 1837: 710–716. doi: 10.1016/j.bbabio.2013.09.003

2. Inoue K, Tsukamoto T, Sudo Y. Molecular and evolutionary aspects of microbial sensory rhodopsins. Biochim Biophys Acta. Elsevier B.V.; 2014; 1837: 562–577. doi: 10.1016/j.bbabio.2013.05.005

3. Spudich JL, Yang CS, Jung KH, Spudich EN. Retinylidene proteins: structures and functions from archaea to humans. Annu Rev Cell Dev Biol. 2000; 16: 365–92. doi: 10.1146/annurev.cellbio.16.1.365 PMID: 11031241

4. Larusso ND, Ruttenberg BE, Singh AK, Oakley TH. Type II opsins: evolutionary origin by internal domain duplication? J Mol Evol. 2008; 66: 417–23. doi: 10.1007/s00239-008-9076-6 PMID: 18392762

5. Mackin KA, Roy RA, Theobald DL. An empirical test of convergent evolution in rhodopsins. Mol Biol Evol. 2014; 31: 85–95. doi: 10.1093/molbev/mst171 PMID: 24077848

6. Shen L, Chen C, Zheng H, Jin L. The evolutionary relationship between microbial rhodopsins and metazoan rhodopsins. ScientificWorldJournal. 2013; 2013: 435651. doi: 10.1155/2013/435651 PMID: 23476135

7. Nathans J, Hogness DS. Isolation, sequence analysis, and intron-exon arrangement of the gene encoding bovine rhodopsin. Cell. 1983; 34: 807–14. PMID: 6194890

8. Mullen E, Johnson AH, Akhtar M. The Identification of Lys-216 as the retinal binding residue in bacteriorhodopsin. FEBS Lett. 1981; 130: 187–193. PMID: 6793396

9. Siddaramappa S, Challacombe JF, Decastro RE, Pfeiffer F, Sastre DE, Giménez MI, et al. A comparative genomics perspective on the genetic content of the alkaliphilic haloarchaeon Natrialba magadii ATCC 43099T. BMC Genomics. 2012; 13: 165. doi: 10.1186/1471-2164-13-165 PMID: 22559199

10. Feuda R, Hamilton SC, McInerney JO, Pisani D. Metazoan opsin evolution reveals a simple route to animal vision. Proc Natl Acad Sci U S A. 2012; 109: 18868–18872. doi: 10.1073/pnas.1204609109 PMID: 23112152

11. Fu H-Y, Lin Y-C, Chang Y-N, Tseng H, Huang C-C, Liu K-C, et al. A novel six-rhodopsin system in a single archaeon. J Bacteriol. 2010; 192: 5866–73. doi: 10.1128/JB.00642-10 PMID: 20802037

12. Baliga NS, Bonneau R, Facciotti MT, Pan M, Glusman G, Deutsch EW, et al. Genome sequence of Haloarcula marismortui: a halophilic archaeon from the Dead Sea. Genome Res. 2004; 14: 2221–34. doi: 10.1101/gr.2700304 PMID: 15520287

13. Bolhuis H, Palm P, Wende A, Falb M, Rampp M, Rodriguez-Valera F, et al. The genome of the square archaeon Haloquadratum walsbyi : life at the limits of water activity. BMC Genomics. 2006; 7: 169. doi: 10.1186/1471-2164-7-169 PMID: 16820047

14. Sudo Y, Ihara K, Kobayashi S, Suzuki D, Irieda H, Kikukawa T, et al. A microbial rhodopsin with a unique retinal composition shows both sensory rhodopsin II and bacteriorhodopsin-like properties. J Biol Chem. 2011; 286: 5967–76. doi: 10.1074/jbc.M110.190058 PMID: 21135094

15. Landau EM, Pebay-Peyroula E, Neutze R. Structural and mechanistic insight from high resolution structures of archaeal rhodopsins. FEBS Lett. 2003; 555: 51–56. doi: 10.1016/S0014-5793(03)01082-2 PMID: 14630318

16. Wang T, Oppawsky C, Duan Y, Tittor J, Oesterhelt D, Facciotti MT. Stable Closure of the Cytoplasmic Half-Channel Is Required for. 2014;

17. Grote M, Engelhard M, Hegemann P. Of ion pumps, sensors and channels—Perspectives on microbial rhodopsins between science and history. Biochim Biophys Acta. Elsevier B.V.; 2014; 1837: 533–545. doi: 10.1016/j.bbabio.2013.08.006

18. Tastan O, Dutta A, Booth P, Klein-Seetharaman J. Retinal proteins as model systems for membrane protein folding. Biochim Biophys Acta. Elsevier B.V.; 2014; 1837: 656–663. doi: 10.1016/j.bbabio.2013.11.021

