Experimentally Defining the Physiology of Freshwater Ultramicrobacteria: acI *Actinobacterial* Light Utilization and Peptide Degradation

By

Jeffrey Dwulit-Smith

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Biophysics)

at the

UNIVERSITY OF WISCONSIN-MADISON

2019

Date of final oral examination: 01/18/2019

The dissertation is approved by the following members of the Final Oral Committee:

Katrina Forest, Professor, Bacteriology/Biophysics

Katherine McMahon, Professor, Bacteriology/Civil and Environmental Engineering

Baron Chanda, Professor, Neuroscience/Biophysics

Alessandro Senes, Associate Professor, Biochemistry/Biophysics

Thomas Brunold, Professor, Chemistry/Biophysics

Acknowledgements

This work would not have been possible without my advisor, Dr. Katrina Forest. I also thank my committee members, Dr. Trina McMahon, Dr. Baron Chanda, Dr. Alessandro Senes, and Dr. Thomas Brunold for helpful guidance throughout my research and key past and present members of the Forest laboratory, Dr. Anna Baker, Dr. Shyamosree Bhattacharya, Katiria Gonzalez-Rivera, Peter Newhouse, Nick Koranda, and Neydis Moreno. They have made my time in lab much more fun and enjoyable. I would like to thank my mother, Diane Dwulit, for her love and support throughout my life. I would not have been able to get to where I am without her, and I owe everything to her.

Table of Contents

List of Figures and Tables

Thesis Abstract

Freshwater is an imperiled finite resource because of constant threats from contamination and climate change. These forces expose freshwater microbiomes to pressures that jeopardize their survival. Consequently, we cannot predict the reactions of bacteria governing the water quality. For example, acI *Actinobacteria* numerically dominate freshwater lakes, but the adaptations that they employ to achieve this success have not been experimentally demonstrated. This work describes three physiological systems in acI *Actinobacteria*: 1) a retinal biosynthesis pathway that provides the chromophore for a functionally-characterized actinorhodopsin and, possibly, for a heliorhodopsin, 2) a complex carotenoid biosynthesis pathway that starts with retinal precursors in all acI clades, and 3) an enzyme that can cleave aspartate-leading dipeptides to free amino acids. All three systems were uncovered by examining acI single-cell genomes and metagenomes and detailing the pathways with chemical structures. The first was tested by heterologous production of enzymes in *Escherichia coli* and analysis of resulting chromophores by HPLC-MS, MS/MS, and UV-Vis spectroscopy. Work confirmed that lycopene and retinal are produced and that retinal binds actino-opsin to form actinorhodopsin. Proton-pumping assays confirmed that the holoprotein creates a proton gradient in response to light. Metatranscriptomics showed that all identified pathway genes for the first and second systems were expressed in environmental acI, with actino-opsin transcripts hundreds of times greater than other transcripts. The third system was characterized by cyanophycin and aspartate-p-nitroaniline cleavage assays and X-ray protein crystallography. The enzyme cleaved the dipeptide ligand to free amino acids but could not cleave cyanophycin polymers. The 1.93 Å resolution structure displayed a canonical serine-histidine-glutamate catalytic triad. Although experimental progress on the physiology of acI has been made, work remains for all systems: 1) characterizing the assembly and function of a heliorhodopsin, 2) characterizing the structure(s) of complex carotenoid(s), and 3) characterizing other potential enzyme ligands such as degraded cyanophycin. Without accurate knowledge of

these and other acI *Actinobacterial* adaptations, bacterial responses to ecosystem imbalances cannot be predicted, and the information cannot be incorporated into climate change models.

Chapter I – Introduction

Freshwater Microbiomes and Ecology

The Earth has less than 0.007% of its freshwater in lakes with most of the remaining sources tied up in glaciers, ice caps, and deep-ground water (1). However, farming and industrialization have consistently threatened these key resources with nutrient runoff, and climate change has altered weather patterns that the ecosystems evolved with. Most currently, infusions of nitrogen- and phosphorus-containing chemicals into freshwater lakes lead to cyanobacterial overgrowth and homeostatic imbalance (2, 3). As rampant growth begins, thick algal mats cover the lake. As the algae count increases, the cells deplete dissolved oxygen, disrupt natural energy cycling, and block light from penetrating into the lake (4, 5). Sudden habitat shifts lead to dark anoxic dead zones as bacteria, invertebrates, and fish struggle to adapt. Humans experience changes immediately in terms of recreational and fishing industry limitations and also more long-term in the form of liver and brain damaging cyanotoxins that survive water treatment and accrue within the potable water system (6, 7). Obviously then, understanding the physiologies of the predominant species in a freshwater lake, not just algae, is of great importance. A better description of the ecological web can then help us mitigate the impacts of climate change by enabling the creation of better ecosystem-response models.

Of course, a single freshwater lake harbors hundreds of species at a certain time, but microbes are a defining pervasive ecosystem level force. However, because microbiomes are extremely diverse, it can be difficult to determine the impact of any given species on the ecosystem. With the rise of massive metagenomic sequencing projects, entire biomes are now sampled, sequenced, and scrutinized to determine what species are most prevalent, most often by genome assembly and homology analysis (8–11). This step is critical in determining the bacteria that can impact the ecosystem on a large-scale and that are not necessarily easily visible even under microscopes. This type of investigation has revealed the presence of five major bacterial classifications in lakes.

acI *Actinobacteria*

Freshwater lakes contain five commonly found types of bacteria, *Actinobacteria, Proteobacteria, Bacteroidetes, Cyanobacteria, and Verrucomicrobia* (10). Remaining groups make up only 2.6% of the total sequences in genomic databases, and the most abundant group is by far the *Actinobacteria*. Within the *Actinobacteria*, 16S rRNA genes from freshwater lakes were distinct from the "typical" soiling-dwelling and marine organisms. As a result of phylogenetic comparisons and clustering, the freshwater group has been divided into the acI, acII, acIII, and acIV designations (12). Within these subdivisions are the even more finely grained distinctions of acI-A to -C, acII-A to -D, and acIV-A and -B, and beyond the lettering are divisions by numbers. It should be noted that these rough parallels with Linnaean taxonomy will be the primary way of discussing organisms throughout (i.e. for the *Actinobacterium* designated acI-B1, acI is the lineage (family), B is the clade (genus), and 1 is the tribe (species).

One of the most prevalent groups of *Actinobacteria* in freshwater lakes is acI *Actinobacteria*. This lineage is ubiquitous and abundant across freshwater ecosystems including freshwater lakes like Lake Mendota in Madison, Wisconsin (10). The group is comprised of gram-positive ultramicrobacteria that are overwhelmingly abundant, sometimes reaching over 50% of all bacteria (12). acI *Actinobacteria* are some of the physically smallest free-living organisms in the lake at just 0.1μm^3, and their low-GC genomes are streamlined to 1.2 to 1.5 MBp (8–11, 13). For comparison, an acI *Actinobacterial* cell is physically about seven-times smaller than the average *Escherichia coli* with a genome of about 5 MBp. Presumably, acI genomes encode a bare minimum of heterotrophic machinery for many metabolic pathways and likely are auxotrophic for many compounds (9, 14). They may even be considered saprovores in that they salvage many of their resources from free-floating small molecules and breakdown products. Genes that likely encode amino acid and sugar membrane importers and internal degradative enzymes dot their genomes. Even though much is known about their genome composition via annotations, little experimental evidence for acI physiology exists. This is because acI are not currently axenically

culturable (11, 15), although this hurdle may soon be overcome (16). Nevertheless, sequencing of the acI lineage at both the single cell and metagenomic levels has allowed them to be classified into three clades (acI-A, acI-B, acI-C) (9). Within these clades are further distinctions into tribes (e.g. acI-AI or acI-BI).

Although streamlined genomes with key primary metabolic cassettes explain some of the acI *Actinobacterial* competitive advantage and allow classification, empilimnal dominance might also require secondary metabolic systems that are not canonically considered part of primary metabolism. Here, I propose three major advantages including: supplementation of energy needs via actinorhodopsin-based photoheterotrophy, carotenoid biosynthesis, and aspartate dipeptide salvage. Together, these three adaptations go far in describing the previously experimentally characterized acI *Actinobacteria*.

Actinorhodopsins

All members of the acI *Actinobacteria* are proposed to bolster heterotrophic growth using phototrophy because their genomes encode actino-opsins (*actR)* (9, 14, 17, 18). Light utilization is feasible because ActR is the putative platform protein for actinorhodopsin (holo-ActR). Rhodopsins are seven-transmembrane-helix bundles with N-terminus-out and C-terminus-in topology. These proteins form holoproteins because they bind retinal at an internal lysine via a Schiff base (19). The retinal-binding pocket, specifically a tuning residue sidechain, determines the absorption profile of the holoprotein (20). Generally, these proteins absorb light in the 450-550 nm range, but others like the well-studied bacteriorhodopsin function best at 670 nm (21). Most commonly, the captured photon energy triggers the *cis-trans* isomerization of retinal, which completes a proton transfer pathway through the bundle interior and thus across the membrane. However, alternative functions of rhodopsins include sodium and chloride ion transport and activation of transducer proteins in the cytoplasm (22). Recently, a new class of rhodopsin, the heliorhodopsin, was discovered by screening a freshwater fosmid library derived from Lake

Kinneret for retinal binding proteins (23, 24). This rhodopsin type has an inverted topology compared to all other rhodopsins and a yet undetermined function. All that is known is that the holoprotein photocycles and pumps a protein only halfway (i.e. a proton is transferred to the retinal, but not further). Even though the contributions of heliorhodopsins to host organism physiology are unknown, it still requires the same retinal cofactor that is usually derived from carotenoids.

Carotenoids

The retinal that ultimately enables rhodopsins to function is derived from carotenoids, colorful molecules that act as photoprotectants, cofactors, pigments, and membrane stabilizers throughout nature (25, 26). The initial building blocks of carotenoids are the five-carbon monomers isopentenyl diphosphate and dimethylallyl diphosphate. These units can be enzymatically polymerized and modified to form simple carotenoids like lycopene and β-carotene and various complex chromophores with oxygen substituents and/or cyclizations, like astaxanthin and pittosporumxanthin C (27, 28). Glycosylations, radicals, and epoxides have also been identified. The absorbances of these molecules range from the blue to yellow (300-500 nm) because of conjugated-bond networks between ranging between three and fifteen bonds; the longer the system, the more red the chromophore appears (27). Additionally, simple and complex carotenoids act as potent oxidation sinks because of the conjugated bond networks. Deinococcus-Thermus bacteria demonstrate that organisms evolve to harness the combinatorial nature of carotenoid synthesis for their protective properties (29). In addition to shielding from oxidation, carotenoids in these organisms allow life in extreme heat, pressure, and radiation environments. Bond networks in other organisms also enable energy transduction and transfer to other chromophores. For example, photosystems (30), like the cyanobacterial photosynthetic apparatus, and antenna rhodopsins (31–33), like xanthorhodopsin, both rely on the extra energy funneled by carotenoids to boost activity. Photosystems rely on xanthophylls and carotenes for

resonance energy transfer into chlorophylls, while antenna rhodopsins use ketocarotenoids for transfer to retinal. Specific examples include zeaxanthin (34) and salinixanthin (35), respectively. However, for organisms to synthesize carotenoids all the necessary genes must be present.

Nitrogen Utilization via Peptidases

In general, bacteria monitor intracellular nitrogen concentrations using 2-oxoglutarate as an indicator of nitrogen starvation. To do this, bacteria express transcription factors (e.g. NtcA, Global Nitrogen Regulator) that upregulate the appropriate metabolic machinery (e.g. *nrt* genes for nitrate transport) (36). Nitrogen starvation involves the lack of nitrogen-containing molecules including simple inorganic compounds such as nitrates and dinitrogen and more complex macromolecules like proteins and DNA. However, the second class of molecules are long polymers. The individual units must be freed before the amino acids and nitrogenous bases can be used for cellular process. The salvage of amino acids in acI may be particularly important because they seem to scavenge much of their carbon and nitrogen from dissolved lake sources like amino acids, peptides, and polyamines (9, 14).

Cells are commonly able to hydrolyze proteins to free amino acids. Many of the initial enzymes in this process are nonspecific (i.e. they cleave bonds between a variety of amino acid combinations and act on peptides of varying lengths). As a result, smaller peptides and even free amino acids can be produced. However, as nonspecific hydrolysis continues, more uniform pools of smaller peptides with defined lengths accumulate. Consequently, specifically-acting enzymes are needed to quickly and efficiently complete the degradation process to amino acids (37). While some of these enzymes exhibit broad activity on short peptides with varied compositions like the initial nonspecific peptidases, others have restricted substrate pools (e.g. only dipeptides with specific leading amino acids can be cleaved) (38).

Even though known substrate specificities exist for enzymes, classifying these peptidases is complex because of protein sequence similarity. Most enzymes cleave main chain peptide

bonds between canonical amino acids, but there are also some that can hydrolyze the main chain peptide bond of aspartate-arginine monomers in cyanophycin, a ~25 – 100 kDa polymer produced inside cyanobacteria as insoluble granules (39, 40). Much actino-opsins, acI *Actinobacteria* are also predicted to encode cyanophycinases, the enzymes used for cyanophycin breakdown. However, no experimental evidence showing this has been published.

Despite this difficulty in enzyme function and, accordingly, substrates, all peptidases can be classified as either aspartic, glutamic, metallo, cysteine, serine, or threonine proteases. Of note are the serine enzymes that contain a catalytic serine and sometimes histidine and glutamate as a triad. This triad is seen the cyanophycinase, CphB, that breaks down the previously mentioned cyanophycin polymer and the dipeptidase, PepE, that strictly salvages the previously alluded to short peptides with specific sequence (aspartate-X, where X is any amino acid) (41). One such gene that matches the criteria for both enzymes is found in acI *Actinobacterial* genomes. Additionally, different databases have annotated it with both functions, yet neither activity has been experimentally confirmed.

Summary

This chapter has described the basics of fresh water acI *Actinobacteria* and three systems that organisms use for ecosystem survival. In the next three chapters, this work will specifically describe how freshwater acI *Actinobacteria* have evolved to 1) incorporate retinal biosynthesis and rhodopsin activity into their physiology 2) incorporate carotenoid biosynthesis machinery into their genomes 3) manifest dipeptide cleavage though a serine-histidine-glutamate catalytic triad. A final chapter will discuss how work in these areas can be expanded in the future.

