2017

Examining Interactions between Opsins and Carotenoid Biosynthetic Proteins in Halophilic Archaea

Alexandru M. Plesa
Colby College

Examining Interactions between Opsins and Carotenoid Biosynthetic Proteins in Halophilic Archaea

Alexandru M. Pleşa

Biology Department

Colby College

Waterville, Maine

May 10, 2017

A thesis submitted to the faculty of the Honors Thesis Committee in partial fulfillment of the graduation requirements for the Degree of Bachelor of Arts with honors in Biology: Cell and Molecular Biology/Biochemistry

Ronald Peck, Advisor

Russell Johnson, Reader

Kevin Rice, Reader

Table of Contents

Abstract

Organisms have evolved numerous specialized molecules for constantly responding to environmental changes. Examples of such molecules are the light-driven proton-pump rhodopsins, such as bacteriorhodopsin (BR) and cruxrhodopsin (cR), and the carotenoid pigments, such as retinal and bacterioruberin. In halophilic Archaea, retinal can covalently bind bacterioopsin (BO) and cruxopsin (CO) to form the corresponding protein complexes, and its biosynthesis is indirectly controlled by the activity of the lycopene elongase (Lye) enzyme, which converts lycopene, a retinal precursor, to a form of bacterioruberin. Interestingly, opsins were shown to inhibit the activity of Lye, thereby promoting retinal biosynthesis and indirectly regulating the apoprotein-cofactor stoichiometry. This is a newly described regulatory mechanism, and, considering the importance of the problem it addresses, we set to determine the protein domains involved in the opsin-Lye inhibition. Using a fusion protein approach, we determined that a 52 amino acid domain in Lye, a 2 amino acid section in BO, and 34 and 43 amino acid regions in CO are required for the studied interaction. Furthermore, we compared the proteins' tertiary structures and found supporting evidence for the validity of our identified regions and for the localization of the interaction at the interface of the lipid bilayer and the cytoplasm. Future studies could further investigate this recently discovered regulatory mechanism by identifying the participating protein amino acids more precisely and by searching for homologous domains in other biological systems.

Introduction

Environments are constantly changing, and, in order to survive, organisms have to respond to these changes. To achieve this function, many life forms have evolved specialized molecules to mediate environmental interactions. Two types of such molecules are opsins and carotenoids.

Opsins

Opsins are light sensing proteins found in all domains of life: Archaea, Bacteria, and Eukarya[1]. Their structure consists of seven transmembrane (TM) α-helices that form a binding pocket for a light-reactive chromophore, which confers the protein's light sensitivity. The chromophore is a vitamin-A based retinaldehyde that forms a covalent bond to a lysine residue from the seventh TM helix (Figure 1). These proteins have been classified in two functionally similar families, type I and type II, based on primary sequence alignments. Type I opsins (microbial opsins) are found in both prokaryotes and eukaryotes, while type II opsins are present only in higher eukaryotes[2,3].

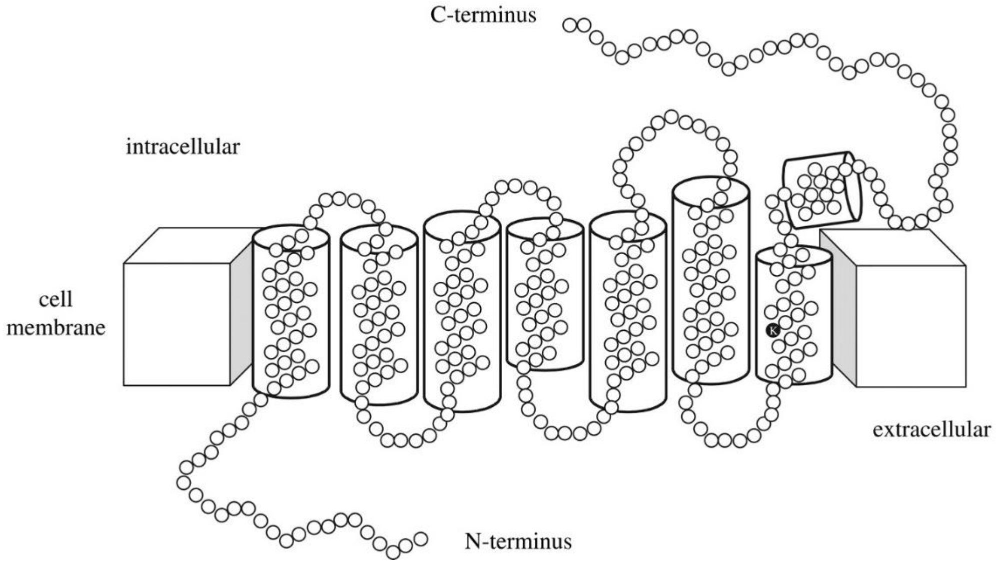

Figure 1. Representation of the general structure of opsins[4]. Highlighted lysine is the retinal binding residue.

On Earth, most life is ultimately dependent on light energy, making light one of the most important environmental signals[5]. As such, most organisms require light-sensing molecules for either energy conversion or signal transduction. Opsins can perform both these functions by acting as G protein-coupled receptors (GPCR) (type II opsins involved in visual perception and circadian rhythms), photoreceptors (type I opsins: sensory rhodopsins I and II), or light-driven ion pumps (type I opsins: bacteriorhodopsin, cruxrhodopsin)[6,7].

Bacteriorhodopsin (BR) was the first discovered microbial opsin, and it is produced by the halophilic Archaeon *Halobacterium salinarum*[3,8]. BR is a complex consisting of the bacterioopsin (BO) apoprotein covalently bound to the retinal cofactor. It is a small (26 kDa) 7-TM light-driven ion pump that uses the light-induced photoisomerization of all-*trans* retinal to 13-*cis* retinal to generate a proton gradient. The gradient is used for the ATP synthase catalyzed conversion of ADP to ATP and for other functions such as active transport and flagellar rotation [9–11]. In *H. salinarum*, BR is used for energy conversion when oxygen levels are too low to sustain aerobic respiration[12].

The archaeal opsin cruxrhodopsin (cR), a homolog of BR, generates a proton gradient in a related halophilic Archaeon, *Haloarcula vallismortis*[13,14]. However, despite this primary function similarity, there is only 48-54% primary sequence identity between the two opsins, which raises the possibility that the two proteins have other, different regulatory functions. Evidence of this possibility comes from the recently resolved crystal structure of cR which displays several structural differences from BR[15].

BR is a stable protein over a wide range of temperature, pH, and salt concentrations. The stability of this opsin along with its light-sensing function have made it the subject of many applications in biology and biotechnology[16–18]. Some of these applications include: optogenetics[19,20], light sensors[21–23], data storage[24,25], and artificial retinal implants[26,27]. cR has not been studied as extensively as BR, but due to the shared light-sensing property, this protein has also been used in neuronal silencing[28,29]. Due to the large number of applications for opsins, there is currently an active interest to discover more about their regulation and functions in their native organisms.

Carotenoids

Carotenoids are organic pigments synthesized by archaea, bacteria, fungi, plants, and some animals (aphids and spider mites)[30,31]. Many life forms produce carotenoids for their roles in a wide range of processes from UV protection and free radical defense to membrane stabilization and photosynthesis[32–37]. One of these carotenoids is bacterioruberin, a C_{50} pink-colored pigment found in various species of microbes, whose function still remains widely unknown.

There have been many studies aimed at discovering the role of bacterioruberin in microbial organisms, but little agreement has been reached due to contradictory results. Some experiments showed that bacterioruberin is involved in DNA protection against UV damage, and that its biosynthesis is enhanced by light exposure[32,38]. However, preliminary results (Peck, unpublished) suggest that, in *Haloferax volcanii*, there is no correlation between bacterioruberin production and UV cytotoxicity protection. Similarly, recent results (Peck, unpublished) indicate that the presence of bacterioruberin has no effect on osmotic stress resistance. This finding is contrary to previous studies, which proposed that bacterioruberin plays a role in osmotic shock protection, on the basis of data showing that this carotenoid reinforces membrane structure[35].

While there is currently no convincing information on the function of bacterioruberin in the microbial cell, this carotenoid and many others have been used in the food, cosmetics and pharmaceutical industries for their coloring, antioxidant, and potential anti-cancer properties[39–41].

For example, recent studies have reported that carotenoids have a significant antiproliferative activity against HepG2 human cell cancer lines, and there have been patents filed for the use of halobacteria extracts as tumor reduction treatments[42]. Moreover, bacterioruberin was shown to have a significantly higher free-radical scavenging capacity than β-carotene, a potent antioxidant[43].

Another important role of carotenoids is their function as cofactors for various proteins involved in many processes like gas transport, energy conservation, light-sensing and signal transduction. While these protein complexes are of large importance for the homeostasis of an organism, little is known about how cofactor biosynthesis and apoprotein expression are coordinated to achieve an appropriate stoichiometry for a functional complex[44]. When this optimal ratio of apoprotein to cofactor is not achieved, either of the molecules can accumulate and lead to pathologies like porphyria (heme precursors accumulation) and retinitis pigmentosa (opsins aggregation)[45,46]. Motivated by the importance of an optimal ratio of protein complex components, we used rhodopsins as model proteins to study the mechanisms involved in regulating cofactor biosynthesis and apoprotein expression that lead to the formation of functional complexes.

Opsins and carotenoids in halophilic Archaea

In its native organism, *H. salinarum*, bacterioopsin (BO) binds retinal in a 1:1 stoichiometry to form a light-driven proton complex (BR)[47]. Retinal, the cofactor, is synthesized in a multi-step process from a geranylgeranyl pyrophosphate precursor (Figure 2)[48]. During low-oxygen conditions, *H. salinarum* can increase BR biosynthesis up to 50-fold in only a few hours of growth[49]. This BR induction requires a corresponding increase in the production of both apoprotein (BO) and cofactor (retinal), which is achieved by higher transcription levels of the BO gene (*bop*) and of other genes encoding retinal biosynthetic enzymes[50]. However, this does not fully explain how the appropriate stoichiometry between BO and retinal is maintained.