19. Zhang F, Vierock J, Yizhar O, Fenno LE, Tsunoda S, Kianianmomeni A, et al. The microbial opsin family of optogenetic tools. Cell. 2011; 147: 1446–57. doi: 10.1016/j.cell.2011.12.004 PMID: 22196724

20. Becker EA, Seitzer PM, Tritt A, Larsen D, Krusor M, Yao AI, et al. Phylogenetically driven sequencing of extremely halophilic archaea reveals strategies for static and dynamic osmo-response. PLoS Genet. 2014; 10: e1004784. doi: 10.1371/journal.pgen.1004784 PMID: 25393412

21. McCarren J, DeLong EF. Proteorhodopsin photosystem gene clusters exhibit co-evolutionary trends and shared ancestry among diverse marine microbial phyla. Environ Microbiol. 2007; 9: 846–58. doi: 10.1111/j.1462-2920.2006.01203.x PMID: 17359257

22. Spudich JL. The multitalented microbial sensory rhodopsins. Trends Microbiol. 2006; 14: 480–7. doi: 10.1016/j.tim.2006.09.005 PMID: 17005405

23. Xu Y, Zhou P, Tian X. Characterization of two novel haloalkaliphilic archaea Natronorubrum bangense gen. nov., sp. nov. and Natronorubrum tibetense gen. nov., sp. nov. Int J Syst Bacteriol. 1999; 49: 261–266. PMID: 10028271

24. Xu Y, Wang Z, Xue Y, Zhou P, Ma Y, Ventosa A, et al. Natrialba hulunbeirensis sp. nov. and Natrialba chahannaoensis sp. nov., novel haloalkaliphilic archaea from soda lakes in Inner Mongolia Autonomous Region, China. Int J Syst Evol Microbiol. 2001; 51: 1693–1698. PMID: 11594597

25. Schiöth HB, Fredriksson R. The GRAFS classification system of G-protein coupled receptors in comparative perspective. Gen Comp Endocrinol. 2005; 142: 94–101. doi: 10.1016/j.ygcen.2004.12.018 PMID: 15862553

26. Brown LS. Eubacterial rhodopsins—Unique photosensors and diverse ion pumps. Biochim Biophys Acta. Elsevier B.V.; 2014; 1837: 553–561. doi: 10.1016/j.bbabio.2013.05.006

27. Zhukovsky EA, Robinson PR, Oprian DD. Transducin activation by rhodopsin without a covalent bond to the 11-Cis-Retinal chromophore. Science (80-). 1991; 251: 558–560.

28. Schweiger U, Tittor J, Oesterhelt D, Manuscript R, November R. Bacteriorhodopsin Can Function without a Covalent Linkage between Retinal and Protein. 1994; 535–541.

29. Friedman N, Druckmann S, Lanyi J, Needleman R, Lewis A, Ottolenghi M, et al. A covalent link between the chromophore and the protein backbone of bacteriorhodopsin is not required for forming a photochemically active pigment analogous to the wild type. Biochemistry. 1994; 33: 6–11.

30. Pebay-Peyroula E, Royant A, Landau EM, Navarro J. Structural basis for sensory rhodopsin function. Biochim Biophys Acta. 2002; 1565: 196–205. PMID: 12409195

31. Ishchenko A, Round E, Borshchevskiy V, Grudinin S, Gushchin I, Klare JP, et al. Ground state structure of D75N mutant of sensory rhodopsin II in complex with its cognate transducer. J Photochem Photobiol B. Elsevier B.V.; 2013; 123: 55–8. doi: 10.1016/j.jphotobiol.2013.03.008

32. Gordeliy VI, Labahn J, Moukhametzianov R, Efremov R, Granzin J, Schlesinger R, et al. Molecular basis of transmembrane signalling by sensory rhodopsin II-transducer complex. Nature. 2002; 419: 484–7. doi: 10.1038/nature01109 PMID: 12368857

33. Li H, Leung K-S, Wong M-H. idock: A multithreaded virtual screening tool for flexible ligand docking. Comput Intell Bioinforma Comput Biol (CIBCB), 2012 IEEE Symp. 2012; 77–84.