References

1. Shiklomanov IA (ed. Gleick PH) 1993. "World Fresh Water Resouces" in Water in Crisis: A Guide to the World's Fresh Water Resources. Oxford University Press, New York.

2. Steinman AD, Isely ES, Thompson K. 2015. Stormwater runoff to an impaired lake: impacts and solutions. Environ Monit Assess 187: 549.

3. Murdock JN, Shields FD, Lizotte RE. 2013. Periphyton responses to nutrient and atrazine mixtures introduced through agricultural runoff. Ecotoxicology 22:215–230.

4. de Figueiredo DR, Azeiteiro UM, Esteves SM, Gonçalves FJM, Pereira MJ. 2004. Microcystin-producing blooms—a serious global public health issue. Ecotoxicol Environ Saf 59:151–163.

5. Merel S, Walker D, Chicana R, Snyder S, Baurès E, Thomas O. 2013. State of knowledge and concerns on cyanobacterial blooms and cyanotoxins. Environ Int 59:303–327.

6. Codd GA, Morrison LF, Metcalf JS. 2005. Cyanobacterial toxins: risk management for health protection. Toxicol Appl Pharmacol 203:264–272.

7. Falconer IR, Burch MD, Steffensen DA, Choice M, Coverdale OR. 1994. Toxicity of the blue-green alga (cyanobacterium) *Microcystis aeruginosa* in drinking water to growing pigs, as an animal model for human injury and risk assessment. Environ Toxicol Water Qual 9:131–139.

8. Garcia SL, McMahon KD, Martinez-Garcia M, Srivastava A, Sczyrba A, Stepanauskas R, Grossart H-P, Woyke T, Warnecke F. 2013. Metabolic potential of a single cell belonging to one of the most abundant lineages in freshwater bacterioplankton. ISME J 7:137-47.

9. Ghylin TW, Garcia SL, Moya F, Oyserman BO, Schwientek P, Forest KT, Mutschler J, Dwulit-Smith J, Chan L-K, Martinez-Garcia M, others. 2014. Comparative single-cell genomics reveals potential ecological niches for the freshwater acI *Actinobacteria* lineage. ISME J 8:2503-16.

10. Newton RJ, Jones SE, Eiler A, McMahon KD, Bertilsson S. 2011. A Guide to the Natural History of Freshwater Lake Bacteria. Microbiol Mol Biol Rev 75:14–49.

11. Neuenschwander SM, Ghai R, Pernthaler J, Salcher MM. 2018. Microdiversification in genome-streamlined ubiquitous freshwater *Actinobacteria*. ISME J 12:185–198.

12. Warnecke F, Sommaruga R, Sekar R, Hofer JS, Pernthaler J. 2005. Abundances, identity, and growth state of *Actinobacteria* in mountain lakes of different UV transparency. Appl Environ Microbiol 71:5551–5559.

13. Kang I, Kim S, Islam MR, Cho J-C. 2017. The first complete genome sequences of the acI lineage, the most abundant freshwater *Actinobacteria*, obtained by whole-genome-amplification of dilution-to-extinction cultures. Sci Rep 7:42252.

14. Hamilton JJ, Garcia SL, Brown BS, Oyserman BO, Moya-Flores F, Bertilsson S, Malmstrom RR, Forest KT, McMahon KD. 2017. Metabolic Network Analysis and Metatranscriptomics Reveal Auxotrophies and Nutrient Sources of the Cosmopolitan Freshwater Microbial Lineage acI. MSystems 2:e00091–17.

15. Garcia SL, McMahon KD, Grossart H-P, Warnecke F. 2014. Successful enrichment of the ubiquitous freshwater acI *Actinobacteria*. Environ Microbiol Rep 6:21–27.

16. Kim S, Kang I, Seo J, Cho J. 2018. Culturing the ubiquitous freshwater *Actinobacterial* acI lineage by supplying a biochemical “helper” catalase. bioRxiv. August 7, 2018.

17. Sharma AK, Sommerfeld K, Bullerjahn GS, Matteson AR, Wilhelm SW, Jezbera J, Brandt U, Doolittle WF, Hahn MW. 2009. Actinorhodopsin genes discovered in diverse freshwater habitats and among cultivated freshwater *Actinobacteria*. ISME J 3:726.

18. Sharma AK, Zhaxybayeva O, Papke RT, Doolittle WF. 2008. Actinorhodopsins: proteorhodopsin-like gene sequences found predominantly in non-marine environments. Environ Microbiol 10:1039–1056.

19. Spudich JL, Yang C-S, Jung K-H, Spudich EN. 2000. Retinylidene proteins: structures and functions from archaea to humans. Annu Rev Cell Dev Biol 16:365–392.

20. Man D, Wang W, Sabehi G, Aravind L, Post AF, Massana R, Spudich EN, Spudich JL, Béjà O. 2003. Diversification and spectral tuning in marine proteorhodopsins. EMBO J 22:1725–1731.

21. Lozier RH, Bogomolni RA, Stoeckenius W. 1975. Bacteriorhodopsin: a light-driven proton pump in Halobacterium Halobium. Biophys J 15:955–962.

22. Beja O, Lanyi JK. 2014. Nature's toolkit for microbial rhodopsin ion pumps. Proc Natl Acad Sci 111:6538–6539.

23. Pushkarev A, Inoue K, Larom S, Flores-Uribe J, Singh M, Konno M, Tomida S, Ito S, Nakamura R, Tsunoda SP, Philosof A, Sharon I, Yutin N, Koonin EV, Kandori H, Béjà O. 2018. A distinct abundant group of microbial rhodopsins discovered using functional metagenomics. Nature 558:595–599.

24. Singh M, Inoue K, Pushkarev A, Béjà O, Kandori H. 2018. Mutation Study of Heliorhodopsin 48C12. Biochemistry 57:5041–5049.

25. Olson JA, Krinsky NI. 1995. Introduction: the colorful, fascinating world of the carotenoids: important physiologic modulators. FASEB J 9:1547–1550.

26. Gruszecki WI, Strzałka K. 2005. Carotenoids as modulators of lipid membrane physical properties. Biochim Biophys Acta BBA - Mol Basis Dis 1740:108–115.

27. Britton G, Liaaen-Jensen S, Pfander H (ed). 2004. Carotenoids. Springer Basel AG, Basel, Switzerland

28. Graham JE, Bryant DA. 2009. The Biosynthetic Pathway for Myxol-2' Fucoside (Myxoxanthophyll) in the Cyanobacterium Synechococcus sp. Strain PCC 7002. J Bacteriol 191:3292–3300.

29. Tian B, Hua Y. 2010. Carotenoid biosynthesis in extremophilic Deinococcus–Thermus bacteria. Trends Microbiol 18:512–520.

30. Grotjohann I, Fromme P. 2005. Structure of cyanobacterial Photosystem I. Photosynth Res 85:51–72.

31. Luecke H, Schobert B, Stagno J, Imasheva ES, Wang JM, Balashov SP, Lanyi JK. 2008. Crystallographic structure of xanthorhodopsin, the light-driven proton pump with a dual chromophore. Proc Natl Acad Sci 105:16561–16565.

32. Balashov SP. 2005. Xanthorhodopsin: A Proton Pump with a Light-Harvesting Carotenoid Antenna. Science 309:2061–2064.

33. Balashov SP, Imasheva ES, Choi AR, Jung K-H, Liaaen-Jensen S, Lanyi JK. 2010. Reconstitution of Gloeobacter Rhodopsin with Echinenone: Role of the 4-Keto Group. Biochemistry 49:9792–9799.

34. Zakar T, Laczko-Dobos H, Toth TN, Gombos Z. 2016. Carotenoids Assist in Cyanobacterial Photosystem II Assembly and Function. Front Plant Sci 7:295

35. Lutnaes BF, Oren A, Liaaen-Jensen S. 2002. New C_{40} -Carotenoid Acyl Glycoside as Principal Carotenoid in *Salinibacter ruber* , an Extremely Halophilic Eubacterium. J Nat Prod 65:1340–1343.

36. Leigh JA, Dodsworth JA. 2007. Nitrogen Regulation in Bacteria and Archaea. Annu Rev Microbiol 61:349–377.

37. Miller CG. 1996. Protein Degradation and Proteolytic Modification, p. 938–954. Am. Soc. Microbiol., Washington, DC.

38. Carter TH, Miller CG. 1984. Aspartate-Specific Peptidases in *Salmonella* typhimurium: Mutants Deficient in Peptidase E. J Bacteriol 159:453-459.

39. Simon RD, Weathers P. 1976. Determination of the structure of the novel polypeptide containing aspartic acid and arginine which is found in cyanobacteria. Biochim Biophys Acta BBA - Protein Struct 420:165–176.

40. Simon RD. 1971. Cyanophycin Granules from the Blue-Green Alga Anabaena cylindrica: A Reserve Material Consisting of Copolymers of Aspartic Acid and Arginine. Proc Natl Acad Sci 68:265–267.

41. Lassy RAL, Miller CG. 2000. Peptidase E, a Peptidase Specific for N-Terminal Aspartic Dipeptides, Is a Serine Hydrolase. J Bacteriol 182:2536–2543.

Chapter II – Retinal Synthesis and Rhodopsins in acI *Actinobacteria*

Publication Statement

Parts of this chapter have been published as: Dwulit-Smith JR, Hamilton JJ, Stevenson DM, He S, Oyserman BO, Moya-Flores F, Garcia SL, Amador-Noguez D, McMahon KD, Forest KT. (2018). acI *Actinobacteria* Assemble a Functional Actinorhodopsin with Natively Synthesized Retinal*. Applied and Environmental Microbiology. 84 (24): e01678-18. doi: 10.1128/AEM.01678-18*

Scientific Contribution Statements

Dwulit-Smith JR performed all bioinformatics and biochemical experiments and contributed to HPLC-MS operation and method development.

Hamilton JJ analyzed metagenomic data and calculated RPKM.

Stevenson DM contributed to HPLC-MS method development and system operation.

He S collected environmental acI samples and generated the rhodopsin phylogeny.

Oyserman BO collected environmental acI samples.

Moya-Flores F collected environmental acI samples and performed RNA extraction and sequencing.

Garcia SL helped select and obtain genome AAA278-O22 for gene expression.

Amador-Noguez D, McMahon KD, and Forest KT advised the experimental plans.

Abstract

Genomic evidence suggests that acI *Actinobacteria* assemble actinorhodopsins for light utilization from actino-opsin platform proteins. In addition to the opsin, acI would also require retinal, a carotenoid-derived chromophore, to complete the holoprotein. Here, I map genes for carotenoid production and show that some acI produce retinal using a β-carotene oxygenase. Consequently, these cells can form actinorhodopsins. I verify this by assembling an acI actinorhodopsin in *Escherichia coli* and measuring its proton-pumping during illumination. Transcriptomics of Lake Mendota confirm that all retinal pathway genes are expressed by acI and that actino-opsin transcripts are at least hundreds of times more abundant than the other transcripts. Actino-opsin transcripts are also among the top five most highly expressed acI genes. Furthermore, I show that some acI contain recently discovered heliorhodopsins, unique proton-capturing inverted-topology rhodopsins. Determining if the heliorhodopsin serves as an additional adaptation to freshwater survival is a next step in defining how acI *Actinobacteria* mitigate the challenges of their ecosystems.

Introduction

Visible-wavelength photoreceptors are pervasive light sensitive proteins that transform light signals into responses useful to the host organism. Organisms from all types of life contain these valuable systems. All these systems share a unifying architecture: a protein platform apoprotein acts as a covalent attachment or binding site for a chromophore to form a light-responsive holoprotein. The chromophore is a small molecule that contains an extensive conjugated bond network that allows the absorption of wavelengths specific to each chromophore. Well-studied examples of these holoproteins are flavoproteins, biliproteins, and especially retinylidene proteins. While the first examples have many different types of proteins, retinylidene proteins are usually rhodopsins.

All rhodopsins regardless of organism are opsin apoproteins that contain seven transmembrane helices that wrap into a bundle such that helices α1 and α7 are next to each other (1, 2). Generally, the N-terminus of the protein faces the cytoplasm while the C-terminus ends on the other side of the membrane. Another unifying feature of all opsins is that they covalently bind retinal at a lysine via a Schiff base inside of the bundle (1). The amino acid pocket alters the electronic organization of the bound retinal leading to large changes in absorbance compared to retinal alone (i.e. red, green, and even blue maximal absorption) (3). Rhodopsin absorbances contrast with the native retinal absorption in the ultraviolet region around 380 nm. Although the main chromophore unites all rhodopsins, their functions can widely vary (4). While some holoproteins create intracellular signaling cascades as in eukaryotic visual rhodopsins or cyanobacterial sensory rhodopsins (5), others allow the movement of ions through the membrane. These ion shuttling rhodopsins include multidirectional cation channels, inward chloride pumps, outward sodium pumps, and most commonly outward proton pumps (6). This last class of ion moving rhodopsins is very prevalent in the microbial word.

Microbial proton pumps were first discovered in Archaea and named bacteriorhodopsins (7). However, work has moved far beyond bacteriorhodopsin to many other branches of the

rhodopsin family, like proteo-, xantho-, and actinorhodopsins (3, 8, 9). While all these specific holoproteins pump protons outwards to create a proton gradient using a set of photocycle intermediates, their photocycling pathway times can vary.

The photocycle of rhodopsins can be divided into a few active states: ground, K, L, M1, M2, M2' N, N', and O (10, 11). The cycle canonically starts at the ground state where the retinal is in an all-trans configuration. Once light of the appropriate wavelength interacts with the holo-protein, 13-cis retinal isomerization occurs, and the cis-configuration persists until the O state. Once the rhodopsin is in this state, the chromophore has returned to all-trans, but the protein has not fully relaxed back to the ground state sidechain arrangement. The various other excited 13-cis retinal states are separated by temporal rearrangements of the amino acid side chains and differing locations of the proton that crosses the Schiff base. Specifically, the M2' state has released a proton and the N' state has a new proton attached (10). As for photocycle intermediate timings, longer steps always occur later. The differences between rhodopsins exist in the exact state change times and the overall duration of one photocycle. For example, bacteriorhodopsin has timings of 3ps,1µs, 40µs, 5ms, 5ms, and 5ms (times are grouped by major state rather than subdivisions) (11), but an actinorhodopsin from *Candidatus Rhodoluna planktonica* has timings of not determined, 320µs, 590µs, 1.9ms, 13ms, and 136ms (12). The order of magnitude difference in total bacteriorhodopsin and actinorhodopsin cycle duration, ~15ms versus ~152ms, has major consequences for total proton pumping capacity, and robustness of the proton gradient may or may not have major consequences for host organism physiology (13).