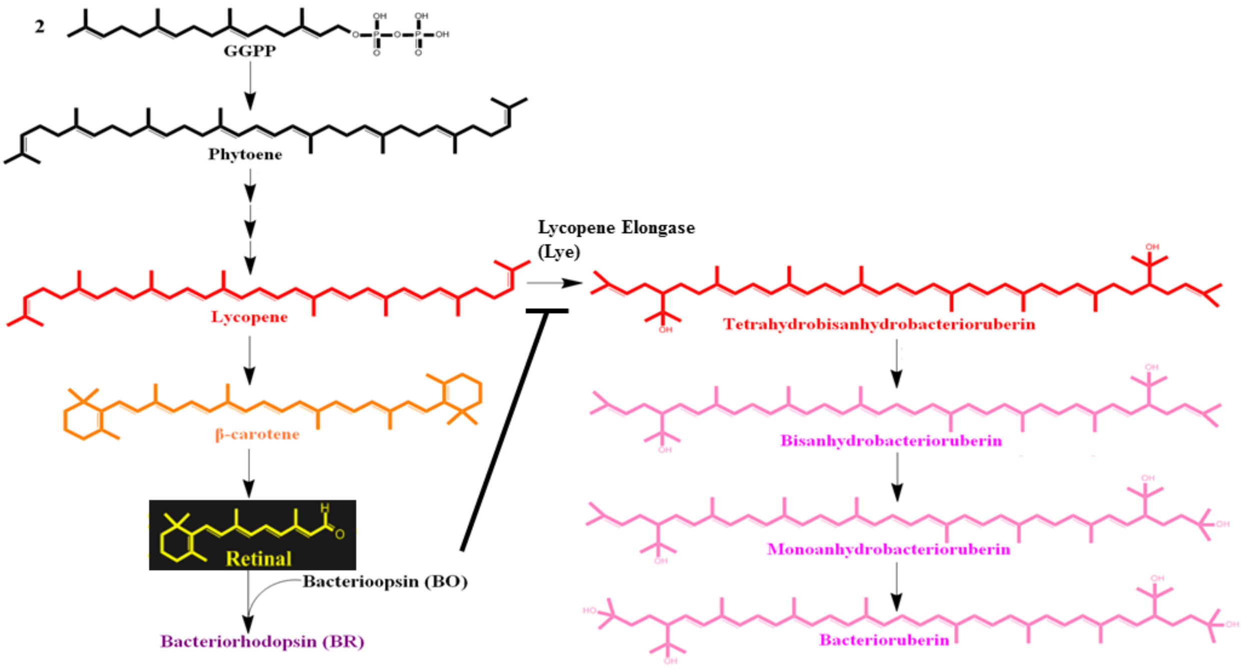

Figure 2. Proposed pathway for retinal and bacterioruberin biosynthesis in *H. salinarum*[44]

Interestingly, previous results suggest that the BO apoprotein itself plays an important role in regulating the production of the retinal cofactor. When not bound to retinal, BO inhibits the activity of lycopene elongase (Lye), an enzyme that catalyzes the committed step in the synthesis of bacterioruberin (Figure 2). Since bacterioruberin has common precursors with retinal, the inhibition of Lye by BO likely acts to promote retinal biosynthesis at the expense of bacterioruberin biosynthesis. Based on this observation and other unpublished data, it has been proposed that BO regulates retinal production through a previously unknown regulatory mechanism that involves a transient interaction between BO and Lye[44].

Furthermore, this interaction between Lye and BO seems to be a specific one. For instance, the Lye homolog of *H. volcanii*, a microbe closely related to *H. salinarum* that does not produce any opsins, is not inhibited by BO, despite a 65% primary sequence identity between the two Lye proteins[44]. Similarly, CO, which shares a 53% primary sequence identity with BO, can inhibit the function of *H. vallismortis* Lye, but has no effect on *H. salinarum* Lye (Peck et al., submitted). This specificity suggests that the opsins experienced selection, which was likely driven by an advantage conferred by the ability to regulate carotenoid biosynthesis through Lye inhibition. The goal of this study was to further characterize the newly described mechanism by determining the protein domains involved in the observed inhibitory interaction, which may provide important insights into cofactor biosynthesis regulation in other biological systems.

Summary of results

Here, we present results that identify the Lye amino acid sequence required for the protein's inhibition by BO. This identified region is largely conserved between the two Lye homologs, with the exception of small sections of disordered structure, which may be responsible for the interaction with BO. We also present a script, Similarity Optimized Backtranslator (SOB), capable of backtranslating amino acid sequences to DNA sequences of high similarity, which we used for generating fusion proteins of BO and CO, through a gene shuffling method. Similarly, we determined the opsin amino acids required for the specific interactions with the Lye homologs and studied their localization using the reported opsin crystal structures. The two proteins had high structural similarity in the identified regions, and, more importantly, all BO, CO, and Lye domains required for inhibition were situated in a similar location, in the cytoplasmic section of the protein between two transmembrane α-helices.

Materials and Methods

Strains, plasmids, genes, and primers

H. volcanii strains and plasmids used in the study are listed in Table 1. Synthetic genes and primers sequences are listed in Table 2. All primers were ordered from IDT (Coralville, IA).

Table 1. *H. volcanii* strains and plasmids used in this study

Name	Description	Ref.
H1209	Δ*mrr*, Δ*pyrE2*, Δ*hdrB*, *pitA* replaced with *pitA* from another organism	[51]
RFP58	H1209 Δ*lye*	this study
RFP131	H1209 Δ*lye::H. vallismortis lye*	this study
RFP152	H1209 Δ*lye::H. salinarum lye*	this study
RFP181	H1209 Δ*lye::lye(Hsal 1-125, Hvol 152-301)*	this study
RPF182	H1209 Δ*lye::lye(Hsal 1-198, Hvol 225-301)*	this study
RFP188	H1209 Δ*lye::lye(Hsal 1-67, Hvol 94-301)*	this study
RFP192	H1209 Δ*lye::lye(Hsal 1-125, Hvol 152-301)* with pRFP121 (empty vector)	this study
RFP193	H1209 Δ*lye::lye(Hsal 1-125, Hvol 152-301)* with pRFP126 (*bop* expression)	this study
RFP194	H1209 Δ*lye::lye(Hsal 1-198, Hvol 225-301)*with pRFP121 (empty vector)	this study
RFP195	H1209 Δ*lye::lye(Hsal 1-198, Hvol 225-301)* with pRFP126 (*bop* expression)	this study
RFP200	H1209 Δ*lye::lye(Hsal 1-67, Hvol 94-301)* with pRFP121 (empty vector)	this study
RFP201	H1209 Δ*lye::lye(Hsal 1-67, Hvol 94-301)* with pRFP126 (*bop* expression)	this study
pMPK408	Plasmid for *H. salinarum* gene replacements - has *E. Coli* vector with *ura3, MevR* & polylinker	[52]
pTA131	*H. volcanii* integrative plasmid for generating knockouts	[53]
pTA963	*H. volcanii* overexpression plasmid	[51]
pRFP126	pTA963 – with *bop* for gene overexpression	this study
pRFP128	pTA131 – plasmid to insert genes into the *H. volcanii lye* locus	this study
pRFP147	pMPK408 – plasmid to insert genes into the *H. salinarum lye* locus	this study
pRFP158	pRFP147 – *lye(Hsal 1-125, Hvol 152-301)*	this study
pRFP177	pRFP147 – *lye(Hsal 1-198, Hvol 225-301)*	this study
pRFP184	pRFP147 – *lye(Hsal 1-67, Hvol 94-301)*	this study
pRFP237	pRFP128 – *lye(Hsal 1-125, Hvol 152-301)*	this study
pRFP238	pRFP128 – *lye(Hsal 1-198, Hvol 225-301)*	this study
pRFP241	pRFP128 – *lye(Hsal 1-67, Hvol 94-301)*	this study
pRFP268	pUC57 – *GS_bop*	this study
pRFP269	pUC57 – *GS_cop3*	this study

Table 2. DNA sequences of the synthetic genes and primers used in this study

Name	Sequence
GS_bop	GCCATGCCACTGGAATCTGCACATATGCTCGAGCTCCTCCCGACCGCGGTCGAGGGCGTCTCGCAGGCGCAGATCACCGGCCGCCCCGAGTGGATCTGGCTCGCGCTCGGCACCGCGCTCATGGGCCTCGGCACGCTCTACTTCCTCGTGAAGGGCATGGGCGTGTCCGACCCGGACGCGAAGAAGTTCTACGCCATCACCACCCTCGTCCCCGCGATCGCGTTCACCATGTACCTCTCGATGCTGCTCGGCTACGGCCTCACCATGGTCCCGTTCGGGGGCGAGCAGAACCCGATCTACTGGGCGCGCTACGCGGACTGGCTCTTCACCACCCCGCTCCTCCTCCTCGACCTCGCCCTCCTCGTGGACGCGGACCAGGGCACCATCTTGGCGCTCGTCGGCGCCGACGGCATCATGATCGGCACCGGCCTCGTCGGGGCCCTCACGAAGGTCTACTCGTACCGCTTCGTCTGGTGGGCCATCTCGACCGCGGCCATGCTCTACATCCTCTACGTCCTCTTCTTCGGCTTCACCTCGAAGGCGGAGTCGATGCGCCCGGAGGTCGCCTCGACCTTCAAGGTCCTCCGCAACGTCACCGTCGTCCTCTGGTCGGCCTACCCGGTCGTGTGGCTCATCGGCTCCGAGGGCGCCGGCATCGTCCCCCTCAACATCGAGACCCTGCTCTTCATGGTCCTCGACGTCAGCGCGAAGGTCGGCTTCGGCCTCATCCTCCTCCGCTCGCGCGCGATCTTCGGCGAGGCGGAGGCGCCGGAGCCGTCGGCCGGCGACGGCGCGGCCGCGACCTCGGACTGAGAATTCCGATTCCCAGAATGTAAGCGATTCCCAGAATGTAAG
GS_cop3	GCCATGCCACTGGAATCTGCACATATGCCGGCGCCGGAGGGCGAGGCGATCTGGCTCTGGCTCGGCACCGCGGGCATGTTCCTCGGCATGCTCTACTTCATCGCGAGGGGCTGGGGCGAGACCGACTCGCGCCGCCAGAAGTTCTACATCGCCACCATCCTCATCACCGCGATCGCGTTCGTCAACTACCTCGCGATGGCGCTCGGCTTCGGCCTCACCATCGTCGAGATCGCGGGCGAGCAGCGCCCGATCTACTGGGCGCGCTACTCGGACTGGCTCTTCACCACCCCGCTCCTCCTCTACGACCTCGGCCTCCTCGCGGGCGCGGACCGGAACACCATCTCGTCGCTCGTCAGCCTCGACGTCCTCATGATCGGCACCGGCCTCGTCGCGACCCTCTCGGCGGGCTCGGGCGTCCTCTCGGCGGGCGCGGAGCGCCTCGTCTGGTGGGGCATCTCGACCGCGTTCCTGCTCGTCCTCCTCTACTTCCTCTTCTCCTCGCTCTCGGGCCGCGTCGCGGACCTCCCGTCGGACACCCGCTCGACCTTCAAGACCCTCCGCAACCTCGTCACCGTCGTCTGGTTGGTCTACCCGGTCTGGTGGCTCGTCGGCACCGAGGGCATCGGCCTCGTCGGCATCGGCATCGAGACCGCGGGCTTCATGGTCATCGACCTCGTCGCGAAGGTCGGCTTCGGCATCATCCTCCTCCGCTCGCACGGCGTCCTCGACGGGGCGGCGGAGACCACCGGCGCCGGCGCGACCGCGACCGCGGACTGAGAATTCCGATTCCCAGAATGTAAGCGATTCCCAGAATGTAAG
RP183	AAAACATATGATGTTCCGGTATCTGTTCGTGT
RP246	AAAATCTAGACTTCGGGCTCGGCGTCTACTATC
RP268	CGCTCTCGAAGCTGTTTCTC
RP269	AAAAATGCGATGGTCCAGAG
RP273	AAAACATATGCCATTGACGAGCCTCCA
RP274	ACGGAGTACAGCGCACCCCCGTTTCGGTTCAAGACGA
RP275	GGGTGCGCTGTACTCCGT
RP276	TTCGACCGCGACGTCGACGAAGCGAACCCGAAGAAG
RP277	GTCGACGTCGCGGTCGAA
RP278	ACCACTGCGACCGCGCTGGGCGAGCGGCGGACCTA
RP279	CAGCGCGGTCGCAGTGGT
RP410	AAAAAGATCTTTAGCCATTGACGAGCCTCCA
RP449	CATGCCACTGGAATCTGCAC
RP450	GGGAATCGCTTACATTCTGGG

Cultivation conditions

H. volcanii was grown at 40°C in Hv-YPC[53] liquid medium or agar supplemented with thymidine (40 μg/mL). *Escherichia coli* DH5α (New England Biolabs, Ipswich, MA) was grown in LB medium or agar supplemented with ampicillin (50 μg/ml) at 37°C. Liquid cultures were grown with shaking at 250 rpm.