34. Tan Q, Zhu Y, Li J, Chen Z, Han GW, Kufareva I, et al. Structure of the CCR5 chemokine receptor-HIV entry inhibitor maraviroc complex. Science. 2013; 341: 1387–90. doi: 10.1126/science.1241475 PMID: 24030490

35. Haga K, Kruse AC, Asada H, Yurugi-Kobayashi T, Shiroishi M, Zhang C, et al. Structure of the human M2 muscarinic acetylcholine receptor bound to an antagonist. Nature. Nature Publishing Group; 2012; 482: 547–51. doi: 10.1038/nature10753

36. Rasmussen SGF, Choi H-J, Fung JJ, Pardon E, Casarosa P, Chae PS, et al. Structure of a nanobody-stabilized active state of the β(2) adrenoceptor. Nature. Nature Publishing Group; 2011; 469: 175–80. doi: 10.1038/nature09648

37. Lebon G, Warne T, Edwards PC, Bennett K, Langmead CJ, Leslie AGW, et al. Agonist-bound adenosine A2A receptor structures reveal common features of GPCR activation. Nature. Nature Publishing Group; 2011; 474: 521–5. doi: 10.1038/nature10136

38. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, et al. The RAST Server: rapid annotations using subsystems technology. BMC Genomics. 2008; 9: 75. doi: 10.1186/1471-2164-9-75 PMID: 18261238

39. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 2004; 5: 113. doi: 10.1186/1471-2105-5-113 PMID: 15318951

40. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004; 32: 1792–7. doi: 10.1093/nar/gkh340 PMID: 15034147

41. Clamp M, Cuff J, Searle SM, Barton GJ. The Jalview Java alignment editor. Bioinformatics. 2004; 20: 426–7. doi: 10.1093/bioinformatics/btg430 PMID: 14960472

42. Huelsenbeck JP, Ronquist F. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics. 2001; 17: 754–5. PMID: 11524383

43. Ronquist F, Huelsenbeck JP. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 2003; 19: 1572–1574. doi: 10.1093/bioinformatics/btg180 PMID: 12912839

44. Rambaut A. FigTree v12.05 [Internet]. 2012 p. http://tree.bio.ed.ac.uk/software/figtree/.

45. Letunic I, Bork P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. Bioinformatics. 2007; 23: 127–8. doi: 10.1093/bioinformatics/btl529 PMID: 17050570

46. Letunic I, Bork P. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. Nucleic Acids Res. 2011; 39: W475–8. doi: 10.1093/nar/gkr201 PMID: 21470960

47. Seitzer P, Huynh TA, Facciotti MT. JContextExplorer: a tree-based approach to facilitate cross-species genomic context comparison. BMC Bioinformatics. BMC Bioinformatics; 2013; 14: 18. doi: 10.1186/1471-2105-14-18 PMID: 23324080

48. Gushchin I, Reshetnyak A, Borshchevskiy V, Ishchenko A, Round E, Grudinin S, et al. Active state of sensory rhodopsin II: structural determinants for signal transfer and proton pumping. J Mol Biol. Elsevier Ltd; 2011; 412: 591–600. doi: 10.1016/j.jmb.2011.07.022

49. Yarov-Yarovoy V, DeCaen PG, Westenbroek RE, Pan C-Y, Scheuer T, Baker D, et al. Structural basis for gating charge movement in the voltage sensor of a sodium channel. Proc Natl Acad Sci U S A. 2012; 109: E93–102. doi: 10.1073/pnas.1118434109 PMID: 22160714

50. Yarov-Yarovoy V, Schonbrun J, Baker D. Multipass membrane protein structure prediction using Rosetta. Proteins. 2006; 62: 1010–25. doi: 10.1002/prot.20817 PMID: 16372357

51. Barth P, Schonbrun J, Baker D. Toward high-resolution prediction and design of transmembrane helical protein structures. Proc Natl Acad Sci U S A. 2007; 104: 15682–7. doi: 10.1073/pnas.0702515104 PMID: 17905872

52. Hildebrand A, Remmert M, Biegert A, Söding J. Fast and accurate automatic structure prediction with HHpred. Proteins. 2009; 77 Suppl 9: 128–32. doi: 10.1002/prot.22499 PMID: 19626712

53. Söding J, Biegert A, Lupas AN. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res. 2005; 33: W244–8. doi: 10.1093/nar/gki408 PMID: 15980461

54. Wang C, Bradley P, Baker D. Protein-protein docking with backbone flexibility. J Mol Biol. 2007; 373: 503–19. doi: 10.1016/j.jmb.2007.07.050 PMID: 17825317

55. Mandell DJ, Coutsias EA, Kortemme T. Sub-angstrom accuracy in protein loop reconstruction by robotics-inspired conformational sampling. Nat Methods. Nature Publishing Group; 2009; 6: 551–2. doi: 10.1038/nmeth0809-551

56. Rohl CA, Strauss CEM, Misura KMS, Baker D. Protein structure prediction using Rosetta. Methods Enzymol. 2004; 383: 66–93. doi: 10.1016/S0076-6879(04)83004-0 PMID: 15063647