While most proteorhodopsin and xanthorhodopsin work has been done in marine organisms and hyperhalophilic species, actinorhodopsin work has been concentrated in freshwater organisms. Specifically, nearly all acI *Actinobacteria* were shown to encode actino-opsins even though the lineage trifurcates into three clades (acI-A, acI-B, acI-C). Actino-opsins are prototypical in that their sequences display the homology, topology, and chromophore binding of general rhodopsins. However, little work has been done to ensure that all actino-opsins can

natively function as actinorhodopsins (9, 12). Even though acI actino-opsins are highly homologous to xanthorhodopsin, a well-studied and crystallized proton-pump from *Salinbacter ruber*, it is not guaranteed that actino-opsins function as proton pumps, especially *in vivo*. Not only does an organism need to express the opsin gene, but it must also produce or acquire enough retinal to reconstitute the holoprotein. This second fact seems to be the major problem for many organisms like *Rhodoluna lacicola*; that is, they contain easily identifiable opsins but seemingly lack the enzyme to produce retinal, a β-carotene oxygenase (9). While there may be a system for specific retinal uptake, no such mechanism other than passive absorption has been identified.

Yet, the field is constantly changing. Recent work expressing fosmids in the presence of retinal has identified an entirely new and unique class of rhodopsin, the heliorhodopsin (14). Although only one example has been studied so far, many others have been bioinformatically identified across nature (14). Heliorhodopsins are like regular rhodopsins in that they have seven transmembrane helices that form a bundle, and the bundle binds retinal. However, the topology is completely reversed. The N-terminus of the protein points outward while the C-terminus in located in the cytoplasm. Interestingly, the holoproteins respond to light and photocycle properly, but the proton that is shuttled to the Schiff base is never pumped out of the cytoplasm (14). This leaves the question of what the physiological role of these newly discovered rhodopsins is.

Results

All acI ActR contain key amino acids for function. I first determined whether acI ActR sequences were consistent with holo-ActR formation. Indeed, all contain the features for proper rhodopsin structure and function (Fig. 2.01A). Seven predicted helices (α1 to α7) match those found in xantho-opsin, a close homolog of ActR from *Salinibacter ruber* (15). A conserved lysine for Schiff base formation and acidic residues for proton-shuttling across the retinylidene gate are present (7), and a leucine is predicted to tune the absorbance of all acI holo-ActRs to the green

region (16). I also found novel features. A proline in the middle of α4 differentiates ActRs from xantho-opsins, and the residue may serve as a means for better phylogenetic classification (Fig. 2.02). Additionally, 3D structure predictions closely position two cysteines on α1 and α7 in clades A and B, which may allow a disulfide bond to covalently staple the protein together (Fig. 2.01B). acI ActRs also contain glycine near the top of α5, a required feature for binding ketolated antenna carotenoids (8, 17). This glycine replaces a bulky residue, usually tryptophan, found in most rhodopsins, including the well-characterized bacteriorhodopsin.

Seven gene products form an actinorhodopsin pathway. Retinal is almost certainly the chromophore needed to form holo-ActR in acI *Actinobacteria*. I find that in addition to encoding ActR, acI single-cell amplified genomes (SAGs), metagenome-assembled genomes (MAGs), and complete genomes from dilution-to-extinction cultures (Table 2.01) appear to encode genes for enzymes that produce retinal and its carotenoid precursors (Fig. 2.03). I assembled a plausible, complete pathway with protein assignments for forming lycopene, β-carotene, retinal, and subsequently, actinorhodopsin (Fig. 2.04A). Lycopene synthesis in acI *Actinobacteria* requires three major steps, as follows: synthesis of geranylgeranyl-PP (step 1), linkage of two geranylgeranyl-PP molecules to phytoene (step 2), and tetra-desaturation of phytoene to lycopene (step 3), predicted to be carried out by CrtE, CrtB, and CrtI, respectively. Subsequent β-cyclization of lycopene would first produce β-carotene (step 4) and then γ-carotene (step 5). A heterodimeric enzyme composed of CrtYc and CrtYd likely carries out these serial cyclizations, much like *Myxococcus xanthus* β-cyclase genes [36% and 33% identity, trihelical transmembrane topology, PxE(E/D) catalytic motif] (Fig. 2.07) (18–21).The symmetric cleavage of β-carotene by a β-carotene oxygenase, Blh, would then form retinal (step 6). Blh is a putative dioxygenase, based on 27% sequence identity, predicted helical topology, and four histidines for nonheme iron coordination that are shared with the characterized β-carotene dioxygenase from an uncultured marine bacterium (Fig. 2.09) (19, 20, 22). Functional holo-ActR requires retinal to autocatalytically form a Schiff base with the side chain of a conserved lysine in ActR (step 7).

In the genomic context, genes encoding the enzymes for chromophore production are grouped into functional regions (Fig. 2.03). Lycopene production genes (*crtE, crtB, and crtI*) are found as a neighborhood in the order of steps 1, 3, and 2 in all acI clades. Genes for steps 4 to 6 (*crtYc, crtYd*, and *blh*) are found at various distances from the lycopene synthesis cluster. In many cases, the 3' ends of *crtB* and *crtYd* are adjacent but encoded in the opposite sense, thus forming a β-carotene synthesis neighborhood. In other cases, the cyclase genes are instead found near *blh*. In most cases, *actR* is widely separated from other pathway genes. An interesting exception is AAA044-D11, where *actR, crtYc, crtYd,* and *blh* are contiguous (Fig. 2.03).

Metatranscriptomic analysis of pathway gene transcripts in environmental acI populations. For actinorhodopsin assembly in acI cells, actR and retinal synthesis genes must be expressed. To measure gene expression in environmental acI *Actinobacteria*, four metatranscriptome samples were collected across multiple time points from the surface of eutrophic Lake Mendota (Dane County, WI, USA), and RNA was isolated and sequenced. The resulting transcripts were mapped to available acI SAGs and MAGs (Table 2.01) to quantify relative gene expression levels in acI cells. Notably, actR is the most highly expressed acI-A gene and the second most highly expressed acI-B gene (23), as well as the most highly expressed gene from either pathway in each of the three acI clades (Table 2.02). Transcription is also observed for genes in the retinal pathway, but at levels several hundredfold lower than for actR. No transcripts were mapped for blh from acI-C because the gene is not present in any acI-C SAG or MAG or in the first complete acI-C genome obtained by dilution-to-extinction cultures (24).

CrtE, CrtB, and CrtI produce lycopene. I sought to demonstrate whether CrtE, CrtB, and CrtI enzymes from acI can form lycopene as predicted (Fig. 2.04, steps 1 to 3). Therefore, a fast, reproducible method for pigment extraction from whole *Escherichia coli* cells and identification by high-performance liquid chromatography (HPLC)-mass spectrometry (MS) analysis was developed. Relevant genes (*crtE, crtB, and crtI*) from assembly AAA278-O22 were cloned into pCDFDuet1 and expressed from the resulting acI-CrtEBI/- cassette (Table 2.03).

Assembly AAA278-O22 was chosen for carotenoid production work because it contained all relevant genes and source DNA was available (25). The extracted compounds of these cells displayed absorbance maxima which exactly matched those of a lycopene standard at expected positions of 447, 472, and 504 nm (Fig. 2.06A) and displayed an intense red color (data not shown). HPLC-MS elution profiles indicated the presence of cis and all-trans isomer peaks at *m/z* 536.438 (Fig. 2.06B). The retention time of the all-trans species (39.90 min) coincides for sample and standard as the peak with highest intensity. The lycopene assignment was further confirmed by tandem mass spectrometry (MS/MS) fragmentation of the parent ion.

To further characterize the lycopene product of acI CrtE, CrtB, and CrtI enzymes, I tested whether it serves as the substrate in a lycopene cyclization reaction. *Pantoea ananatis* genes *crtE, crtB, crtI,* and *crtY* were expressed in *E. coli* as positive controls (Table 2.03, and Fig. 2.08A). When the *crtY* cyclase gene and acI lycopene pathway genes were coexpressed from pCDFDuet1 acI-CrtEBI/Pa-CrtY, the cellular extract absorbance maxima were 407, 429, and 453 nm (Fig. 2.08B). These dramatically blue-shifted maxima, compared to those of β-carotene, combined with the defined cyclase activity of CrtY, identify the major extract product as β-zeacarotene, β-carotene saturated between C-7' and C-8' (Fig. 2.08B). Thus, CrtE, CrtB, and CrtI from acI produce a chromophore, which serves as a substrate for a lycopene cyclase. Presumably, more β-carotene would be produced after complete desaturation and a second cyclization, which were inefficiently carried out in our heterologous expression system.

Blh produces retinal from β-carotene. The final enzymatic step in retinal biosynthesis is the symmetric cleavage of a cyclized carotene (Fig. 2.04, step 6). This enzyme is not encoded in any acI-C genomes (Fig. 2.03). To show that acI can natively perform retinal synthesis, I tested the enzymatic activity of AAA278-O22 Blh. This blh gene was expressed from pCDFDuet1 Pa-CrtEBIY/acI-Blh (Table 2.03) to ensure a large β-carotene substrate pool. Introduction of Blh yielded yellow cells instead of the intensely orange colored cells observed when β-carotene is abundant (data not shown). HPLC-MS confirmed the presence of retinal at *m/z* 285.221 (Fig.

2.10A). Retinol was also identified via its dehydrated species at *m/z* 269.226 (Fig. 2.10B), as described by Breeman and Huang (47). As is the case for lycopene, geometric isomer peak patterning was observed. Maximal all-trans species retention times match those of retinal and retinol standards at 23.76 min and 23.88 min, respectively. Additionally, MS/MS fragmentation of the appropriate *m/z* species confirmed retinoid identities (Fig. 2.10A and B). Retinol was roughly seven-fold more abundant than retinal in the extract, as judged by MS intensities. Accordingly, the UV-visible (UV-Vis) absorbance profile appears more like that of retinol (Fig. 2.10C). Retinal produced by Blh is likely serving as a potent electron acceptor for the *E. coli* alcohol dehydrogenase, *ybbO* (26, 27).

ActRL06 with retinal forms an active, green light-dependent proton pump. The gold-standard test of a functional rhodopsin is light-dependent activity. Given that acI ActR proteins contain the residues for chromophore binding and proton movement, I tested an opsin from each clade (AAA278-O22, AAA027-L06, and MEE578) for production and chromophore binding in *E. coli.* Each opsin was expressed from pET21b+ alongside the plasmid confirmed for retinal production. ActRL06 was selected for further characterization due to its high expression and efficient retinal binding, as judged from samples captured by metal affinity from detergent-solubilized membranes. Holo-ActRL06 maximally absorbed in the green region at 541 nm (Fig. 2.11A). The covalent attachment of retinal was confirmed by incubation with hydroxylamine hydrochloride, which frees the retinal and leads to production of retinal oxime. For conclusive demonstration that holo-ActRL06 outwardly pumps protons in the presence of light when in a membrane environment, microelectrode pH measurements were performed. During exposure to white light, the pH of a nonbuffered assay solution sharply decreases compared to that in periods of dim red light (Fig. 2.11B). To confirm protons as the source of the steep pH drop, carbonyl cyanide m-chlorophenylhydrazone (CCCP) was added to discharge the proton gradient. CCCP trials did not display light-dependent proton pumping but rather the constant downward drift of

control cells. Thus, the acI holo-ActR is a retinal-bound, green light-dependent proton-pumping rhodopsin.

A putative heliorhodopsin exists in acI. I first determined whether acI HeR sequences were consistent with a published, characterized HeR. Indeed, the HeRL06 contains high homology in both sequence and predicted structural organization (Fig. 2.12). Seven predicted helices (I to VII) match those found in HeR 48C12, the only studied heliorhodopsin (14, 28). A conserved lysine for Schiff base formation is present, but one of the acidic residues for proton-shuttling across the retinylidene gate is missing. Also, the overall topology is predicted to have an N-terminus-in, C-terminus-out arrangement (FIG 2.13). The helix arrangement agrees with the analysis of HeR 48C12 (N-in, C-out) and is opposite to ActRL06. The topology agrees with a high concentration of positively charged residues on the cytoplasmic side of the membrane. In fact, many of the fourteen lysines and arginines present on the cytoplasmic side of HeR 48C12 are present or replaced elsewhere in the acI sequence (Fig. 2.12). As HeRL06 alignment and topology results agreed with the characterized opsins, HeR 48C12 and ActRL06. the gene was cloned and overexpressed with retinal to assess cellular color change (Fig.1.14). White cells took on a red-orange color consistent with holo-HeRL06 formation. Tests for proton gradient formation in response to light were negative when the acI helio-opsin was expressed along with retinal-producing machinery (Fig. 2.15).

Discussion

Prior to this work, there was no biochemical evidence to support the hypothesis that acI *Actinobacteria* use opsin-based phototrophy in freshwater. I have provided experimental verification of an advantage that allows the acI lineage to be so abundant. I describe two favorable environmental adaptations, retinal biosynthesis and light utilization. Transformation of simple isoprenoid precursors into carotenoids and retinal allow ActR to function as a green light-absorbing, outward proton-pumping holo-ActR.

The pathway for retinal and rhodopsin synthesis has been experimentally confirmed using acI *Actinobacterial* proteins. The machinery to start retinal production consists of three enzymes, CrtE, CrtB, and CrtI (steps 1 to 3). The enzymes produce lycopene (Fig. 2.04 and Fig. 2.06) and cluster into the *crtEIB* operon (Fig. 2.03 and Fig. 2.05). The reactions producing γ-carotene and β-carotene (steps 4 to 5) that follow lycopene synthesis are likely carried out by a heterodimeric membrane protein expressed from the *crtYc* and *crtYd* operon (Fig. 2.03, Fig. 2.07). Homologs of these gene products are found in *Myxococcus xanthus*, where they were shown to synthesize a mixture of γ-carotene and β-carotene (18). Therefore, I propose that acI can also synthesize these compounds. Although the ratio of these cyclized products in acI is unknown, it may be intrinsically linked to intracellular concentrations of retinal and complex carotenoid. Unsaturated but cyclized intermediates like β-zeacarotene (Fig. 2.11) may result from differing levels of cyclase compared to those of CrtE, CrtB, and CrtI. I note that a system for stable β-zeacarotene production could prove a fruitful biotechnology tool because monocyclized carotenoids are not readily available for purchase.