Plasmid construction

The Lye fusion genes (pRFP158, pRFP177, pRFP184) were constructed by two-step PCR. Firstly, *Halobacterium sp.* NRC-1[54] genomic DNA was used as template in PCR with primer pairs RP275-RP246 (pRFP158), RP279-RP246 (pRFP177), RP277-RP246 (pRFP184) and *Haloferax sp.* DS2[55] genomic DNA was used as template in PCR with primer pairs RP274-RP273 (pRFP158), RP278-RP273 (pRFP177), and RP276-RP273 (pRFP184). Secondly, the first step PCR products were used as template with the primer pair RP246-RP273, and the resulting amplicons were digested with XbaI and NdeI and ligated into the XbaI-NdeI fragment of pRFP147.

To construct the plasmids pRFP237, pRFP238, and pRFP241 for inserting the Lye fusion genes into the genome of *H. volcanii*, plasmids pRFP158, pRFP177, and pRFI184, respectively, were used as template in PCR with primer pair RP183-RP410. The PCR products were then digested with NdeI and BglII and ligated into the NdeI-BglII 4 kb fragment of pRFP128.

Plasmids pRFP28, pRFP120, pRFP121, pRFP126, pRFP128, and pRFP147 were previously constructed in the Peck lab. The vectors harboring the synthetic genes (pFP268, pRFP269) were ordered from Genscript (Piscataway, NJ).

Haloferax strain construction

The *H. volcanii* strains expressing Lye fusion proteins were constructed by integrating the *lye* genes from plasmids pRFP237, pRFP238, and pRFP241 into the genome of the Δ*lye* strain RFP58 using a previously described gene replacement method[53]. Then, the resulting strains were transformed with plasmid pRFP126 (and pTA963 as a control), which allows *H. volcanii* to express *H. salinarum bop* (Peck et al., submitted), in order to form the corresponding strains RFP192, RFP193, RFP194, RFP195, RFP200, and RFP201.

Structure prediction

The structural models for the two Lye homologs were created using Phyre2 intensive modeling mode[56] and superimposed with SuperPose[57]. Specific regions of the structures were highlighted and the figures were exported to image files using Geneious.

Gene design, shuffling, and transformation

To increase the efficiency of gene shuffling, the genes encoding BO and CO were designed *in silico* for high DNA sequence similarity. This was achieved by loading the amino acid sequences of the two proteins in the SOB Python script we developed.

The ≈800 bp BamHI/XbaI fragments of pRFP268 and pRFP269 were gel-purified and PCR-amplified using primers RP449 and RP450. Gene shuffling was performed on the obtained amplicons as described in Meyer et al.[58] with the following modifications. Phusion High-Fidelity DNA Polymerase (New England Biolabs, Ipswich, MA) was used for all PCR steps. The library amplification was performed using 20 µL of the elution reassembly product in four 50 µL PCR reactions with primer pair RP449-RP450, and the product (GS-library) of this reaction was PCR-purified and concentrated by vacuum centrifugation.

The NdeI-EcoRI digested GS-library was size reduced to 750-1100 bp through gel extraction and ligated into the ≈8300 bp NdeI-EcoRI fragment of pRFP121. This library of plasmids containing the shuffled genes was then transformed in *E. coli* competent cells, which were grown overnight in a 200 mL culture, after a 1h recovery step. Plasmid DNA was extracted from the culture and transformed into the RFP131 and RFP152 strains using the previously mentioned protocol. The white transformed colonies were used for colony PCR to amplify and sequence the shuffled gene (Eurofins Genmoics, Louisville, KY).

Colorimetric assay

Bacterioruberin levels of the strains expressing Lye fusion proteins were analyzed using a colorimetric procedure (Peck et al., submitted). Colonies grown on Hv-YPC solid medium were photographed in a reproducible manner and the pictures' color properties (hue and saturation) were used to determine a ruberin index, which represents a measure of the bacterioruberin to lycopene ratio. A high ruberin index is correlated to a larger amount of bacterioruberin compared to lycopene and indicates a low level of Lye inhibition.

Results

Lye fusion

To determine the Lye amino acid region involved in the inhibitory interaction with BO, we constructed hybrid genes by fusing parts of the *lye* homologs from *H. salinarum* and *H. volcanii* and tested the encoded proteins for inhibition by BO. For this, we integrated the fusion genes into the *lye* locus of *H. volcanii,* and we transformed an expression vector with or without *bop*. Then, we used a colorimetric assay for comparing the bacterioruberin to lycopene ratio between the strains expressing BO and those that were not, to determine which fusion proteins were inhibited (Figure 3). The strains expressing fusion proteins *H. sal* 1-120 – *H.* vol 147-301 and *H. sal* 1-192 – *H. vol* 219-301 had a significantly reduced ruberin index when BO was co-expressed from plasmid, indicating the presence of BO-Lye inhibition (Figure 3). Using these data, we determined that inhibition was always present when the 69-120 region was from *H. salinarum* Lye (Figure 4).

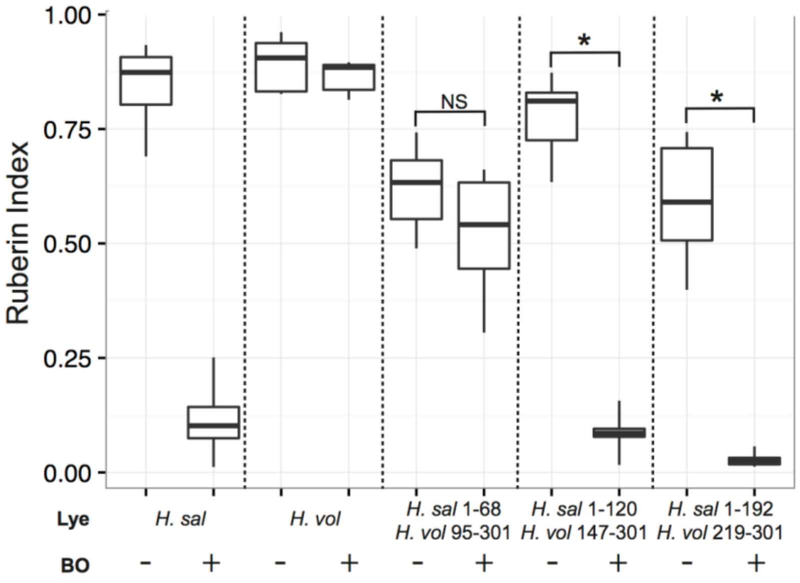

Figure 3. Ruberin index for strains expressing different Lye proteins in the presence and absence of BO. Permutation test. (*, p<0.05)

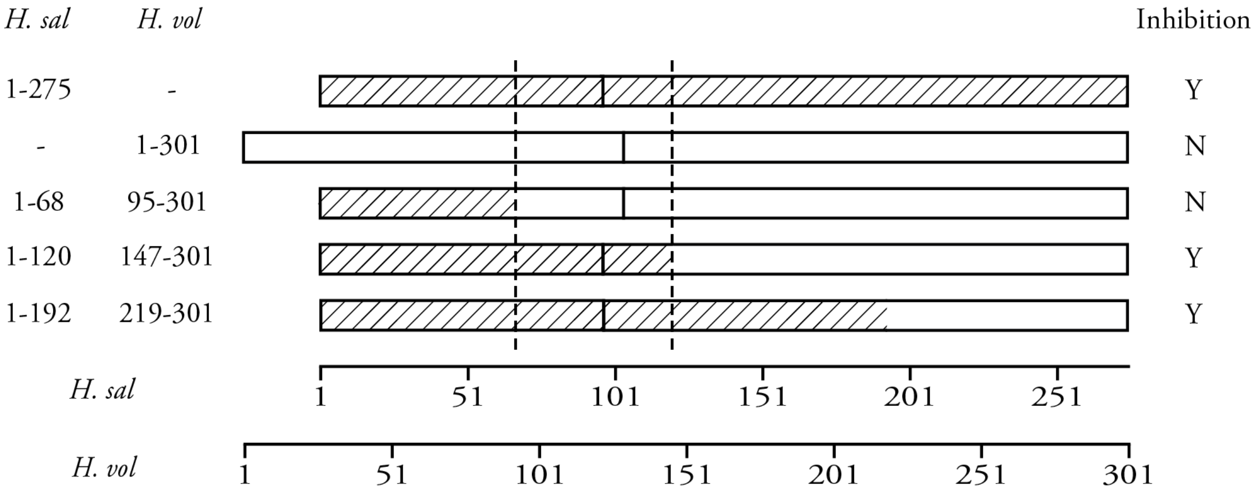

Figure 4. Protein alignment map of the Lye homologs and three of their fusion proteins. The x-axes represent the amino acid positions in the corresponding homolog, the y-axes show the fusion protein composition using residue numbering (left) and enzyme inhibition by BO (right), and the vertical lines within the protein sequences represent gaps in the alignment. The dashed lines mark the *H. salinarum* Lye region required for inhibition by BO.

Lye structure prediction

We generated template-based structure predictions of the two Lye homologs, and compared the identified 52 amino acid region between the two models (Figure 5). Both Lye sequences aligned to 4-hydroxybenzoate octaprenyltransferase with a confidence score of 100% and >85% coverage. The validity of this structural prediction is further confirmed by the fact that both lycopene (Lye substrate) and octaprenyl (4-hydroxybenzoate octaprenyltransferase substrate) are C_{40} molecules with similar conjugated systems. Interestingly, the two protein models have a high structural similarity for the 52 amino acid section of interest, with minor differences in the third TM domain and second inter-helix coil. These differences are located in regions of low model confidence, due to the presence of a disordered section. However, it is worth noting that the only significant region of disorder in each homolog is located within the identified 52 amino acid section.