57. Bonneau R, Strauss CE., Rohl CA, Chivian D, Bradley P, Malmström L, et al. De Novo Prediction of Three-dimensional Structures for Major Protein Families. J Mol Biol. 2002; 322: 65–78. doi: 10.1016/S0022-2836(02)00698-8 PMID: 12215415

58. Chovancova E, Pavelka A, Benes P, Strnad O, Brezovsky J, Kozlikova B, et al. CAVER 3.0: a tool for the analysis of transport pathways in dynamic protein structures. PLoS Comput Biol. 2012; 8: e1002708. doi: 10.1371/journal.pcbi.1002708 PMID: 23093919

59. Borshchevskiy V, Round E, Erofeev I, Weik M, Ishchenko A, Gushchin I, et al. Low-dose X-ray radiation induces structural alterations in proteins. Acta Crystallogr Sect D Biol Crystallogr. 2014; 70: 2675–2685. doi: 10.1107/S1399004714017295

60. Kolbe M. Structure of the Light-Driven Chloride Pump Halorhodopsin at 1.8 Å Resolution. Science (80-). 2000; 288: 1390–1396. doi: 10.1126/science.288.5470.1390

61. MacPyMOL. Schrodinger, LLC;

62. Trott O, Olson A. AutoDock Vilna: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. J Comput Chem. 2010; 31: 455–461. doi: 10.1002/jcc.21334 PMID: 19499576

63. Morris G, Huey R, Lindstrom W, Sanner M, Belew R, Goodsell D, et al. AutoDock4 and AutoDock-Tools4: Automated docking with selective receptor flexibility. J Comput Chem. 2009; 30: 2785–2791. doi: 10.1002/jcc.21256 PMID: 19399780

64. Dallakyan S, Olson A. Small-molecule library screening by docking with PyRx. Hempel J, Williams C, Hong C, editors. Springer New York; 2015.

65. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera—A visualization system for exploratory research and analysis. J Comput Chem. 2004; 25: 1605–1612. doi: 10.1002/jcc.20084 PMID: 15264254

66. Sander T, Freyss J, von Korff M, Rufener C. DataWarrior: An Open-Source Program For Chemistry Aware Data Visualization And Analysis. J Chem Inf Model. 2015; 150202131703006. doi: 10.1021/ci500588j

67. Gütlein M, Karwath A, Kramer S. CheS-Mapper 2. 0 for visual validation of (Q) SAR models. 2014; 1–18.

68. Gu J, Gui Y, Chen L, Yuan G, Lu HZ, Xu X. Use of Natural Products as Chemical Library for Drug Discovery and Network Pharmacology. PLoS One. 2013; 8: 1–10. doi: 10.1371/journal.pone.0062839

69. O'Boyle NM, Banck M, James CA, Morley C, Vandermeersch T, Hutchison GR. Open Babel: An Open chemical toolbox. J Cheminform. Chemistry Central Ltd; 2011; 3: 33. doi: 10.1186/1758-2946-3-33

70. Mysinger MM, Carchia M, Irwin JJ, Shoichet BK. Directory of useful decoys, enhanced (DUD-E): Better ligands and decoys for better benchmarking. J Med Chem. 2012; 55: 6582–6594. doi: 10.1021/jm300687e PMID: 22716043

71. Price MN, Dehal PS, Arkin AP. FastTree 2—approximately maximum-likelihood trees for large alignments. Poon AFY, editor. PLoS One. Public Library of Science; 2010; 5: e9490. doi: 10.1371/journal.pone.0009490

72. Price MN, Dehal PS, Arkin AP. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol. 2009; 26: 1641–50. doi: 10.1093/molbev/msp077 PMID: 19377059

73. Felsenstein J. PHYLIP (Phylogeny Inference Package) version 3.69. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle. 2005;

74. Price M. Fast Tree-Comparison Tools. Distributed by the author. Department of Computational and Theoretical Biology, Lawrence Berkeley National Laboratory.

75. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, et al. The Pfam protein families database. Nucleic Acids Res. 2010; 38: D211–22. doi: 10.1093/nar/gkp985 PMID: 19920124

76. The Global Proteome Machine Organization. X! Tandem CYCLONE [Internet]. 2013. Available: www.thegpm.org

77. Proteome Software Inc. Scaffold v4.0.3 [Internet].

78. Nesvizhskii AI, Keller A, Kolker E, Aebersold R. A Statistical Model for Identifying Proteins by Tandem Mass Spectrometry abilities that proteins are present in a sample on the basis. Anal Chem 2003; 75: 4646–4658.