A key finding of our work is the presence, expression, and robust activity of a retinal-producing oxygenase from the acI gene, *blh*. The enzyme symmetrically cleaves β-carotene to two retinal molecules (Step 6) (Fig. 2.04A and Fig. 2.11) and might additionally use alternative substrates, such as a β-zeacarotene or other β-carotene-like carotenoids, to yield a single retinal. Notably, the *blh* gene is not found in acI-C genomes; in five acI-C genomes ranging in completeness from 25 to 100%, *blh* has never been recovered (Table 2.01) (4). This absence points toward clade-level differentiation with respect to retinal production and subsequent holo-ActR assembly in acI *Actinobacteria*. Specifically, I propose that populations of acI-A and acI-B maintain a high relative abundance in the community throughout the year (29, 30) because they can synthesize their own retinal to support phototrophy; therefore, phototrophy could support heterotrophy. In contrast, acI-C *Actinobacteria* exhibit “bloom and bust” cycles coincident with cyanobacterial blooms (31). It will require further study to determine if this acI-C variability is due

to an inability to harvest light or a dependence on exogenous retinal, potentially from lysed cyanobacteria. Indeed, many species of cyanobacteria encode functional β-carotene oxygenases (22, 32, 33), and retinoid concentrations in eutrophic lakes during cyanobacterial blooms are measurable (34). This proposal is consistent with actinorhodopsin studies in the culturable freshwater organisms *Rhodoluna lacicola* and "*Candidatus Rhodoluna planktonica*" (9, 12). While both carry actR, only the latter contains *blh* and thus exhibits self-sufficient rhodopsin activity in the laboratory. *R. lacicola* is dependent on retinal supplementation. As such, it may be analogous to *blh*-lacking organisms. Indeed, recently reported complete genomes (35) indicate that even some acI-A and acI-B *Actinobacteria* would require an alternative retinal source. This heterogeneity in gene content has major implications for niche filling by acI-A and acI-B members and may dictate community-related growth patterns.

Regardless of whether acI cells synthesize or scavenge retinal, all acI ActR proteins contain the necessary features for proper activity as retinal holoproteins with green maximal absorbance, and I have demonstrated that an example acI ActR functions as a retinylidene protein (Fig. 2.10 and Fig. 2.11). Green absorption correlates with light penetration depth for many freshwater bodies where acI *Actinobacteria* thrive and strengthens the case that holo-ActRs in acI are light-activated proteins in the environment. Native production of holo-ActR would enable acI cells to produce a proton gradient even when central metabolism intermediates are scarce. In addition to showing the proton-pumping activity of holo-ActR, I discovered interesting qualities of ActRs (Fig. 2.01 and Fig. 2.02). A proline in α4 was determined to be a phylogenetically differentiating residue between actino-opsins and other xantho-opsins, like the one in *Salinibacter ruber*. Prolines kink helices, and these residues may have broad structural effects on photocycling times and/or intermediate photocycle structures. Similarly, a disulfide bond joining α1 to α7 in acI-A and acI-B *Actinobacteria* could also impact activity and/or protein stability.

Helio-opsin was also found in several genomes. Because 48C12 is the only characterized example of the newly discovered heliorhodopsins (14, 28), it was used as a model search

sequence. Putative helio-opsins (HeR) were found in several acI genomes including the L06 genome. This more complete genome also provided the functionally characterized ActR (Fig. 2.12, Fig. 2.13) and retinal pathway genes. Additionally, probable retinal binding and probable inability of proton pumping in response to light were demonstrated (Fig. 2.14, Fig. 2.15). However, other factors like low helio-opsin expression, low heliorhodopsin assembly, or other control-mimicking function might be at play in the pumping assay. Because holo-HeRL06 seems to form and does not pump protons in response to light, it is interesting to think about the possible role that the protein plays in both acI physiology and protein function. Perhaps the protein is a signaling rhodopsin that binds an unknown transducer protein. The stimulus may be light or retinal itself. On the other hand, holo-HeR could be a protein that modulates holo-ActR activity by direct interaction.

Methods

acI gene identification and pathway assembly. Annotations for multiple acI SAGs (25) and MAGs were analyzed for genes relating to carotenoids using the Joint Genome Institute's Integrated Microbial Genomes Viewer (36). Translated candidate protein sequences were used to identify homologs in other acI genomes, and those with consistent gene neighborhoods and known carotenoid-related functions were prioritized. After function assignments, two pathways for carotenoid biosynthesis and use in acI *Actinobacteria* were assembled.

acI biomass collection and transcriptomics. Environmental sampling, metatranscriptome sequencing, and gene expression calculations were all performed as previously described (23). Briefly, four samples were collected from within the top 12 m of the Lake Mendota (Dane County, WI) water column and filtered through cheesecloth and onto 0.2µm mixed cellulose filters (Whatman). Filters were immediately frozen in liquid nitrogen and stored at 80°C. Samples were subjected to TRIzol-based RNA extraction, phenolchloroform separation, and RNA precipitation. RNA was further purified using the RNeasy minikit (Qiagen) with on-

column digestion of DNA via the RNase-free DNase set (Qiagen). RNA was then sent to the University of Wisconsin—Madison Biotechnology center for rRNA depletion, cDNA synthesis, and sequencing. rRNA was depleted using the RiboZero rRNA removal kit (bacteria) (Illumina). Samples were then prepared for sequencing using the TruSeq RNA library prep kit v2 (Illumina), pooled in an equimolar ratio, and sequenced on an Illumina HiSeq 2500 platform using 2 100-bp paired-end sequencing. After sequencing, metatranscriptomic reads were trimmed, merged, subjected to in silico rRNA removal, mapped to carotenoid-related genes, and counted. Raw paired-end reads were trimmed using Sickle (37), merged using FLASH (38), and subjected to in silico rRNA removal using SortMeRNA. Sickle was run using default parameters; FLASH was run with a maximum overlap of 100 nucleotides; and SortMeRNA (39) was run using databases for bacterial, archaeal, and eukaryotic rRNA, derived from SILVA v119 and RFAM v12.0 (40). Trimmed and merged reads from all four samples were then pooled and mapped to a single reference FASTA file containing 36 high-quality acI genomes from a larger freshwater genome collection (11). Reads were competitively mapped to genes using BBMap (https://sourceforge .net/projects/bbmap/) with the ambig random and minid 0.95 options. Next, reads mapping to each carotenoid-related gene were counted using hts-count (41) with the intersection_strict option. Within each clade, reads mapping to each carotenoid-related gene were pooled, and gene expression was computed on a reads per kilobase per million (RPKM) basis (42), while also accounting for different gene lengths and total mapped reads for each acI genome. All scripts are found on GitHub (https://github .com/joshamilton/Hamilton_acI_2017/tree/actR).

Plasmid construction. All cloning was performed using Phusion high-fidelity (HF) GC master mix (Thermo Fisher) and the listed primers (Table 2.04). A collection of stable DNA sources for amplification and subcloning was first created. A plasmid containing E. coli-codon-optimized crtE, crtB, crtI, and crtY from Pantoea ananatis (Pa) was obtained from the International Genetically Engineered Machine (iGEM) organization catalog (http://parts.igem.org/Main_Page) (Table 2.03). The genes were amplified by PCR from the iGEM plasmid as a block. acI

biosynthetic genes were amplified by PCR from AAA278-O22 genomic DNA as gene clusters. The L06 opsin sequence was obtained as an E. coli-optimized gene from DNA2.0. All genes were individually amplified by PCR, if needed, and added to a cloning pipeline. Primer design included up to 40 bp of cloning site flanking regions from pCDFDuet-1 (Novagen). Genes were placed into the first site between NcoI and AflII restriction sites and between NdeI and AvrII sites of the second site by ligation-free recombination in E. cloni 10G cells (Lucigen). The insert molarity was up to 10-fold more abundant than that of appropriately digested and purified backbone. After selection on 100 g/ml spectinomycin sulfate LB-Miller agar plates, insert presence was validated by colony PCR using GoTaq Green master mix (Thermo Fisher). Restriction enzyme digestion in lab and Sanger sequencing at the UW Biotechnology Center further confirmed plasmid correctness. pET21b+ (Novagen) was used in the same pipeline, ensuring that genes were in-frame with a C-terminal hexahistidine tag with selection by 100 g/ml disodium carbenicillin.

Chromophore production and extraction. For chromophore production, multiple colonies of freshly transformed BL21(DE3) Tuner E. coli (EMD Millipore) were picked into half-full 2-liter flasks of LB-Miller broth plus 100 g/ml spectinomycin sulfate and shaken in 37°C darkness at 250 rpm for 24 h. ODml (ODml optical density at 600 nm [OD600] dilution volume in milliliters) cells (500 or 1,500) were centrifuged at 3,300 g, washed in 100 mM Tris (pH 6.8), centrifuged again, and frozen in liquid nitrogen. For chromophore extraction, cells stored at 80°C were thawed at 23°C for up to 30 min. Cells were suspended at 500 ODml cells/3 ml acetone (HPLC grade; Sigma), intensely vortexed for 1 min, and incubated on ice for 5 min. After clarification by centrifugation at 8,000 g for 10 min, the chromophore-containing supernatant was added to 1 ml/500 ODml of 23°C NaCl-saturated water (American Chemical Society [ACS] grade; Fisher) and 1 ml/500 ODml 23°C dichloromethane (ACS-grade; ACROS) and vortexed for 1 min. Further centrifugation resulted in a colorless aqueous bottom layer and a colorful organic top layer. The organic layer was evaporated under nitrogen, resulting solids were suspended in acetone for

carotenoids or ethanol for retinoids, and the resulting solution was filtered through a compatible 4-mm nylon 0.22µm filter.

UV-Vis and LC-MS/MS analysis of chromophores. Absorbance spectra were acquired on a Beckman Coulter DU640B spectrophotometer. A Dionex Ultimate 3000 ultra HPLC (UHPLC) coupled by electrospray ionization (ESI; positive mode) to a hybrid quadrupole– high-resolution mass spectrometer (Q Exactive Orbitrap, Thermo Scientific) was used for detection of target compounds based on their accurate masses, mass spectra, and retention times (all matched to purified standards). Liquid chromatography (LC) was based on a published protocol (43). Separation was achieved using a C30 reversed phase, 150 mm 2.5 mm, 3-m particle column (YMC Carotenoid) at a flow rate of 0.2 ml/min. Solvent A consisted of equal parts methanol and water with 0.5% vol/vol acetic acid; solvent B was equal parts methanol and methyl tert-butyl ether with 0.5% vol/vol acetic acid (43). Total run time was 58 min, with the following two gradients: 5 min at 30% B, followed by a 20-min ramp to 100% B and 100% B (retinoids); or 5 min at 30% B, followed by a 25-min ramp to 100% B, and 100% B (carotenoids). MS scans consisted of full positive mode scanning for *m/z* 200 to 600 from 5 min onwards. In addition, MS/MS scans were obtained by isolating fractions with *m/z* values of 537.44548 (lycopene), 285.22129 (retinal), and 269.22639 (retinol). MS/MS fragmentations were performed at a normalized-collision energy (NCE) of 30 with an isolation window of *m/z* 1.4 and a postfragmentation window of *m/z* 50 to approximately 25 above the isolation mass. For all scans, mass resolution was set at *m/z* 35,000 ppm, AGC target was 1 106 ions, and injection time was 40 ms. Settings for the ion source were as follows: auxiliary gas flow rate, 50; sheath gas flow rate, 10; sweep gas flow rate, 2; spray voltage, 3.5 kV; capillary temperature, 350°C; heater temperature, 250°C; and S-lens RF level, 55.0. Nitrogen was used as the nebulizing gas by the ion trap source. Standards for lycopene (L487500, lot 7ANR-20-1; Toronto Research Chemicals, Inc.) -carotene (PHR1239, 3 100 mg, lot LRAA6761; Sigma-Aldrich), retinal (R2500-25MG, lot SLBN4199V; Sigma-Aldrich), and retinol (R7632-25MG, lot BCBP8066V; Sigma-Aldrich) were analyzed along with experimental samples.

Isomer peaks within the mass window result most often from environmentally induced changes during sample preparation. Control cells did not display any significant sustained signal within the time-mass window. Data analysis was performed using Thermo Xcalibur and visualized on MAVEN (44, 45) and Thermo Xcalibur (Thermo Scientific) software. Proposed gene product and actino-opsin analysis. Clustal Omega was used to align proposed biosynthetic enzymes and opsin sequences for identification of protein characteristics (46–48).Opsin sequences were submitted to the I-TASSER server for structure prediction using xantho-opsin (PDB 3DDL:A) as the template (49–51). The Basic Local Alignment Search Tool (BLAST) and TM/HMM v2.0 were used for identity percentage calculation and transmembrane helix estimates, respectively (19, 20). PyMOL was used to visualize the resulting protein structures (52). The actino-opsin phylogeny was reconstructed using opsin protein sequences from bacterial isolate references, SAGs, and MAGs. Opsin protein sequences were aligned with the PROMALS3D multiple sequence and structure alignment tool (53). The alignment was trimmed to exclude positions that contained gaps for more than 30% of the included sequences. Poorly aligned positions and divergent regions were further eliminated by using Gblocks (54). A maximum-likelihood phylogenetic tree was constructed using PhyML v3.0 (55), with the LG substitution model, the gamma distribution parameter estimated by PhyML, and a bootstrap value of 100 replicates. The phylogenetic tree was visualized with Dendroscope v3.2.10 with midpoint root (56). The ActR and other XR sequence subtree was extracted and displayed.

Actinorhodopsin enrichment and analysis. BL21(DE3) Tuner E. coli cells cotransformed with a plasmid expressing Pa-CrtEBIY/acI-Blh and acI-ActRL06 were grown as during carotenoid production, except that disodium carbenicillin was also added to 100 g/ml. All further steps were done under red light or darkness at 4°C and/or on ice using 4°C buffers. A sample of 2,000 ODml cells was harvested and suspended in 5 g/ml lysis buffer (50 mM Tris-HCl [pH 8.0], 300 mM NaCl, 1 50-ml EDTA-free protease inhibitor tablet [Roche], 1 mg/ml lysozyme, 20 g/ml DNase I, 5 mM MgCl2, 130 M CaCl2, and 4 mM phenylmethylsulfonyl fluoride [PMSF]).