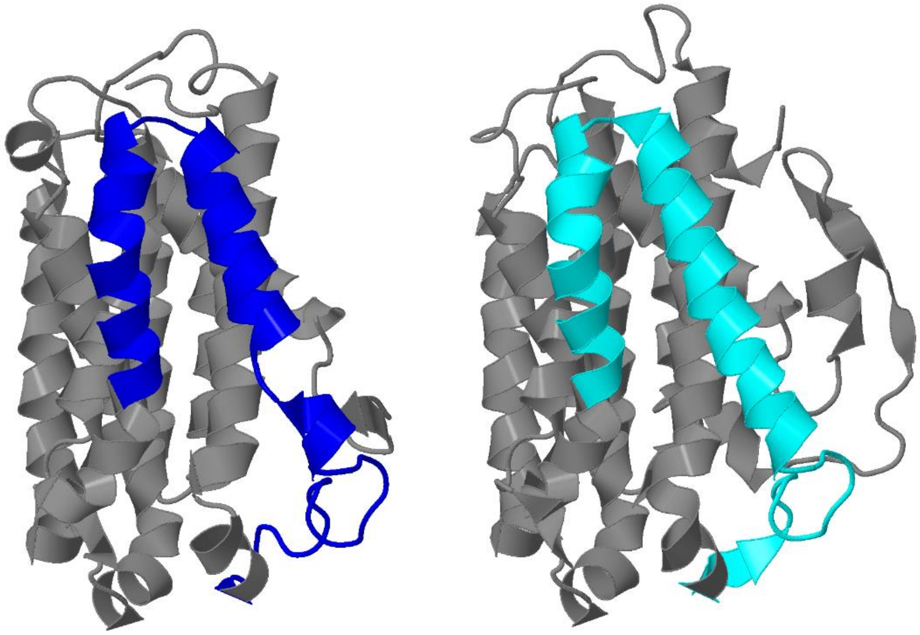

Figure 5. Tertiary structure predictions of *H. salinarum* (left) and *H. volcanii* (right) Lye. Colored regions represent the identified 52 amino acid section of interest. Structures are depicted within the membrane and oriented with the top side towards the extracellular environment.

Similarity optimized backtranslation

We studied the opsin domains required for Lye inhibition, by generating *bop-cop* fusion genes using gene shuffling. This method starts with genes of high similarity and, through progressive hybridization PCR, creates a library of randomly hybridized genes. To achieve the required initial gene sequence similarity, we wrote a Python script (SOB) that takes as input an alignment of two proteins and a specified codon usage (both files in comma separated values format) and returns a file with corresponding DNA sequences of higher similarity compared to random codon usage.

SOB works by backtranslating the two proteins into the codon optimized genes, finding regions of similarity, and then trying to extend them in both directions. More specifically, the algorithm performs pairwise comparisons of the two genes, starting at the first codon. If it finds a match larger than 3 bp, it checks the codons before and after the match site for silent mutations that could increase the similarity of the genes. When such a substitution is available, SOB changes the codons accordingly and starts the process over. The algorithm is repeated until the sequences do not change anymore, or when the maximum number of iterations set by the user is reached (Figure 6).

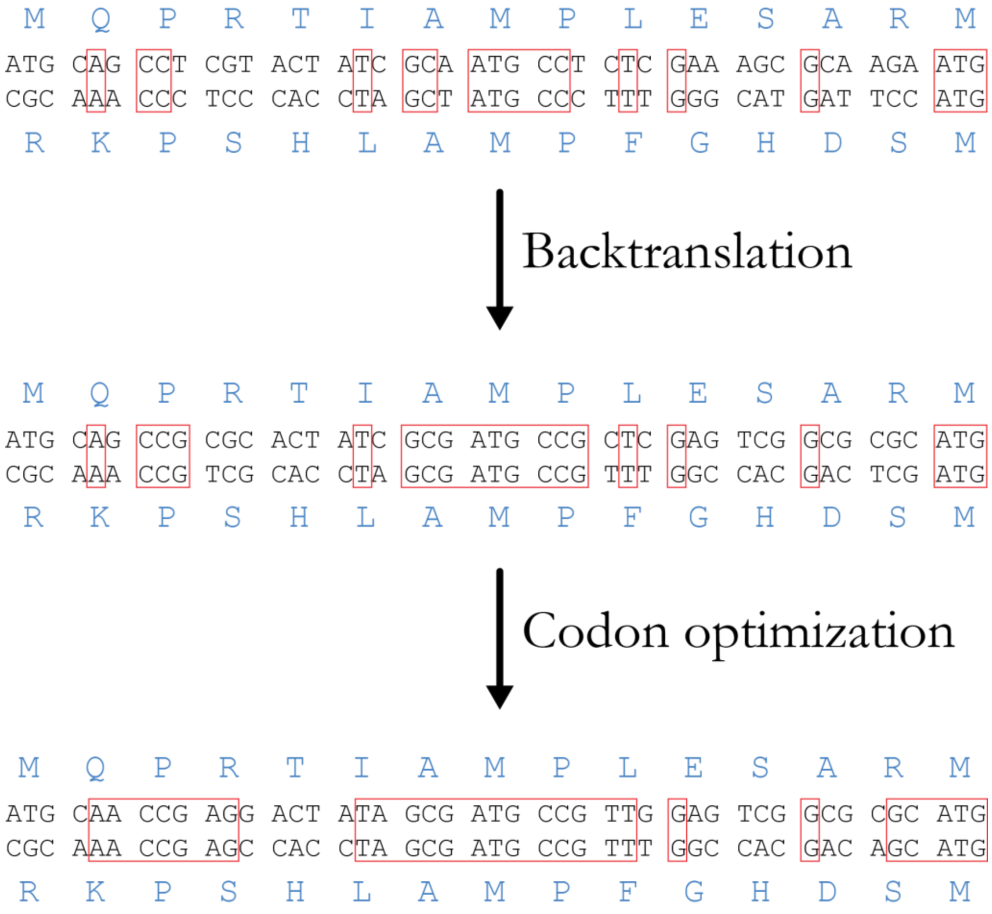
M Q P R T I A M P L E S A R M
ATG CAG CCT CGT ACT ATC GCA ATG CCT CTC GAA AGC GCA AGA ATG
CGC AAA CCC TCC CAC CTA GCT ATG CCC TTT GGG CAT GAT TCC ATG
R K P S H L A M P F G H D S M
Backtranslation
M Q P R T I A M P L E S A R M
ATG CAG CCG CGC ACT ATC GCG ATG CCG CTC GAG TCG GCG CGC ATG
CGC AAA CCG TCG CAC CTA GCG ATG CCG TTT GGC CAC GAC TCG ATG
R K P S H L A M P F G H D S M
Codon optimization
M Q P R T I A M P L E S A R M
ATG CAA CCG AGG ACT ATA GCG ATG CCG TTG GAG TCG GCG CGC ATG
CGC AAA CCG AGC CAC CTA GCG ATG CCG TTT GGC CAC GAC AGC ATG
R K P S H L A M P F G H D S M

Figure 6. SOB proof of concept using a theoretical example. The first line displays the aligned genes for the two proteins of interest, with the corresponding regions of similarity highlighted in the red rectangles. Second and third lines represent the DNA sequences after backtranslation and codon optimization, respectively. It can be observed that similarity regions are extended in both directions at each step.

The initial opsin genes, as sequenced from their native organisms, were 59% similar and had 23.5% gaps in the alignment. After the backtranslation step, the genes reached 69.1% similarity and the gaps percentage reduced to 16.9%. And, after the codon optimization step, the output of SOB consisted of two genes that were 73.3% similar and had 12% gaps in their alignment.[59] More importantly, SOB decreased the distances between the regions of identity in the alignment, thereby increasing the length of similar regions and the probability of crossing-over during gene shuffling, which subsequently increased the efficiency of our library generation (Figure 7).

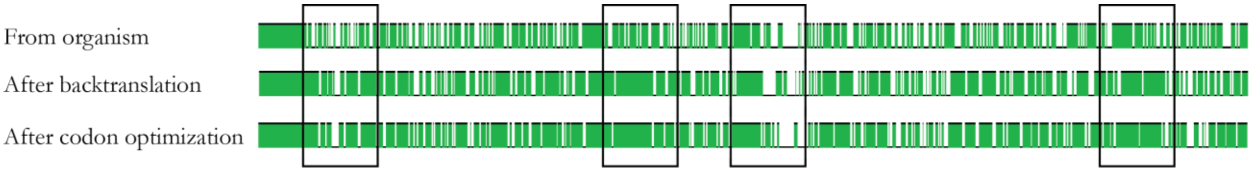

Figure 7. Pairwise sequence identity of the *bop* and *cop* genes at different stages of the optimization process. The rectangles highlight regions where there is a clear increase in sequence similarity between the input and output of our algorithm.

Gene shuffling

To identify the corresponding opsin protein domains required for the studied inhibitory interaction, we chose a fusion protein approach, similar to the one used for the Lye experiments. We used the *bop* and *cop* sequences optimized by SOB to generate a library of diverse hybrid genes, which was transformed in *H. volcanii* strains expressing *H. salinarum* or *H. vallismortis* Lye, and then we screened for constructs that inhibited either of the two homologs. Inhibition was considered present when colonies were closer in color to cream than to pink (Figure 8). Ruberin indices are currently being determined for the identified strains of interest.

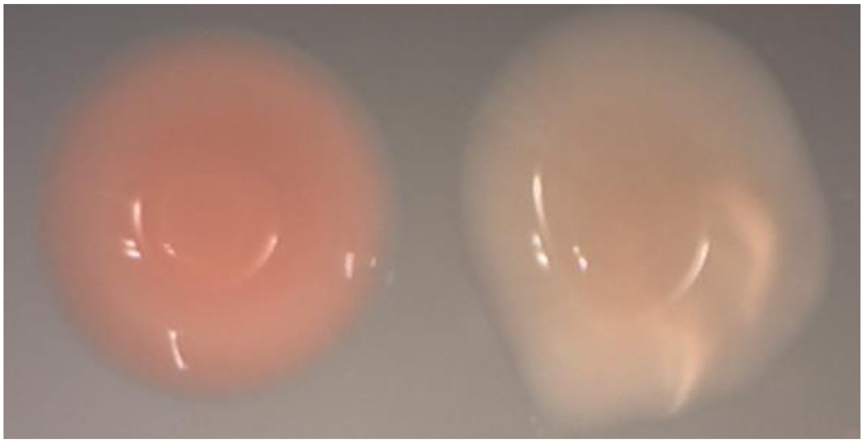

Figure 8. Representative picture of colonies from two strains expressing different opsin fusion proteins. Due to the cream color, the colony on the right was considered to exhibit Lye inhibition by the fusion opsin.

Using the color phenotype, candidate colonies were selected for screening. Their opsin genes were amplified using colony PCR and then sequenced. The sequences were aligned against both opsins to determine the gene's identity. For *H. salinarum* Lye, we determined that inhibition always occurred when amino acids 115 and 116 (Valine and Aspartate in BO and Alanine and Glycine in CO) were from BO. Inhibition was still present when any of the other regions of the protein was individually switched for the corresponding CO domain (Figure 9). Similarly, we observed *H. vallismortis* Lye inhibition whenever two specific domains (27-69 and 102-135) of CO were present in the fusion protein. Using the protein alignment, we further narrowed down the second region to 108-130, based on amino acid matches between the proteins. Individually, the origin of all the other regions, or their presence in some cases, was not observed to be important for inhibiting the *H. vallismortis* Lye homolog. An interesting result is that some of the constructs had a higher frequency in our screen than others (Supplementary Table 1).

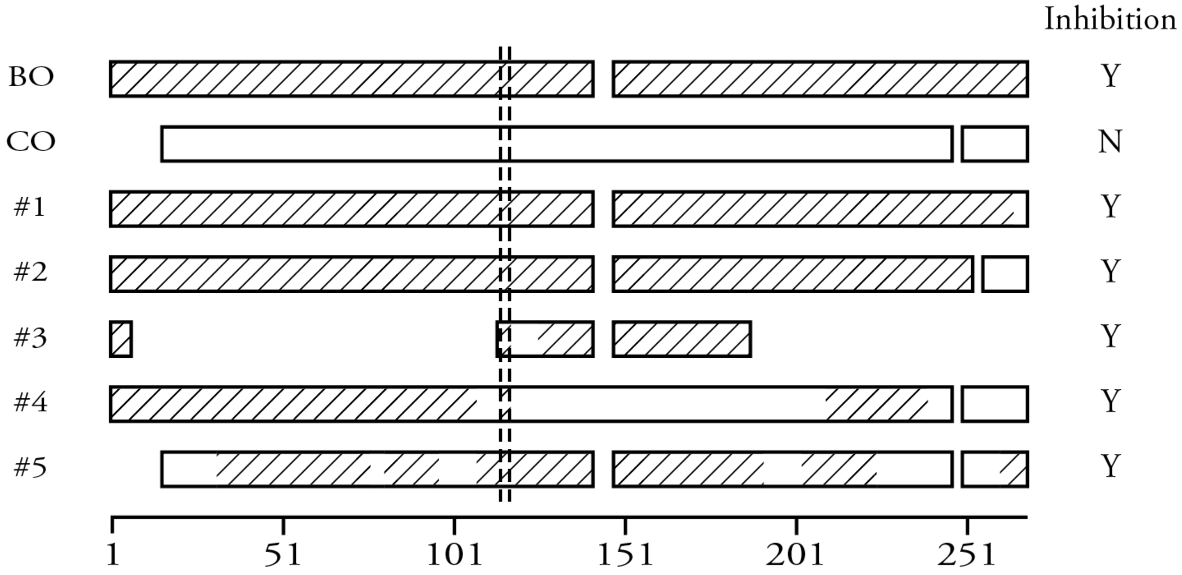

Figure 9. Protein alignment map of opsin fusions that inhibited *H. salinarum* Lye. X-axis represents alignment amino acid position, y-axes show protein names (left) and their inhibition of Lye (right), and blank sections represent alignment gaps. The dashed lines mark the opsin region required for *H. salinarum* Lye inhibition.

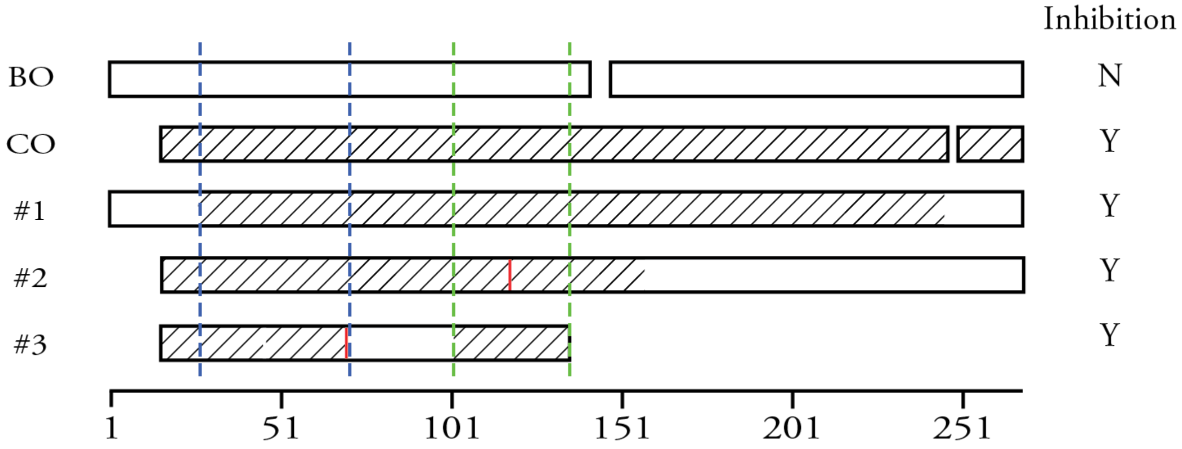

Figure 10. Protein alignment map of opsin fusions that inhibited *H. vallismortis* Lye. X-axis shows alignment amino acid position, y-axes show protein names (left) and their inhibition of Lye (right), blank sections represent alignment gaps, and red lines show substitutions. The dashed lines mark the opsin regions required for *H. vallismortis* Lye inhibition.

Opsin structure comparison

To further study the opsin-Lye interaction, we mapped the identified regions from the gene shuffling experiment onto the previously reported crystal structures of the two protein complexes (Figures 11-12).

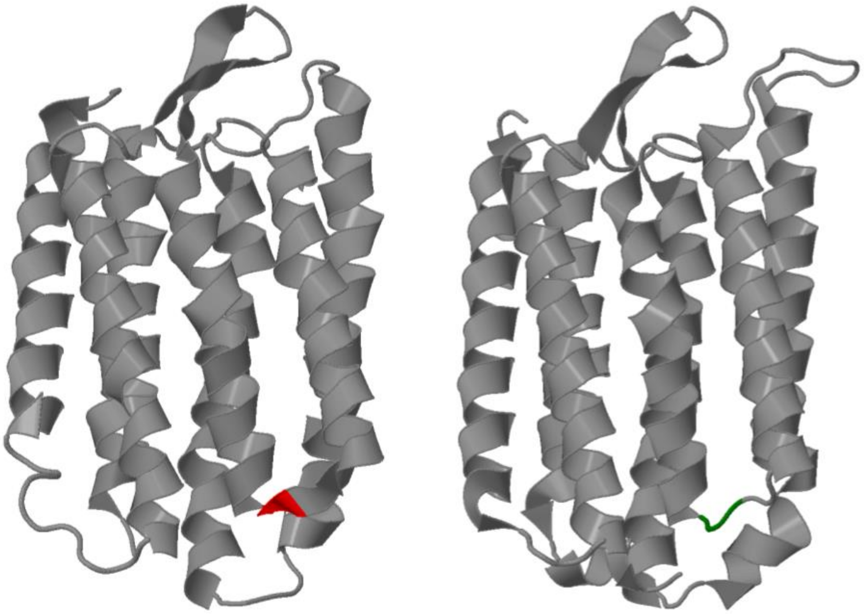

Figure 11. Tertiary structures of BR[60] (left) and cR[15] (right). Colored regions represent the identified two amino acid section required for *H. salinarum* Lye inhibition. Structures are depicted within the membrane and oriented with the top side towards the extracellular environment.

We found a high degree of structural similarity between the two opsins in the sections required for *H. vallismortis* Lye inhibition. The differences between homologs for all three identified regions consisted of small sections of disorder. It is worth noting that the regions required for the Lye interaction include cytoplasmic domains situated between two TM α-helixes, similarly to the previously described Lye regions required for inhibition.

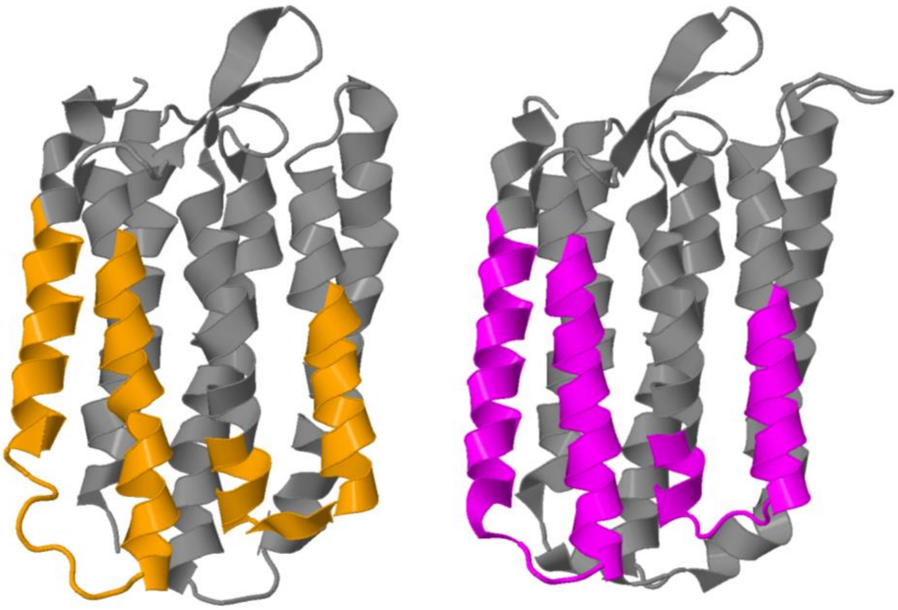

Figure 12. Tertiary structures of BR[60] (left) and cR[15] (right). Colored regions represent the identified 43 and 34 amino acid sections required for *H. vallismortis* Lye inhibition. Structures are depicted within the membrane and oriented with the top side towards the extracellular environment.

Discussion

Lycopene elongase

Lye catalyzes the conversion of lycopene to tetrahydrobisanhydrobacterioruberin as its main primary function. Recent results suggest that the protein also has a regulatory role in the biosynthesis of retinal through its inhibition by opsins. To further characterize this recently discovered interaction and determine the Lye protein domains involved in it, we generated fusion proteins from the *H. salinarum* and *H. volcanii* homologs and tested their inhibition by BO.