Cells were lysed at 16,000 lb/in2 by five passes through a French pressure cell. Sequential centrifugation at 10,000 g for 15 min and 100,000 g for 45 min cleared debris and pelleted membranes. Membranes were loosened with lysis buffer and transferred to a Potter-Evelhjem homogenizer. After homogenization, membranes were diluted with 4°C lysis buffer to 25 ml and recentrifuged at 100,000 g for 45 min. Suspension and homogenization were repeated with solubilization buffer (50 mM Tris-HCl [pH 8.0], 300 mM NaCl, 2% mass/volume [m/v] dodecyl-maltoside, 10 mM imidazole). Membranes were diluted to 15 ml with solubilization buffer and rocked for 18 h in darkness. Centrifugation at 20,000 g for 20 min clarified the material before chromatography. The soluble fraction was loaded at 0.5 ml/min onto an equilibrated 1 ml Ni-nitrilotriacetic acid (NTA) column. The processing profile in column volumes (CV) was as follows: 3 solubilization buffer, 12 wash buffer (50 mM Tris-HCl [pH 8.0], 300 mM NaCl, 0.05% m/v DDM, and 30 mM imidazole), 5 elution buffer (50 mM Tris-HCl [pH 8.0], 300 mM NaCl, 0.05% m/v DDM, and 500 mM imidazole). The first elution fraction was dialyzed against 1 liter 4°C final buffer (10 mM HEPES [pH 7.5], 100 mM NaCl, 0.05% m/v) for 2 h and then analyzed by absorbance spectroscopy. Light and 50 µl of 23°C saturated hydroxylamine HCl was used to selectively remove retinal from 140 l of sample to confirm the Schiff base. The resulting unbound retinal oxime absorbs at a wavelength maximum between those of retinal and retinol, where 247- and 257-nm peaks indicate 11-cis and all-trans oxime species, respectively.

Microelectrode pH trace acquisition. The assay was similar to a published protocol and uses E. coli for expression of holo-ActR (9). All steps were carried out under red light generated by putting red cellophane over a Sylvania F20T12/2364 fluorescent bulb. E. coli cells (500 ODml) were harvested by centrifugation at 3,300 g for 15 min at 4°C. Cells were suspended in 45 ml 23°C assay solution (10 mM NaCl, 10 mM MgSO4 · 7H2O, 100 µM CaCl2 · 2H2O) and centrifuged at 3,300 g for 10 min. The latter wash step was repeated. The final suspension was in 20 ml 23°C assay solution to yield an OD of 25. Cells were incubated for 60 min at 23°C to stabilize in darkness before the onset of the assay. The assay was set up with the sample in a

glass test tube surrounded by a 300-ml 23°C water bath in a 400-ml beaker. A Mettler-Toledo InLab microprobe (model number 51343160) was clamped in place above the sample tube and connected to a datalogger (model number 850060; Sper Scientific). The entire setup was surrounded on four sides by foil-lined cardboard. Two FE15T8 bulbs (15 W, 700 lumens each) in an 11 45 cm housing with 90° reflectors were placed 15 cm from the center of the sample tube, such that light illuminated the entire length of the closest tube side. pH was recorded every second after 3.5 min of equilibration time for 60 min as follows: three cycles of 10 min with fluorescent lights off and 10 min with fluorescent lights on. Carbonyl cyanide m-chlorophenyl hydrazone (CCCP) (L06932, lot 10181844; Alfa Aesar) was added to a final concentration of 20 µM during dark stabilization after 45 min.

Heliorhodopsin identification and analysis. No annotations for helio-opsin existed in any acI SAGs (25) or MAGs as helio-opsins had yet to be discovered when these genomes were deposited. Once HeR 48C12 was found, it was used as a homology search model to identify putative acI HeR (14, 28). Candidate protein sequences were used to identify homologs in other acI genomes, and all protein sequences were analyzed by alignment and transmembrane helix predicted as previously mentioned. HeRL06 was cloned as ActRL06 into a pET21b+ vector (Table 2.03). To test expression and activity, this plasmid replaced the ActRL06 plasmid in the same production and cells were subjected to the same micro pH electrode assay to test proton pumping activity. Additionally, the plasmid was induced alone in BL21(DE3) Tuner *E. coli* cells with 1mM IPTG (from 1 M stock in water) and 10µM retinal (from 10 mM stock in ethanol). Specifically, the cells were grown to OD600 of 0.6 at 37°C and 250rpm in half-full flasks. Induction was started and allowed to occur for 3-4 hr, at which point retinal was added. The cells were allowed to shake for an additional 3-4hr and then harvested as all other cells.

Data availability. All genomic and metatranscriptomic sequences are available through the Integrated Microbial Genomes (IMG) and National Center for Biotechnology Information (NCBI) databases, respectively. Genome sequences can be accessed using IMG taxon ID numbers

provided in Table 2.01. The raw RNA sequences can be found in the Sequence Read Archive (SRA) of the National Center for Biotechnology Information under BioProject accession no. PRJNA362825.

References

1. Spudich JL, Yang C-S, Jung K-H, Spudich EN. 2000. Retinylidene proteins: structures and functions from archaea to humans. Annu Rev Cell Dev Biol 16:365–392.

2. Palczewski K. 2006. G Protein–Coupled Receptor Rhodopsin. Annu Rev Biochem 75:743–767.

3. Man D, Wang W, Sabehi G, Aravind L, Post AF, Massana R, Spudich EN, Spudich JL, Béjà O. 2003. Diversification and spectral tuning in marine proteorhodopsins. EMBO J 22:1725–1731.

4. Beja O, Lanyi JK. 2014. Nature's toolkit for microbial rhodopsin ion pumps. Proc Natl Acad Sci 111:6538–6539.

5. Ishchenko A, Round E, Borshchevskiy V, Grudinin S, Gushchin I, Klare JP, Remeeva A, Polovinkin V, Utrobin P, Balandin T, Engelhard M, Büldt G, Gordeliy V. 2017. New Insights on Signal Propagation by Sensory Rhodopsin II/Transducer Complex. Sci Rep 7: 41811.

6. Pinhassi J, DeLong EF, Béjà O, González JM, Pedrós-Alió C. 2016. Marine Bacterial and Archaeal Ion-Pumping Rhodopsins: Genetic Diversity, Physiology, and Ecology. Microbiol Mol Biol Rev 80:929–954.

7. Subramaniam S. 1999. The structure of bacteriorhodopsin: an emerging consensus. Curr Opin Struct Biol 9:462–468.

8. Balashov SP. 2005. Xanthorhodopsin: A Proton Pump with a Light-Harvesting Carotenoid Antenna. Science 309:2061–2064.

9. Keffer JL, Hahn MW, Maresca JA. 2015. Characterization of an Unconventional Rhodopsin from the Freshwater *Actinobacterium* Rhodoluna lacicola. J Bacteriol 197:2704–2712.

10. Lanyi JK. 2006. Proton transfers in the bacteriorhodopsin photocycle. Biochim Biophys Acta BBA - Bioenerg 1757:1012–1018.

11. Edman K, Nollert P, Royant A, Pebay-Peyroula E. 1999. High-resolution X-ray structure of an early intermediate in the bacteriorhodopsin photocycle 401: 822-6.

12. Nakamura S, Kikukawa T, Tamogami J, Kamiya M, Aizawa T, Hahn MW, Ihara K, Kamo N, Demura M. 2016. Photochemical characterization of actinorhodopsin and its functional existence in the natural host. Biochim Biophys Acta 1857:1900–1908.

13. Kirchman DL, Hanson TE. 2013. Bioenergetics of photoheterotrophic bacteria in the oceans. Environ Microbiol Rep 5:188–199.

14. Pushkarev A, Inoue K, Larom S, Flores-Uribe J, Singh M, Konno M, Tomida S, Ito S, Nakamura R, Tsunoda SP, Philosof A, Sharon I, Yutin N, Koonin EV, Kandori H, Béjà O. 2018. A distinct abundant group of microbial rhodopsins discovered using functional metagenomics. Nature 558:595–599.

15. Luecke H, Schobert B, Stagno J, Imasheva ES, Wang JM, Balashov SP, Lanyi JK. 2008. Crystallographic structure of xanthorhodopsin, the light-driven proton pump with a dual chromophore. Proc Natl Acad Sci 105:16561–16565.

16. Ozaki Y, Kawashima T, Abe-Yoshizumi R, Kandori H. 2014. A Color-Determining Amino Acid Residue of Proteorhodopsin. Biochemistry 53:6032–6040.

17. Balashov SP, Imasheva ES, Choi AR, Jung K-H, Liaaen-Jensen S, Lanyi JK. 2010. Reconstitution of Gloeobacter Rhodopsin with Echinenone: Role of the 4-Keto Group. Biochemistry 49:9792–9799.

18. Iniesta AA, Cervantes M, Murillo FJ. 2008. Conversion of the lycopene monocyclase of Myxococcus xanthus into a bicyclase. Appl Microbiol Biotechnol 79:793–802.

19. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403–410.

20. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes (ed. F. Cohen). J Mol Biol 305:567–580.

21. Arrach N, Fernández-Martín R, Cerdá-Olmedo E, Avalos J. 2001. A single gene for lycopene cyclase, phytoene synthase, and regulation of carotene biosynthesis in Phycomyces. Proc Natl Acad Sci 98:1687–1692.

22. Kim Y-S, Kim N-H, Yeom S-J, Kim S-W, Oh D-K. 2009. *In Vitro* Characterization of a Recombinant Blh Protein from an Uncultured Marine Bacterium as a β-Carotene 15,15′-Dioxygenase. J Biol Chem 284:15781–15793.

23. Hamilton JJ, Garcia SL, Brown BS, Oyserman BO, Moya-Flores F, Bertilsson S, Malmstrom RR, Forest KT, McMahon KD. 2017. Metabolic Network Analysis and Metatranscriptomics Reveal Auxotrophies and Nutrient Sources of the Cosmopolitan Freshwater Microbial Lineage acI. MSystems 2:e00091–17.

24. Kang I, Kim S, Islam MR, Cho J-C. 2017. The first complete genome sequences of the acI lineage, the most abundant freshwater *Actinobacteria*, obtained by whole-genome-amplification of dilution-to-extinction cultures. Sci Rep 7:42252.

25. Ghylin TW, Garcia SL, Moya F, Oyserman BO, Schwientek P, Forest KT, Mutschler J, Dwulit-Smith J, Chan L-K, Martinez-Garcia M, others. 2014. Comparative single-cell

genomics reveals potential ecological niches for the freshwater acI *Actinobacteria* lineage. ISME J 8:2503-16.

26. Jang H-J, Ha B-K, Zhou J, Ahn J, Yoon S-H, Kim S-W. 2015. Selective retinol production by modulating the composition of retinoids from metabolically engineered E. coli. Biotechnol Bioeng 112:1604–1612.

27. Rodriguez GM, Atsumi S. 2014. Toward aldehyde and alkane production by removing aldehyde reductase activity in Escherichia coli. Metab Eng 25:227–237.

28. Singh M, Inoue K, Pushkarev A, Béjà O, Kandori H. 2018. Mutation Study of Heliorhodopsin 48C12. Biochemistry 57:5041–5049.

29. Newton RJ, McMahon KD. 2011. Seasonal differences in bacterial community composition following nutrient additions in a eutrophic lake: Seasonal variation in nutrient-amended microcosms. Environ Microbiol 13:887–899.

30. Hall MW, Rohwer RR, Perrie J, McMahon KD, Beiko RG. 2017. Ananke: temporal clustering reveals ecological dynamics of microbial communities. PeerJ 5:e3812.

31. Berry MA, Davis TW, Cory RM, Duhaime MB, Johengen TH, Kling GW, Marino JA, Den Uyl PA, Gossiaux D, Dick GJ, Denef VJ. 2017. Cyanobacterial harmful algal blooms are a biological disturbance to Western Lake Erie bacterial communities. Environ Microbiol 19:1149–1162.

32. Marasco EK, Vay K, Schmidt-Dannert C. 2006. Identification of Carotenoid Cleavage Dioxygenases from *Nostoc* sp. PCC 7120 with Different Cleavage Activities. J Biol Chem 281:31583–31593.

33. Ahrazem O, Gómez-Gómez L, Rodrigo MJ, Avalos J, Limón MC. 2016. Carotenoid Cleavage Oxygenases from Microbes and Photosynthetic Organisms. Int J Mol Sci. Oct 17: 1781-1819

34. Wu X, Jiang J, Hu J. 2013. Determination and Occurrence of Retinoids in a Eutrophic Lake (Taihu Lake, China): Cyanobacteria Blooms Produce Teratogenic Retinal. Environ Sci Technol 47:807–814.

35. Neuenschwander SM, Ghai R, Pernthaler J, Salcher MM. 2018. Microdiversification in genome-streamlined ubiquitous freshwater *Actinobacteria*. ISME J 12:185–198.

36. Markowitz VM, Chen I-MA, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Jacob B, Huang J, Williams P, Huntemann M, Anderson I, Mavromatis K, Ivanova NN, Kyrpides NC. 2012. IMG: the integrated microbial genomes database and comparative analysis system. Nucleic Acids Res 40:D115–D122.

37. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33).

38. Magoc T, Salzberg SL. 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27:2957–2963.

39. Kopylova E, Noé L, Touzet H. 2012. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics 28:3211–3217.

40. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2012. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 41:D590–D596.

41. Anders S, Pyl PT, Huber W. 2015. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics 31:166–169.

42. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5:621–628.

43. Breemen RBV, Huang CR. 1996. High-performance liquid chromatography-electrospray mass spectrometry of retinoids. FASEB J 10:1098–1101.

44. Clasquin MF, Melamud E, Rabinowitz JD. 2012. LC-MS Data Processing with MAVEN: A Metabolomic Analysis and Visualization Engine. Baxevanis, AD, Petsko, GA, Stein, LD, Stormo, GD (eds.), Current Protocols in Bioinformatics. John Wiley & Sons, Inc., Hoboken, NJ, USA.

45. Melamud E, Vastag L, Rabinowitz JD. 2010. Metabolomic Analysis and Visualization Engine for LC−MS Data. Anal Chem 82:9818–9826.

46. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Soding J, Thompson JD, Higgins DG. 2014. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7:539–539.

47. McWilliam H, Li W, Uludag M, Squizzato S, Park YM, Buso N, Cowley AP, Lopez R. 2013. Analysis Tool Web Services from the EMBL-EBI. Nucleic Acids Res 41:W597–W600.

48. Li W, Cowley A, Uludag M, Gur T, McWilliam H, Squizzato S, Park YM, Buso N, Lopez R. 2015. The EMBL-EBI bioinformatics web and programmatic tools framework. Nucleic Acids Res 43:W580–W584.