With this approach, we localized a domain required for inhibition within a 52 amino acid region of Lye. While the data have not confirmed this section to be sufficient to confer a Lye homolog the ability to be inhibited by BO, the results indicate that the 69-120 region of *H. salinarum* Lye is necessary for this interaction to occur. Future studies will test a fusion protein consisting of *H. volcanii* Lye with the identified 52 amino acid domain from the *H. salinarum* homolog. If this construct is inhibited by BO, the inhibitory interaction can be confidently localized to the identified region. Moreover, to increase the resolution of our search for the amino acids involved in this inhibitory interaction, we plan on performing gene shuffling experiments with the two Lye homologs.

The structure prediction studies revealed interesting features of the identified region of interest. Firstly, the 52 amino acid section presents high structural similarity between the two Lye homologs, indicating that the region required for inhibition may be further narrowed down. Secondly, this domain of interest seems to include the only significant section of disorder in each of the proteins. Disordered domains are sections of proteins that lack a fixed structure, but which are often functionally important. Lastly, the identified region spans both part of the third TM helix and the cytoplasmic side, raising the possibility that the inhibition takes places at the interface between the lipid bilayer and the cytoplasm, which is further strengthened by the fact that BO is an integral membrane protein.

Taken together, the fusion protein analysis along with the structure prediction results suggest that the amino acids involved in *H. salinarum* Lye's inhibition by BO reside within the 69-120 domain. Nevertheless, it might be the case that the predicted model is somewhat different than the 3D structure of the protein and that the interaction does not require specific amino acid residues, but rather a certain overall tertiary structure. Such limitations are inherent to our experimental design and weaken the confidence of our results, but they will be addressed in the future Lye shuffling experiments.

SOB and gene shuffling

Gene shuffling is a high-throughput method for generating a library of hybrid genes. The key parameter that determines the diversity and size of the library is the similarity of the starting genes. In order to increase the efficiency of our shuffling, we wrote a script for backtranslating two aligned protein sequences to high similarity DNA sequences. The tested peptides (Figure 6) confirmed the ability of our algorithm to find matching sections ≥ 3 bp and extend them in both directions by optimizing codon usage for higher similarity. However, despite its proper functioning, we are aware of the many limitations SOB has as a potential bioinformatics tool.

One such limitation is the limited accessibility, evident in the lack of a user-friendly interface and in the input requirement for an alignment of the two proteins in a specific format. These features can be improved by having the script read multiple file formats, and perform its own protein alignment using an online tool like BLASTP[61]. Furthermore, currently, SOB is not guaranteed to output the two most similar DNA sequences given any two proteins. At the expense of computational time, the script could start the algorithm from any position on the alignment not just from the first codon, and it could also search for any codon changes that increase sequence similarity, not just those that extend previous regions of identity. Such modifications, along with expanding the backtranslation process to include all possible codon usage rankings, would significantly increase the probability that the algorithm finds the global minimum (the two DNA sequences of highest similarity) as opposed to a local minimum.

Nevertheless, despite the limitations of our script, SOB has proved effective in generating a diverse library of fusion proteins (Supplementary Table 1). Based on the frequency of the sequenced shuffled opsins, it is clear that, at least for the proteins that inhibited Lye, there was a bias towards a BO N-terminus or a CO middle section (Figures 9,10). We believe that these results can be explained by the high similarity terminal regions of the two proteins, which had a higher probability of cross-over than the rest of the peptide.

Bacterioopsin and cruxopsin

In the cell, bacterioopsin and cruxopsin covalently bind retinal to form the light-driven proton pumps BR and cR, respectively, which convert light energy to chemical energy. Apart from this important function, the two opsins also play a role in retinal biosynthesis by inhibiting Lye, the enzyme that catalyzes the committed step in bacterioruberin synthesis. This interaction represents a recently described regulatory mechanism aimed at maintaining the appropriate ratio of apoprotein to cofactor. Using gene shuffling, we studied the protein regions, from both BO and CO, involved in this observed interaction with Lye. However, the inhibition screen was performed using solely colony color, and the bacterioruberin levels were not yet confirmed by the colorimetric assay used in the previous experiments. As such, these conclusions are less convincing than the ones regarding the Lye domains required for BO mediated inhibition.

Our results localized the BO amino acids required for inhibiting *H. salinarum* Lye to V115 and D116. The corresponding CO amino acids are A115 and G116, which suggests that the studied interaction might be ionic in nature. More interestingly, these two amino acids were localized in a cytoplasmic inter-TM region, which strengthens the hypothesis that the inhibitory interaction occurs at the interface between the membrane and the cytoplasm. Nevertheless, our data did not validate the two amino acids as the only ones necessary for inhibition, a hypothesis that remains to be tested in future studies.

The screen for *H. vallismortis* Lye inhibiting opsins was not as successful in localizing the necessary region for interaction as the *H. salinarum* Lye screen, since it narrowed down the search to two regions of 34 and 43 amino acids, as opposed to only one. This difference was likely due to the lower number of screened colonies for *H. vallismortis* Lye inhibition, compared to that for the inhibition of the *H. salinarum* homolog. Nonetheless, these domains exhibited high structural similarity between the two opsins, and also included cytoplasmic inter-TM regions.

It is worth mentioning that all the identified opsin and Lye protein regions required for the inhibitory interaction are situated within the same general location, in the cytoplasmic side of the protein between two TM regions (Figure 13). This suggests that the opsin and Lye proteins could come in close proximity to each other and have a direct inhibitory interaction leading to bacterioruberin biosynthesis inhibition and higher levels or retinal. Future research could test this hypothesis through co-immunoprecipitation and crystallization experiments.

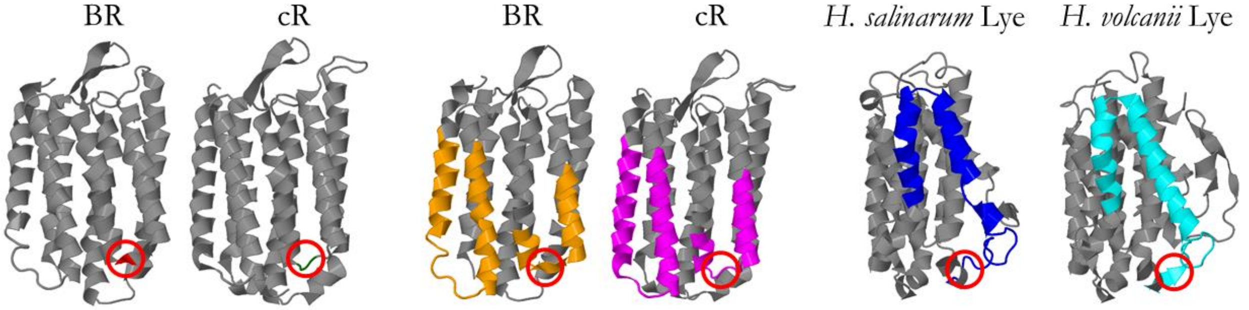

Figure 13. Structure comparison between the opsins and Lye homologs. The colored regions represent the identified sections required for inhibition (left – *H. salinarum* Lye inhibition; middle – *H. vallismortis* Lye inhibition; right – BO inhibition), while the red circles show the potential interaction sites. Proteins are depicted within the membrane and oriented with the top side towards the extracellular environment.

Conclusions

Overall, our analysis identified specific domains of the studied proteins that were required for the opsin-Lye inhibitory interaction. Future directions will focus on determining the exact protein regions responsible for this regulatory mechanism. With the information provided in this study, fusion proteins can be specifically designed to test the identified regions, and amino acid substitutions could more precisely localize the interacting amino acids. Lastly, a bioinformatics study can be performed to identify similar domains in other biological systems and potentially elucidate the mechanisms responsible for the regulation of apoprotein-cofactor stoichiometry.

Acknowledgments

I would like to thank Professor Ronald Peck for his unwavering support and guidance throughout my research experience. Since the spring of my first year at Colby College, he has been a great mentor and has helped me develop as a scientist.

Professors Russell Johnson and Kevin Rice have been great advisors and teachers, and I would like to thank them for the helpful discussions and feedback regarding my honors thesis.

I am also indebted to Serena Graham and Professor David Angelini for their contributions to method development and data collection. Many members of the Peck lab have also provided technical assistance and helpful feedback including Emily Shaw, Adam Lavertu, Katie Metayer, Erika Smith, Tevis Vitale, Ray Nakada, and Christine Wamsley.

Lastly, I would like to thank my parents, Cristian and Cornelia, my brother, Andrei, and my friends for their constant encouragement and support.

The research reported in this project was supported by an Institutional Development Award from the National Institute of General Medical Sciences (NIGMS) of the National Institutes of Health under grant number P20GM103423. Additional funding was provided by the Colby College Natural Science Division and award 5GM094725-02 from NIGMS to Ronald F. Peck.

References

1. Yoshizawa, S. *et al.* Functional characterization of flavobacteria rhodopsins reveals a unique class of light-driven chloride pump in bacteria. *Proc. Natl. Acad. Sci. U. S. A.* **111,** 6732–7 (2014).

2. Larusso, N. D., Ruttenberg, B. E., Singh, A. K. & Oakley, T. H. Type II Opsins: Evolutionary Origin by Internal Domain Duplication? *J. Mol. Evol.* **66,** 417–423 (2008).

3. Spudich, J. L., Yang, C.-S., Jung, K.-H. & Spudich, E. N. Retinylidene Proteins: Structures and Functions from Archaea to Humans. *Annu. Rev. Cell Dev. Biol.* **16,** 365–392 (2000).

4. Peirson, S. N., Halford, S. & Foster, R. G. The evolution of irradiance detection: melanopsin and the non-visual opsins. *Philos. Trans. R. Soc. B Biol. Sci.* **364,** 2849–2865 (2009).

5. Shichida, Y. & Matsuyama, T. Evolution of opsins and phototransduction. *Philos. Trans. R. Soc. B Biol. Sci.* **364,** 2881–2895 (2009).

6. Plachetzki, D. C., Degnan, B. M. & Oakley, T. H. The Origins of Novel Protein Interactions during Animal Opsin Evolution. *PLoS One* **2,** e1054 (2007).

7. Terakita, A. The opsins. *Genome Biol.* **6,** 213 (2005).

8. Grote, M. & O'Malley, M. A. Enlightening the life sciences: the history of halobacterial and microbial rhodopsin research. *FEMS Microbiol. Rev.* **35,** 1082–1099 (2011).

9. Luecke, H., Schobert, B., Richter, H. T., Cartailler, J. P. & Lanyi, J. K. Structure of bacteriorhodopsin at 1.55 A resolution. *J. Mol. Biol.* **291,** 899–911 (1999).

10. Luecke, H., Schobert, B., Richter, H. T., Cartailler, J. P. & Lanyi, J. K. Structural changes in bacteriorhodopsin during ion transport at 2 angstrom resolution. *Science (80-.).* **286,** 255–261 (1999).

11. Hoff, W. D., Jung, K. H. & Spudich, J. L. Molecular mechanism of photosignaling by archaeal sensory rhodopsins. *Annu. Rev. Biophys. Biomol. Struct.* **26,** 223–258 (1997).

12. Hartmann, R., Sickinger, H. D. & Oesterhelt, D. Anaerobic growth of halobacteria. *Proc. Natl. Acad. Sci. U. S. A.* **77,** 3821–5 (1980).

13. Kitajima, T. *et al.* Novel bacterial rhodopsins from *Haloarcula vallismortis*. *Biochem. Biophys. Res. Commun.* **220,** 341–345 (1996).

14. Becker, E. A. *et al.* A large and phylogenetically diverse class of type 1 opsins lacking a canonical retinal binding site. *PLoS One* **11,** (2016).

15. Chan, S. K. *et al.* Crystal structure of cruxrhodopsin-3 from *Haloarcula vallismortis*. *PLoS One* **9,** (2014).

16. Trivedi, S., Prakash Choudhary, O. & Gharu, J. Different Proposed Applications of Bacteriorhodopsin. *Recent Pat. DNA Gene Seq.* **5,** 35–40 (2011).

17. Saeedi, P. *et al.* Potential applications of bacteriorhodopsin mutants. *Bioengineered* **3,** 326–328 (2012).

18. Wagner, N. L., Greco, J. A., Ranaghan, M. J. & Birge, R. R. Directed evolution of bacteriorhodopsin for applications in bioelectronics. *J. R. Soc. Interface* **10,** 20130197–20130197 (2013).

19. Deisseroth, K. Optogenetics. *Nat. Methods* **8,** 26–29 (2011).

20. Berndt, A., Yizhar, O., Gunaydin, L. A., Hegemann, P. & Deisseroth, K. Bi-stable neural state switches. *Nat. Neurosci.* **12,** 229–234 (2009).

21. Hellingwerf, K. J., Arents, J. C., Scholte, B. J. & Westerhoff, H. V. Bacteriorhodopsin in liposomes. II. Experimental evidence in support of a theoretical model. *Biochim. Biophys. Acta* **547,** 561–82 (1979).

22. Wang, W. W., Knopf, G. K. & Bassi, A. S. Photoelectric properties of a detector based on dried bacteriorhodopsin film. *Biosens. Bioelectron.* **21,** 1309–1319 (2006).

23. Hong, F. T. & Mauzerall, D. Interfacial Photoreactions and Chemical Capacitance in Lipid Bilayers. *Proc. Natl. Acad. Sci.* **71,** 1564–1568 (1974).

24. Imhof, M., Rhinow, D. & Hampp, N. Two-photon polarization data storage in bacteriorhodopsin films and its potential use in security applications. *Appl. Phys. Lett.* **104,** 81921 (2014).

25. Greco, J. A., Wagner, N. L., Ranaghan, M. J., Rajasekaran, S. & Birge, R. R. in *Biomolecular Information Processing* 33–59 (Wiley-VCH Verlag GmbH & Co. KGaA, 2012). doi:10.1002/9783527645480.ch3

26. Frydrych, M., Silfsten, P., Parkkinen, S., Parkkinen, J. & Jaaskelainen, T. Color sensitive retina based on bacteriorhodopsin. *Biosystems.* **54,** 131–40 (2000).

27. Tukiainen, T., Lensu, L. & Parkkinen, J. in *Advances in Brain, Vision, and Artificial Intelligence* 94–103 (Springer Berlin Heidelberg). doi:10.1007/978-3-540-75555-5_10

28. Chuong, A. S. *et al.* Noninvasive optical inhibition with a red-shifted microbial rhodopsin. *Nat. Neurosci.* **17,** 1123–1129 (2014).

29. Chow, B. Y. *et al.* High-performance genetically targetable optical neural silencing by light-driven proton pumps. *Nature* **463,** 98–102 (2010).

30. Moran, N. A. & Jarvik, T. Lateral Transfer of Genes from Fungi Underlies Carotenoid Production in Aphids. *Science (80-.).* **328,** 624–627 (2010).

31. Altincicek, B., Kovacs, J. L. & Gerardo, N. M. Horizontally transferred fungal carotenoid genes in the two-spotted spider mite *Tetranychus urticae*. *Biol. Lett.* **8,** 253–257 (2012).

32. Shahmohammadi, H. R. *et al.* Protective roles of bacterioruberin and intracellular KCl in the resistance of *Halobacterium salinarium* against DNA-damaging agents. *J. Radiat. Res.* **39,** 251–262 (1998).

33. Holt, N. E. *et al.* Carotenoid cation formation and the regulation of photosynthetic light

harvesting. *Science (80-.).* **307,** 433–436 (2005).

34. Miller, N. J., Sampson, J., Candeias, L. P., Bramley, P. M. & Rice-Evans, C. A. Antioxidant activities of carotenes and xanthophylls. *FEBS Lett.* **384,** 240–242 (1996).

35. Lazrak, T. *et al.* Comparison of the effects of inserted C40- and C50-terminally dihydroxylated carotenoids on the mechanical properties of various phospholipid vesicles. *Biochim. Biophys. Acta* **903,** 132–41 (1987).

36. Dundas, I. D. & Larsen, H. A study on the killing by light of photosensitized cells of *Halobacterium salinarum*. *Arch. Mikrobiol.* **46,** 19–28 (1963).

37. Naguib, Y. M. A. Antioxidant activities of astaxanthin and related carotenoids. *J. Agric. Food Chem.* **48,** 1150–1154 (2000).

38. El-Sayed, W. S. M. *et al.* Effects of Light and Low Oxygen Tension on Pigment Biosynthesis in *Halobacterium salinarum*, Revealed by a Novel Method to Quantify Both Retinal and Carotenoids. *Plant Cell Physiol* **43,** 379–383 (2002).

39. de la Vega, M., Sayago, A., Ariza, J., Barneto, A. G. & León, R. Characterization of a bacterioruberin-producing Haloarchaea isolated from the marshlands of the Odiel river in the southwest of Spain. *Biotechnol. Prog.* **32,** 592–600 (2016).

40. Rodrigo-Banos, M., Garbayo, I., Vilchez, C., Bonete, M. J. & Martinez-Espinosa, R. M. Carotenoids from Haloarchaea and their potential in biotechnology. *Marine Drugs* **13,** 5508–5532 (2015).

41. Lazrak, T. *et al.* Bacterioruberins reinforce reconstituted *Halobacterium* lipid membranes. *BBA - Biomembr.* **939,** 160–162 (1988).

42. Abbes, M. *et al.* Biological properties of carotenoids extracted from *Halobacterium halobium* isolated from a Tunisian solar saltern. *BMC Complement. Altern. Med.* **13,** 255 (2013).

43. Yatsunami, R. *et al.* Identification of carotenoids from the extremely halophilic archaeon *Haloarcula japonica*. *Front. Microbiol.* **5,** (2014).

44. Dummer, A. M. *et al.* Bacteriopsin-Mediated Regulation of Bacterioruberin Biosynthesis in *Halobacterium salinarum*. *J. Bacteriol.* **193,** 5658–5667 (2011).

45. Dombeck, T. A. & Satonik, R. C. The Porphyrias. *Emerg. Med. Clin. North Am.* **23,** 885–899 (2005).

46. Hartong, D. T., Berson, E. L. & Dryja, T. P. Retinitis pigmentosa. *Lancet (London, England)* **368,** 1795–809 (2006).

47. Sumper, M. & Herrmann, G. Biosynthesis of purple membrane: control of retinal synthesis by bacterio-opsin. *FEBS Lett.* **71,** 333–336 (1976).

48. Peck, R. F., Johnson, E. A. & Krebs, M. P. Identification of a Lycopene -Cyclase Required for Bacteriorhodopsin Biogenesis in the Archaeon *Halobacterium salinarum*. *J. Bacteriol.* **184,** 2889–2897 (2002).

49. Oesterhelt, D. & Stoeckenius, W. Rhodopsin-like Protein from the Purple Membrane of *Halobacterium halobium*. *Nat. New Biol.* **233,** 149–152 (1971).

50. Ni, B. F., Chang, M., Duschl, A., Lanyi, J. & Needleman, R. An efficient system for the synthesis of bacteriorhodopsin in *Halobacterium halobium*. *Gene* **90,** 169–72 (1990).

51. Allers, T., Barak, S., Liddell, S., Wardell, K. & Mevarech, M. Improved strains and plasmid vectors for conditional overexpression of His-tagged proteins in *Haloferax volcanii*. *Appl. Environ. Microbiol.* **76,** 1759–69 (2010).

52. Cline, S. W. & Doolittle, W. F. Efficient transfection of the archaebacterium *Halobacterium halobium*. *J. Bacteriol.* **169,** 1341–4 (1987).

53. Allers, T., Ngo, H.-P., Mevarech, M. & Lloyd, R. G. Development of additional selectable markers for the halophilic archaeon *Haloferax volcanii* based on the leuB and trpA genes. *Appl. Environ. Microbiol.* **70,** 943–53 (2004).

54. Ng, W. V *et al.* Genome sequence of *Halobacterium* species NRC-1. *Proc. Natl. Acad. Sci. U. S. A.* **97,** 12176–81 (2000).

55. Hartman, A. L. *et al.* The Complete Genome Sequence of *Haloferax volcanii* DS2, a Model Archaeon. *PLoS One* **5,** e9605 (2010).

56. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. E. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **10,** 845–858 (2015).

57. Maiti, R., Van Domselaar, G. H., Zhang, H. & Wishart, D. S. SuperPose: a simple server for sophisticated structural superposition. *Nucleic Acids Res.* **32,** W590–W594 (2004).

58. Meyer, A. J., Ellefson, J. W. & Ellington, A. D. in *Current Protocols in Molecular Biology* 15.12.1-15.12.7 (John Wiley & Sons, Inc., 2014). doi:10.1002/0471142727.mb1512s105

59. Rice, P., Longden, I. & Bleasby, A. EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet.* **16,** 276–277 (2000).

60. Pebay-Peyroula, E., Rummel, G., Rosenbusch, J. P. & Landau, E. M. X-ray structure of bacteriorhodopsin at 2.5 angstroms from microcrystals grown in lipidic cubic phases. *Science* **277,** 1676–81 (1997).

61. Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25,** 3389–402 (1997).

Supplementary Information

Supplementary Table 1. Structural composition of the shuffled genes. S1-S5 are the sequences displayed in Figure 9, while SU1- SU3 are unused sequences. V1-V3 are the sequences displayed in Figure 10, while UV1 is an unused sequence. In each of the two screens, one of the sequencing reactions did not work.