49. Roy A, Kucukural A, Zhang Y. 2010. I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc 5:725–738.

50. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. 2015. The I-TASSER Suite: protein structure and function prediction. Nat Methods 12:7–8.

51. Zhang Y. 2008. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics 9:40.

52. Schrödinger, LLC. The PyMOL Molecular Graphics System, V2.0.0. Schrödinger, LLC, New York, NY.

53. Pei J, Kim B-H, Grishin NV. 2008. PROMALS3D: a tool for multiple protein sequence and structure alignments. Nucleic Acids Res 36:2295–2300.

54. Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17:540–552.

55. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59:307–321.

56. Huson DH, Scornavacca C. 2012. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. Syst Biol 61:1061–1067.

Figures and Tables

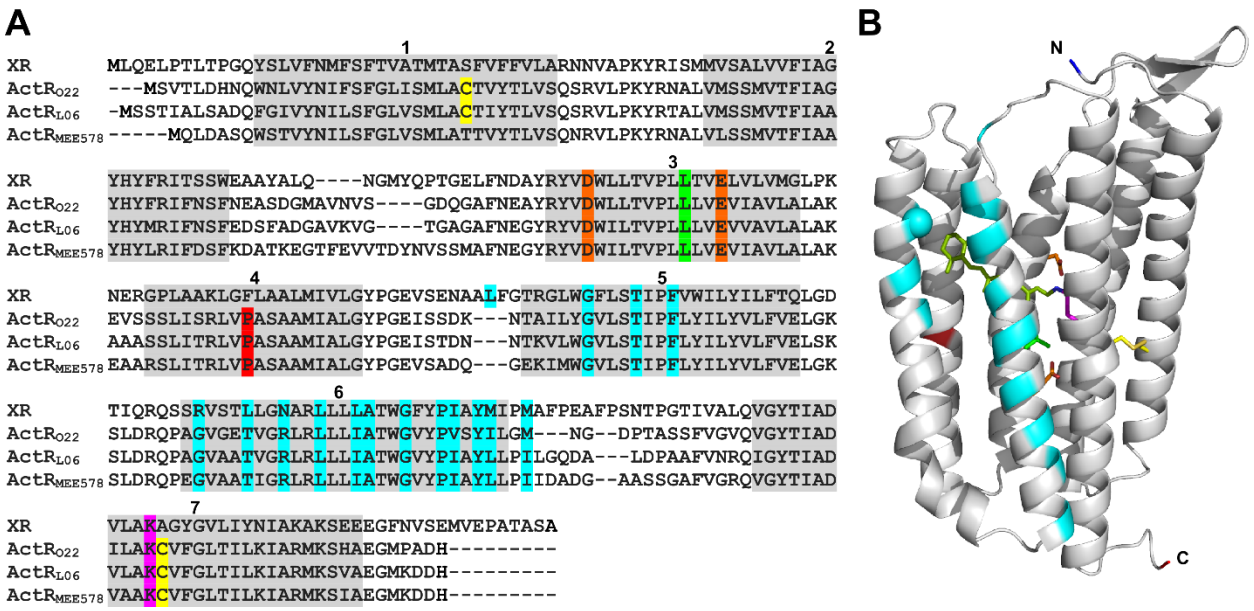


FIG 2.01 Features of ActRs from acI clades A, B, and C. (A) ActRs from clade A (ActRO22), clade B (ActRL06), and clade C (ActRMEE578) are aligned to a homolog, xantho-opsin (XR) from *Salinibacter ruber*. Proteins are subscripted with genome shorthand (Table 2.01). Features are highlighted: predicted helices (gray, numbered), cysteines (yellow), main proton shuttles (orange), main absorbance tuner (green), Schiff base lysine (magenta), possible antenna carotenoid residues (cyan), and proline indicative of ActR versus xantho-opsins (red). Antenna carotenoid residues are based on amino acids in proximity to salinixanthin in the crystal structure of xanthorhodopsin (PDB 3DDL). (B) Three-dimensional (3D) structure prediction for ActR with key residue side chains displayed according to the color coding in panel A. The alpha-carbon of glycine near the top of α5 that would allow antenna binding is shown as a sphere, the Schiff base retinal (lime green) and predicted disulfide bond in clades A and B are modeled, and the N and C termini are labeled.

FIG 2.02 Xantho-opsin phylogeny groupings provide clear evidence for a distinct grouping of action-opsins. A conserved proline in α4 differentiates actino-opsins (ActRs) from other xantho-opsins (XRs) in the supergroup. Bolded opsins are aligned in Fig. 2.01A.

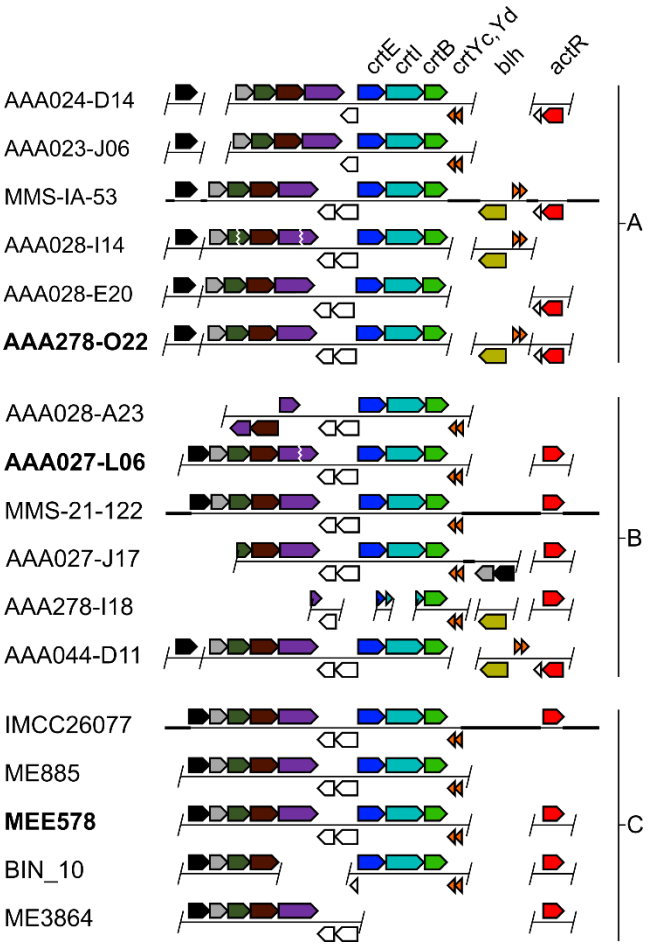


FIG 2.03 Rhodopsin-and-retinal-related genes in the genomic context. acI genomic contigs or genomes from clades A, B, and C are primarily labeled by shorthand notation (i.e., designation after "*Actinobacterium* SCGC" or "int.metabat."). Genes are arrows pointing in the direction of transcription. A slash indicates a contig boundary, which may or may not end immediately after the pictured gene, a vertical zig-zag indicates contig ends that have been manually paired, and a thick horizontal bar indicates a longer region of DNA not represented. Uncolored genes are neighboring genes that may or may not be functionally associated with the carotenoid-related genes. Boldface labels indicate a gene source for this study. Relevant Integrated Microbial Genomes (IMG) database locus tags for this study from AAA278-O22 are A278O22DRAFT_00010530-10510 (crtE to crtB), A278O22DRAFT_00007000-6980 (crtYd to blh), and A278O22DRAFT_00001360 (actR).

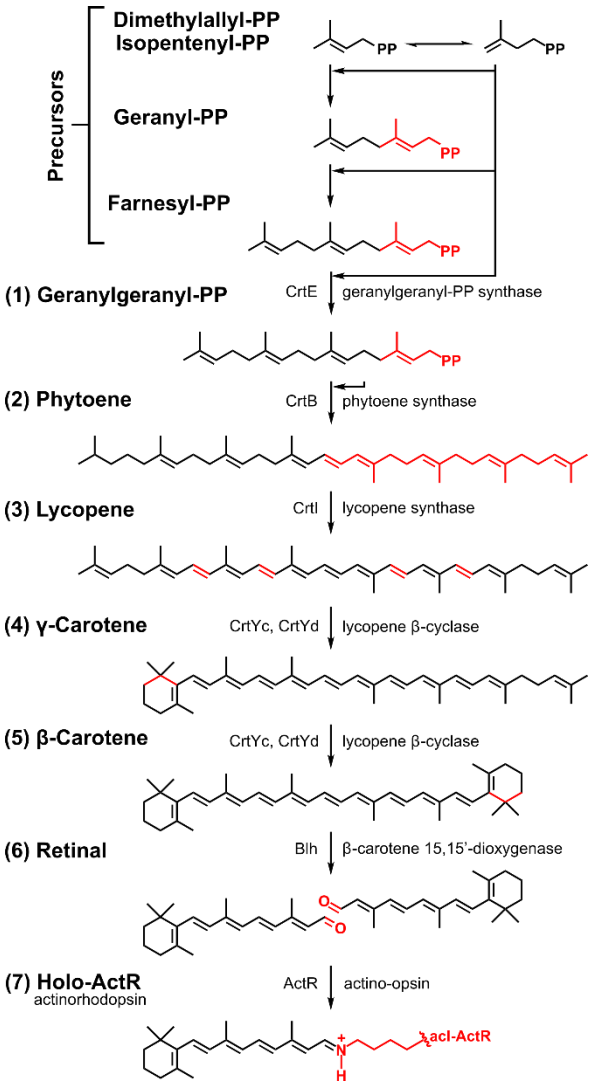


FIG 2.04 Predicted rhodopsin-and-retinal-related pathways in acI. The actinorhodopsin synthesis pathway requires isopentenyl precursors to be assembled into carotenoids (1 to 3), which are modified (4 to 5) and cleaved to produce retinal (6). A Schiff base forms between retinal and a lysine of acI-ActR to form acI-holo-ActR (7). Chemical changes are shown in red, proteins from this study with their predicted functions are shown to the sides of progress arrows, and product names are shown to left of their chemical structures. Step 6 cannot be performed in acI *Actinobacteria* that lack Blh; retinal must be exogenously sourced.

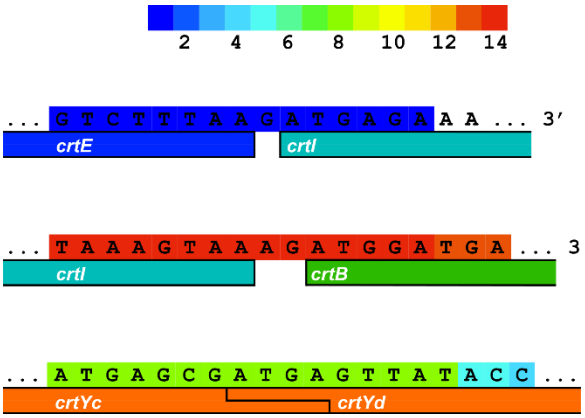


FIG 2.05 Intergenic transcripts that map to rhodopsin-and-retinal-related genes in genome ME885. Operons are evident for lycopene (*crtE* through *crtB*) and lycopene cyclase (*crtYc* and *crtYd*) biosynthesis in an acI-C member. The color key indicates of the number of times a base was covered. Transcripts and genes may continue beyond the edges of the DNA window.

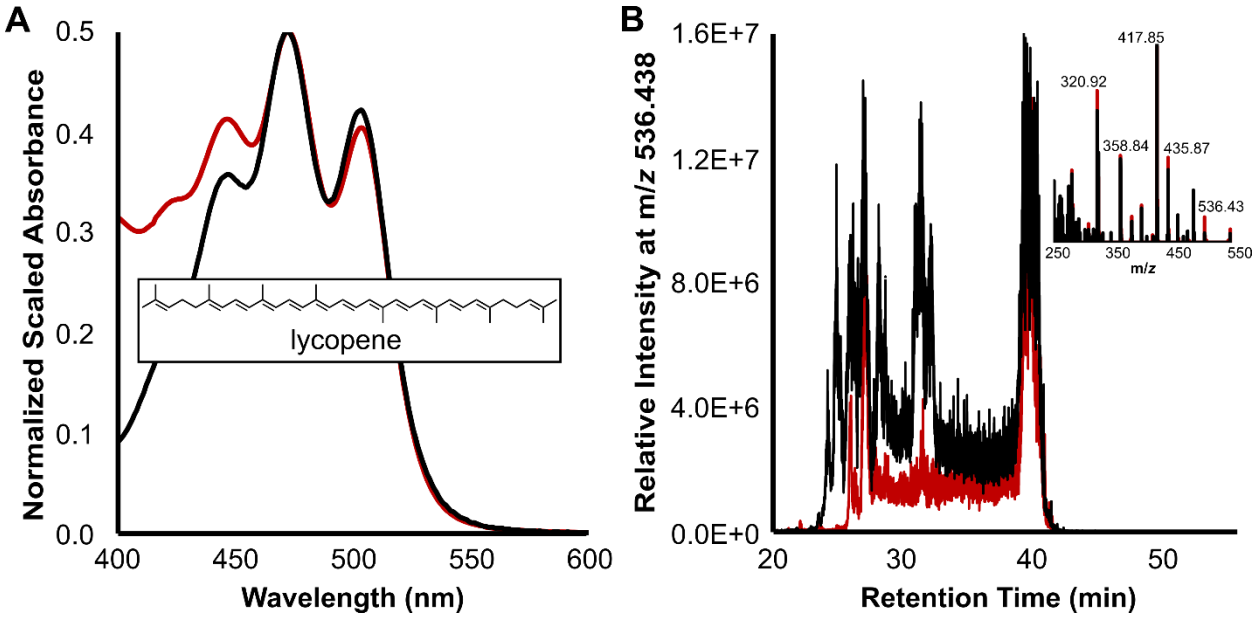


FIG 2.06 CrtE, CrtB, and CrtI from acI members catalyze lycopene formation. (A) Spectra of a control extract subtracted from an extract from cells containing acI enzymes (CrtE, CrtB, CrtI) (black) and standard lycopene (red). All extracts are in acetone. The inset depicts lycopene, and maxima for both curves are at 447, 472, and 504 nm. (B) HPLC-MS data for the same samples. The inset depicts MS/MS data from all-trans peak retention times.

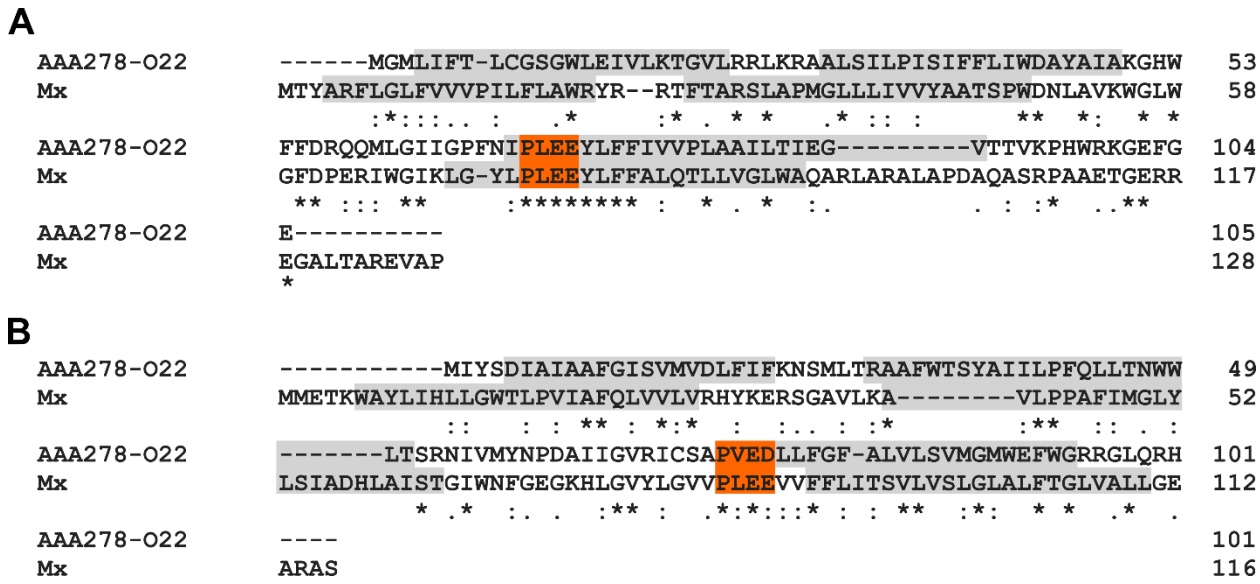


FIG 2.07 Alignments for lycopene cyclase chains (acI-CrtYc and acI-CrtYd). Proteins from acI genome AAA278-O22 are aligned to CrtYc (A) and CrtYd (B) from *Myxococcus xanthus* (Mx), which form the functionally characterized heterodimeric β-cyclase. In this and following five figures, identical (*), highly similar (:), and slightly similar (.) residues are indicated, and transmembrane helix predictions are highlighted in gray. The catalytic PxE(E/D) motif is highlighted in orange.

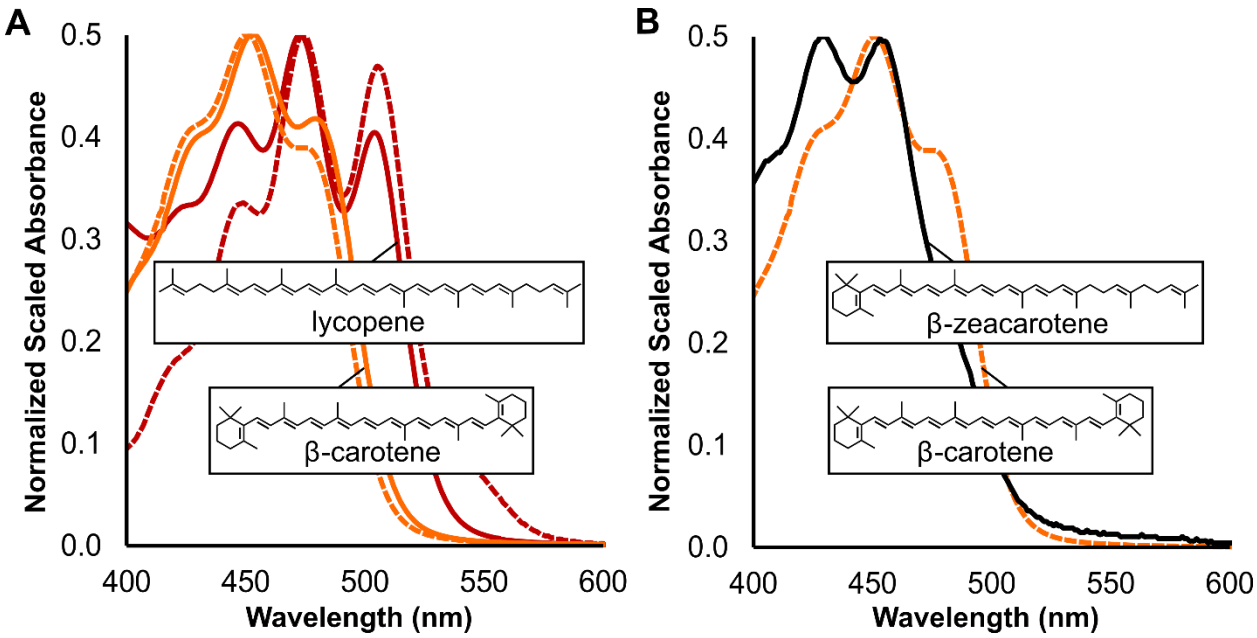


FIG 2.08 A lycopene-like carotenoid produced by CrtE, CrtB, and CrtI from acI members is cyclized by CrtY from *Pantoea ananatis*. (A) Spectra of a control extract subtracted from an extract from cells containing *Pantoea ananatis* enzymes (CrtE, CrtB, and CrtI; red dash), standard lycopene (red), a control extract subtracted from an extract from cells containing *Pantoea ananatis* enzymes (CrtE, CrtB, CrtI, and CrtY; orange dash), and standard -carotene (orange). All extracts are in acetone. The insets depict lycopene and -carotene, and the maxima are 472/474 nm and 453/451 nm. (B) The spectrum for synthesized -carotene as in panel A (orange dash), and the spectrum of a control extract subtracted from an extract from cells containing acI enzymes (CrtE, CrtB, and acI-CrtI) and CrtY from *Pantoea ananatis*. All extracts are in acetone. The insets depict -zeacarotene and -carotene, and the shifted maxima are at 429 and 453 nm.

```
Umb66A03        ---------------------------------MGLMLIDWCALA--LVVFIGLPHGALD   25
AAA278-O22      MEMAKLKTFSRVRTFSSAIVAVAIALSIVFSSWLGVDSLNWQVVMAVVALAIGIPHGALD   60
                                                :*:  ::* .:   :.: **:******
Umb66A03        AAISFSMISSAKRIARLAGILLIYLLLATAFFLIWYQ--LPAFSLLIFLLISIIHFGMAD   83
AAA278-O22      HLVTLPKAQPLKM----AIFIAIY--VAIALIAIWAILQWNVWGFIAVVIMSATHFGIGD  114
                  :::   .  *     * :: **  :* *:: **      .:.:: .:::*  ***:.*
Umb66A03        FNASPSKL----------KWPHIIAHGGVVTVWLPLIQKNEVTKLFSILTNGPTPILWDI  133
AAA278-O22      SAFISELNRLKGIQSHLPIWAYA-PAAGALPVVIPLVNSRSTDALQKV---NSELINWHH  170
                      .              * :    .*.: * :**::....  * .:   .   * *.
Umb66A03        LLIFFL--CWSIGVCLHTYETLRSKHYNIAFELIGLIFLAWYAPPLVTFATYFCFIHSRR  191
AAA278-O22      GYTSELQIAVAVIATLSAMTLLSKKRYRDLLDVALLAALASVAPPLVAFAVYFGCWHAMR  230
                     *  . :: . * :   * .*:*.  :::  *  **  *****:**.**   *: *
Umb66A03        HFSFVWKQLQHMSSKKMMIGSA------IILSCTSWLIGGGIYFFLNSKMI---ASEAAL  242
AAA278-O22      HTARLSSLLPRSLA-AYEAGNSWQAFRSAVIPGLPALIGTLIFVALLAGFSHNNVSDSFL  289
                * : : . * :  :     *.:       ::     ***  *:. * : :    .*:: *
Umb66A03        QTVFIGLAALTVPHMILIDFIFRPHSSRIKIKN-----                        275
AAA278-O22      WLTLVTIWALTVPHMMVTAKLDR-----AALKNKSHLR                        322
                  .:: : *******::   : *       :**
```

FIG 2.09 Alignment for 15-15'-β-carotene dioxygenase (acI-Blh). The protein from acI genome AAA278-O22 is aligned to a functionally characterized 15-15'-β-carotene dioxygenase from Uncultured Marine Bacterium 66A03 (Umb66A03). The histidines indicated for non-heme iron binding are highlighted in orange.

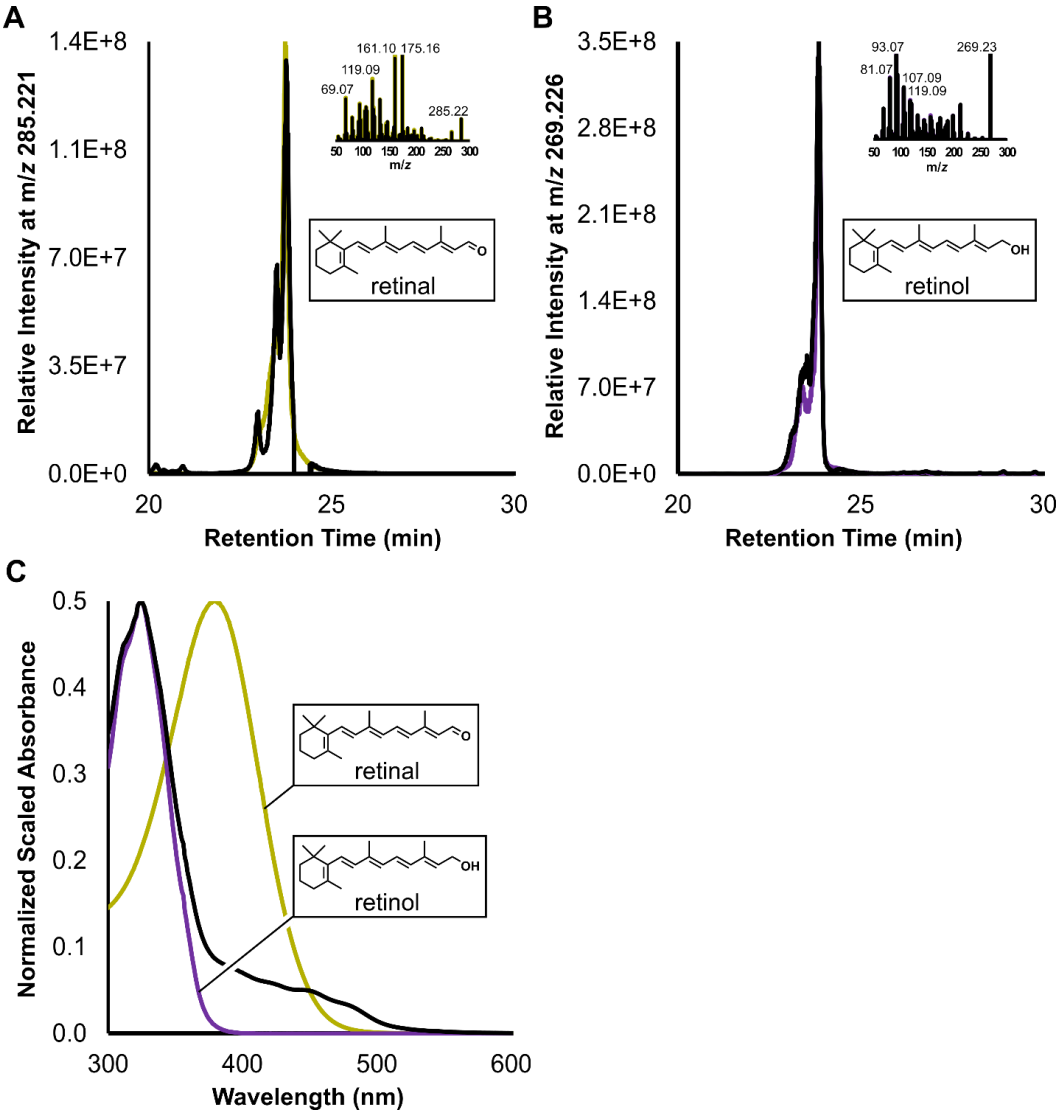


FIG 2.10 acI-Blh is a -carotene oxygenase that catalyzes retinal formation. (A) HPLC-MS data for Pa-EBIY/ acI-Blh extract (black) and a retinal standard (yellow). The inset depicts MS/MS data from all-trans peak retention times. (B) HPLC-MS and MS/MS (inset) data for the same experimental extract but with standard retinol spectra (purple). Retinol forms a dehydrated species in positive mode, and the species was used as a proxy for retinol detection (C) Spectra of a control extract subtracted from an extract from cells containing *Pantoea ananatis* enzymes (CrtE, CrtB, CrtI, and CrtY) and Blh from acI (black), standard retinal (yellow), and standard retinol (purple). All extracts are in ethanol. Insets depict retinal and retinol to highlight conjugation differences, and important maxima are at 379, 324, and 325 nm.