Id	BO aa	CO aa	Frequency	Comments
S1	1-263	264-268	25%	
S2	1-252	256-268	16.7%	3 amino acid gap
S3	1-5; 114-117; 126-187;	118-125	5%	Frameshift after 187, contains the following non-opsin amino acids: RPSRPSATSSPSSGRPTRSCGSSAP RAPASSPSTSRPCSSWSSTSARWR SASASSSSARARSSARRRRRSRRP ATARPRPRTENSCSPGDPLVLER
S4	1-107; 115-117; 210-239	108-114; 118-209; 240-268	1.7%	More pink than others, but still cream-colored
S5	32-76; 81-96; 108-191; 203-224; 261-268	6-31; 77-80; 97-107; 192-202; 225-260	1.7%	
SU1	1-268		45%	
SU1	1-260	261-268	1.7%	
SU3		6-73	1.7%	Inexplicable cream color phenotype. Probably due to an unknown mutation.
V1	1-26; 246-268	27-245	73.3%	
V2	158-268	6-117; 119-157	6.7%	Position 118 is W (Q in BO, R in CO)
V3	46; 71-101	6-45; 47-69; 102-135	6.7%	
VU1	77-117; 253-263	6-76; 118-213; 218-253; 264-268	6.7%	Very pink. Insert of ADADHDA between positions 213 and 218; aa 253 counted twice because protein has D (253 in CO) and E (253 in BO)

Supplementary Data 1. SOB source code

```
#Author: Alex Pleşa
#Project: Gene shuffling
#Date: 9/15/16


###Read alignment
alignment <- read.csv(file="Alignment.csv",head=TRUE,sep=",")
alignment <- data.frame(lapply(alignment, as.character), stringsAsFactors=FALSE)


###Read genetic code
genCode <- read.csv(file="genCode.csv",head=TRUE,sep=",")
genCode <- data.frame(lapply(genCode, as.character), stringsAsFactors=FALSE)


###Create data frames for all sequences
gene1_seq<- matrix(,nrow=1,ncol=3*length(alignment[,1]))
gene2_seq<- matrix(,nrow=1,ncol=3*length(alignment[,2]))


###Generate first DNA sequence
variant_g1<-""
for(i in 1:length(alignment[,1])){
  if(is.na(alignment[i,1])){
    variant_g1 = paste(variant_g1,"   ",sep="")
  }
  else{
    variant_g1 = paste(variant_g1,genCode[2,grep(alignment[i,1],colnames(genCode))],sep="")
  }
}
for(i in 1:nchar(variant_g1)){
  gene1_seq[i]=substr(variant_g1,i,i)
}
colnames(gene1_seq)=1:as.numeric((3*length(alignment[,1])))


###Generate second DNA sequence
variant_g2<-""
for(i in 1:length(alignment[,2])){
  if(is.na(alignment[i,2])){
    variant_g2 = paste(variant_g2,"   ",sep="")
  }
  else{
    variant_g2 = paste(variant_g2,genCode[2,grep(alignment[i,2],colnames(genCode))],sep="")
  }
}
for(i in 1:nchar(variant_g2)){
  gene2_seq[i]=substr(variant_g2,i,i)
}
colnames(gene2_seq)=1:as.numeric((3*length(alignment[,2])))
```

```
###Similarize the sequences
iteration=0; #intialize variable that counts the iterations
repeat{ #repeat this until nothing changes anymore or it has hit the max iterations
  stop=0;#nothing changed
  start=0;
  end=0;
  iteration=iteration+1;

  for(i in 1:length(gene1_seq)){ #loop through the genes

    if(gene1_seq[i]==gene2_seq[i]){ #if the nucleotides are the same

      if(start==0){ #if start is not set, set it here
        start=i;
      }
      end=i; #make this the end of the island
    }
    else {
      if((end-start)>=2){ #if length is larger than 3 bp
        changed=change(gene1_seq,gene2_seq,start,end);#see if you can exchange any codons to increase similarity
        if(changed[1]!=0){
          gene1_seq[changed[2]]=substr(genCode[changed[5]+1,changed[3]],1,1);
          gene1_seq[changed[2]+1]=substr(genCode[changed[5]+1,changed[3]],2,2);
          gene1_seq[changed[2]+2]=substr(genCode[changed[5]+1,changed[3]],3,3);
          gene2_seq[changed[2]]=substr(genCode[changed[6]+1,changed[4]],1,1);
          gene2_seq[changed[2]+1]=substr(genCode[changed[6]+1,changed[4]],2,2);
          gene2_seq[changed[2]+2]=substr(genCode[changed[6]+1,changed[4]],3,3);
          stop=1;#something changed
        }

        if(changed[7]!=0){
          gene1_seq[changed[8]]=substr(genCode[changed[11]+1,changed[9]],1,1);
          gene1_seq[changed[8]+1]=substr(genCode[changed[11]+1,changed[9]],2,2);
          gene1_seq[changed[8]+2]=substr(genCode[changed[11]+1,changed[9]],3,3);
          gene2_seq[changed[8]]=substr(genCode[changed[12]+1,changed[10]],1,1);
          gene2_seq[changed[8]+1]=substr(genCode[changed[12]+1,changed[10]],2,2);
          gene2_seq[changed[8]+2]=substr(genCode[changed[12]+1,changed[10]],3,3);
          stop=1;#something changed
        }
      }
      start = 0; #set start and end to 0
      end = 0;
    }

    if((i==end)&&(end==length(gene1_seq))&&(start!=1)){ #special case for C terminus
      changed=change(gene1_seq,gene2_seq,start,end);#see if you can exchange any codons to increase similarity
      if(changed[1]!=0){
        gene1_seq[changed[2]]=substr(genCode[changed[5]+1,changed[3]],1,1);
        gene1_seq[changed[2]+1]=substr(genCode[changed[5]+1,changed[3]],2,2);
        gene1_seq[changed[2]+2]=substr(genCode[changed[5]+1,changed[3]],3,3);
        gene2_seq[changed[2]]=substr(genCode[changed[6]+1,changed[4]],1,1);
        gene2_seq[changed[2]+1]=substr(genCode[changed[6]+1,changed[4]],2,2);
        gene2_seq[changed[2]+2]=substr(genCode[changed[6]+1,changed[4]],3,3);
        stop=1;#something changed
      }
    }
  }

  if((stop==0)||(iteration==1000)){#if nothing changed, don't repeat
    break
  }
}


###Function that finds alternative codon pairings to increase sequence identity
change <- function(gene1, gene2, start, end){
  s_c = ceiling((start-1)/3)*3-2; #get the position of the codon before the island
  s_c1 = paste(gene1[s_c],gene1[s_c+1],gene1[s_c+2],sep=""); #get gene 1 first mismatched codon before island
  s_c2 = paste(gene2[s_c],gene2[s_c+1],gene2[s_c+2],sep=""); #get gene 2 first mismatched codon before island
  s_a1_col = grep(s_c1,genCode); #get column of AA of gene1 of first mismatched codon before island
  s_a2_col = grep(s_c2,genCode); #get column of AA of gene2 of first mismatched codon before island

  e_c = ceiling((end+1)/3)*3-2; #get the position of the codon after the island
  e_c1 = paste(gene1[e_c],gene1[e_c+1],gene1[e_c+2],sep=""); #get gene 1 first mismatched codon after island
  e_c2 = paste(gene2[e_c],gene2[e_c+1],gene2[e_c+2],sep=""); #get gene 2 first mismatched codon after island
  e_a1_col = grep(e_c1,genCode); #get column of AA of gene1 of first mismatched codon after island
  e_a2_col = grep(e_c2,genCode); #get column of AA of gene2 of first mismatched codon after island

  #variable that signals if anything has changed and stores values for changing codons
  changed <- c(0,s_c,s_a1_col,s_a2_col,0,0,0,e_c,e_a1_col,e_a2_col,0,0);

  #Find most similar codons in the 2 columns for the 1st mismatched codon before island
  if(start!=1){#if the start of the island is not on the first bp, check on the left
    score <- matrix(0L,nrow=as.numeric(genCode[1,s_a1_col]),ncol=as.numeric(genCode[1,s_a2_col]));#score
matrix for all codon combinations
    for(i in 1:as.numeric(genCode[1,s_a1_col])){ #loop through first AA codon posibilities
      c1=genCode[i+1,s_a1_col];
      for(j in 1:as.numeric(genCode[1,s_a2_col])){ #loop through second AA codon posibilities
        c2=genCode[j+1,s_a2_col];
        score[i,j]=get_score(1,c1,c2);
      }
    }

    if(max(score)>get_score(1,s_c1,s_c2)){ #replace codons with 1st pair if more similar than previous pair
      inds = which(score==max(score), arr.ind=TRUE);
      changed[1]=1; #there is an alternative
      changed[5]=inds[1,1]; #codon position for gene 1
      changed[6]=inds[1,2]; #codon position for gene 2
    }
  }

  #Find most similar codons in the 2 columns for the 1st mismatched codon after island
  if(end!=length(gene1)){
    #score matrix for all codon combinations
    score <- matrix(0L,nrow=as.numeric(genCode[1,e_a1_col]),ncol=as.numeric(genCode[1,e_a2_col]));
    for(i in 1:as.numeric(genCode[1,e_a1_col])){ #loop through first AA codon posibilities
      c1=genCode[i+1,e_a1_col];
      for(j in 1:as.numeric(genCode[1,e_a2_col])){ #loop through second AA codon posibilities
        c2=genCode[j+1,e_a2_col];
        score[i,j]=get_score(2,c1,c2);
      }
  }

  if(max(score)>get_score(2,e_c1,e_c2)){ #replace codons with 1st pair if more similar than previous pair
   inds = which(score==max(score), arr.ind=TRUE);
   changed[7]=1; #there is an alternative
   changed[11]=inds[1,1]; #codon position for gene 1
   changed[12]=inds[1,2]; #codon position for gene 2
  }
 }

 return (changed);
}


###Function that returns the similarity score of 2 codons
get_score <- function(type,c1,c2){
 s = 0;

 if(type==1){
  for(k in 3:1){ #loop through codon positions from righ to left
   if(substr(c1,k,k)==substr(c2,k,k)){ #if they match increase score by k
    s=s+k*k;#scoring scheme is 3rd nucletodies match=9; 2nd nucleotides match=4, 1st nucleotides match=1
   }
  }
 }

 else{
  for(k in 1:3){ #loop through codon positions from left to right
   if(substr(c1,k,k)==substr(c2,k,k)){ #if they match increase score by k
    s=s+9/k/k;#scoring scheme is 1st nucletodies match=9; 2nd nucleotides match=9/4, 3rd nucleotides match=1
   }
  }
 }

 return (s);
}


#Write the two genes to a .txt file
sink("output.txt")
cat(as.character(colnames(alignment[1])))
cat(">\n")
cat(paste(gene1_seq,sep="",collapse=""))
cat("\n\n")
cat(as.character(colnames(alignment[2])))
cat(">\n")
cat(paste(gene2_seq,sep="",collapse=""))
sink()
```