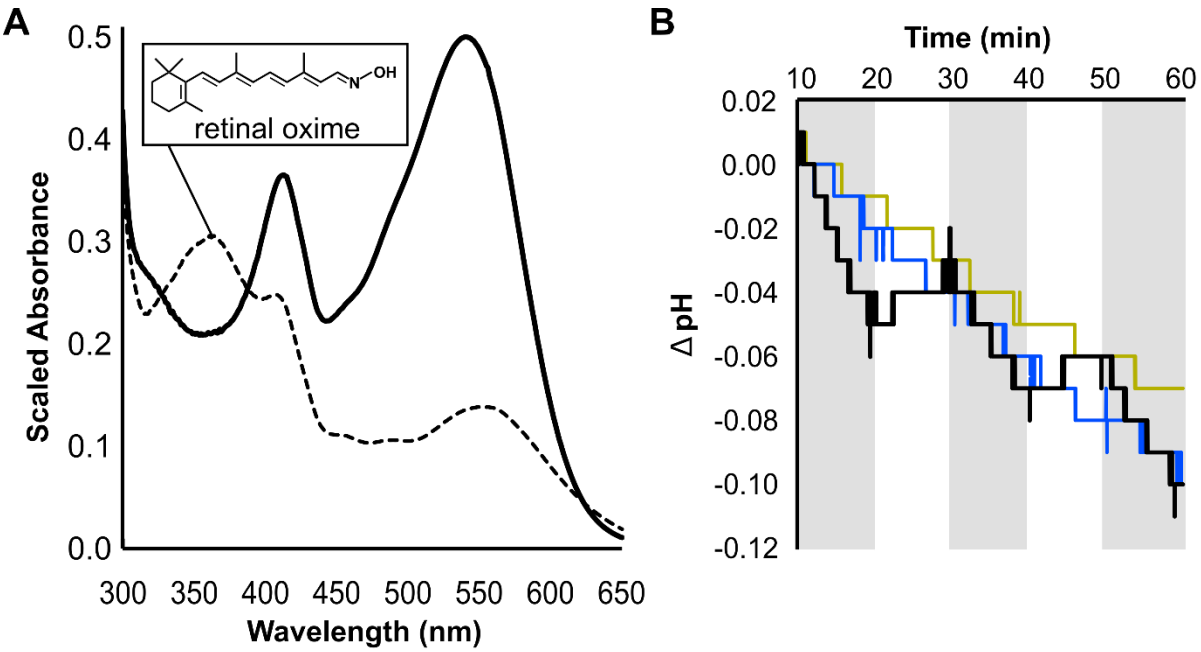


FIG 2.11 holo-ActRL06 is a light-driven proton pump. (A) Spectra for a holo-ActRL06 enrichment purification from cells also producing retinal using *Pantoea ananatis* CrtE, CrtB, and CrtI, and Blh from acI members without (solid) and with (dashed) hydroxylamine hydrochloride. The retinal oxime that results from hydroxylamine incubation is shown, and the maximum absorbance is at 358 nm. (B) Microelectrode pH traces for retinal-producing cells expressing as in panel A with (black, blue) or without (yellow) acI-ActRL06. CCCP is present in the blue trace. The room was illuminated by a white light with a red plastic overlay for the entire assay, and shaded regions represent extra direct illumination by white light. Curves are from one of three independent experiments.

```
48C12   MAKPTVKEIKSLQNFNRIAGVFHLLQMLAVLALANDFALPMTGTYLNGPPGTTFSAPVVI
acI     MSKPI--TATSLRKVNIYAGVLHLAQMIAVLALSSDFTLPINATYMSGPPGTTYAAPVTL
        *:**     .**::.*  ***:** **:*****:.**:**:..**:.******::***.:

        LETPVGLAVALFLGLSALFHFIVSSGNFFKRYSASLMKNQNIFRWVEYSLSSSVMIVLIA
        FETPIGLTVAIFLGLSSIAHFIVASPKFFPRYSAGLAAKRNYFRWVEYSISSSVMIVLIA
        :***:**:**:*****:: ****:* :** ****.*  ::* *******:**********

        QICGIADIVALLAIFGVNASMILFGWLQEKYTQPKDGDLLPFWFGCIAGIVPWIGLLIYV
        QVTGVSDITAIISIFGVNASMILFGWLQEKYENPGSGGWLPYIFGCITGIIPWLALCFYV
        *: *::**.*:::****************** :* .*. **: ****:**:**:.* :**

        IAPGSTSDVAVPGFVYGIIISLFLFFNSFALVQYLQYKGKGKWSNYLRGERAYIVLSLVA
        FGIGGAGETKAPTFVYVVVLTIFLFFNSFALVQFLQYKMVGKWSDYLRGERTYITLSLIA
        :. *.:.:. .* *** :::::*********:****  ****:******:**.***:*

        KSALAWQIFSGTLIPA-
        KSALAWQIFANTLIPPV
        *********:.****
```

FIG 2.12 Alignment for helio-opsin (acI-HeR). An HeR from acI-B (ActRL06) is aligned to the only characterized HeR from an unknown freshwater organism, which islikely *Actinobacterial* and designated 48C12. The Schiff base lysine for retinal attachment is highlighted in yellow.

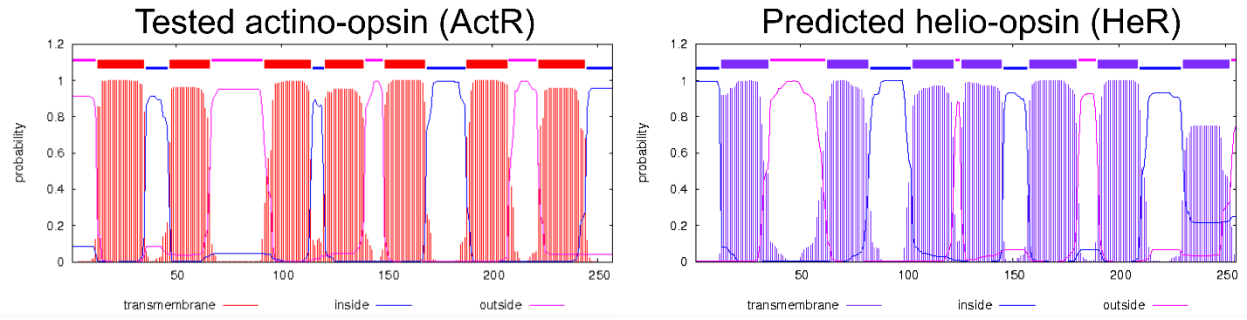


FIG 2.13 Helio-opsin from acI has an N-in, C-out topology. A functionally characterized ActR from acI-B (ActRL06) and the HeR from the same genome are both predicted to have seven helices (thick lines) but reversed topology (thin lines – inside (blue), outside (pink)). TMHMM Server v. 2.0 was used for transmembrane prediction and image generation.

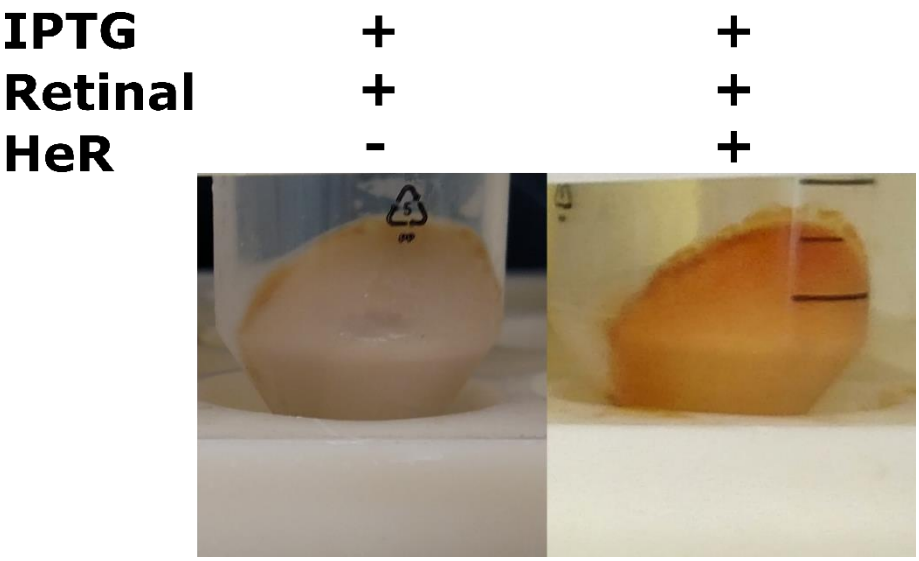


FIG 2.14 HeRL06 (helio-opsin) appears to bind retinal to form holo-HeRL06 (heliorhodopsin). Cells overexpressing HeR from acI-B (AAA-027L06) appear red-orange in the presence of retinal.

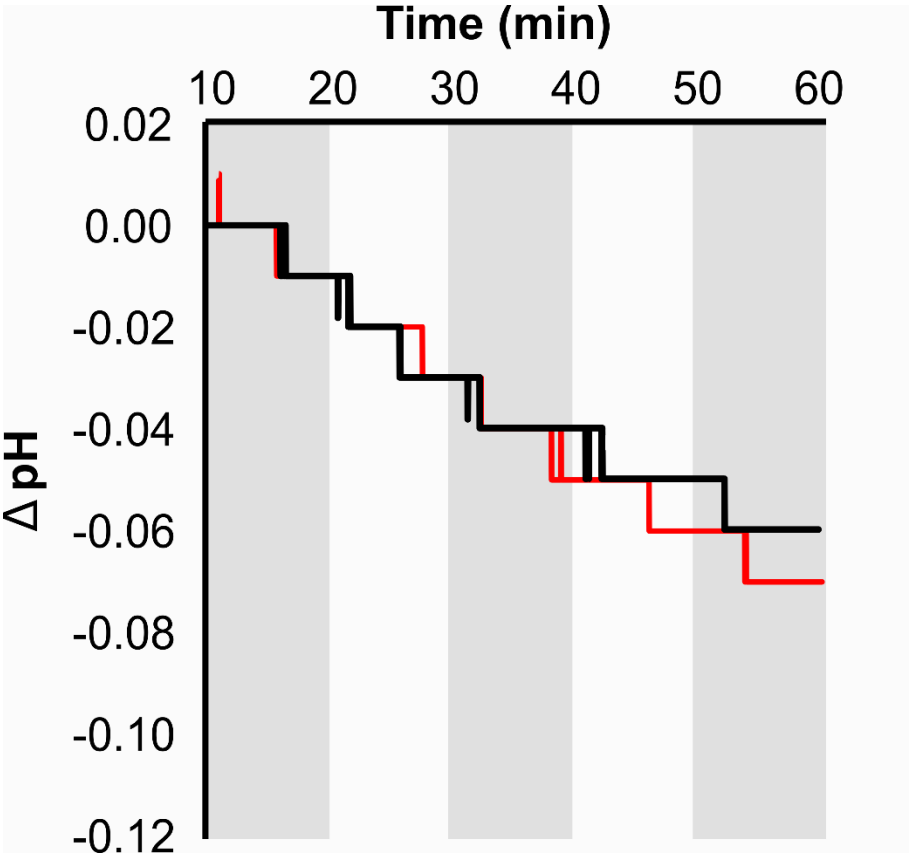


FIG 2.15 holo-HeRL06 is not a light-driven proton pump. Microelectrode pH traces for retinal-producing cells with (black) or without (red) acI-HeRL06. The room was illuminated by a white light with a red plastic overlay for the entire assay, and shaded regions represent extra direct illumination by white light. Curves are from one of three independent experiments.

TABLE 2.01 Metadata of acI genomes referred to in this study

Classification	Genome Name	IMG Taxon ID	Recovered Genome Size (MB)	Contigs	Genes	Completeness (%)[i]	Contamination (%)[i]
acI-A	AAA024-D14[a]	2264265190	0.78	82	892	58.4	0.98
	AAA023-J06[a]	2236661001	0.70	98	818	43.7	0.00
	MMS-IA-53[b]	2775506795	1.33	1	1386	100	0.00
	AAA028-I14[c]	2545555832	0.78	54	848	50.8	0.00
	AAA028-E20[c]	2602042079	0.88	51	953	60.2	0.00
	AAA278-O22[a]	2236661007	1.14	43	1238	78.2	0.42
acI-B	AAA028-A23[a]	2236661004	0.83	64	913	67.7	0.19
	AAA027-L06[d]	2505679121	1.16	75	1282	87.4	4.37
	MMS-21-122[b]	2775506799	1.24	1	1281	100	0.00
	AAA027-J17[a]	2236661002	0.97	81	1094	73.1	0.84
	AAA278-I18[a]	2236661006	0.94	54	1037	69.0	0.17
	AAA044-D11[c]	2524023189	1.15	30	1214	75.2	0.84
acI-C	IMCC26077[e]	2602042021	1.52	1	1572	100.	0.00
	ME885[c]	2582580608	0.87	78	906	55.7	4.76
	ME578[c]	2556921504	0.36	32	390	25.9	3.45
	BIN_10[g]	NA[h]	0.98	95	1171	75.7	0.00
	ME3864[c]	2582580572	0.79	70	839	53.0	3.78

[a] Ghylin, T. W. et al (2014). Comparative single-cell genomics reveals potential ecological niches for the freshwater acI *Actinobacteria* lineage. The ISME Journal, 8(12), 2503–2516. Single amplified genomes (SAGs) recovered from one of three lakes including Lake Mendota, using fluorescence activated cell sorting, subsequent multiple displacement amplification (MDA) followed by sequencing and assembly. Additional details are available in the original publication.

[b] Neuenschwander S.M., Ghai R., Pernthaler J., Salcher M. M. (2018) Microdiversification in genome-streamlined ubiquitous freshwater *Actinobacteria*. The ISME Journal, 12(1), 185-198. This genome is a MAG derived a dilution culture enriched in acI-B cells. The DNA was obtained using MDA products that were sequenced as a metagenome, assembled, and binned manually with Sanger sequencing to close gaps. Additional details are available in the original publication.

[c] Hamilton J. J. et al. 2017. Metabolic network analysis and metatranscriptomics reveal auxotrophies and nutrient sources of the cosmopolitan freshwater microbial lineage acI. mSystems 2:e00091-17. Note that the names in the IMG record are actually MEint.metabat.885,

MEint.metabat.578, and MEint.3864, respectively. MAGs were recovered from combined assemblies (i.e. bioinformatically pooled metagenomes from 94 samples collected from the top 12 m of Lake Mendota at the deep hole) and binned using differential coverage based on metagenomes mapped to the combined assembly and based on k-mer frequencies. Details can be found in the original publication and S. Roux, L-K.Chan, R. Egan, R. R. Malmstrom, K. D. McMahon, M. B. Sullivan (2017) “Ecogenomics of freshwater virophages and their giant virus hosts assessed through time-series metagenomics”. *Nature Communications*. 8: article 858.

[d] Garcia, S. L. et al. (2013). Metabolic potential of a single cell belonging to one of the most abundant lineages in freshwater bacterioplankton. The ISME Journal, 7(1), 137–147. This genome was a SAG recovered from Lake Stechlin in Germany, using fluorescence activated cell sorting, subsequent multiple displacement amplification (MDA) followed by sequencing and assembly. Additional details are available in the original publication of Garcia et al 2013.

[e] Kang I., Kim S., Islam M. R., Cho J-C. (2017). The first complete genome sequences of the acI lineage, the most abundant freshwater *Actinobacteria*, obtained by whole-genome-amplification of dilution-to-extinction cultures. Sci Rep 7. This genome is a MAG derived a dilution culture enriched in acI-C cells. The DNA was obtained using MDA products that were sequenced as a metagenome, assembled, and binned manually with Sanger sequencing to close gaps. Additional details are available in the original publication of Kang et al 2017.

[g] Tsementzi, D., Poretsky, R. S., Rodriguez-R, L. M., Luo, C., & Konstantinidis, K. T. (2014). Evaluation of metatranscriptomic protocols and application to the study of freshwater microbial communities. Environmental Microbiology Reports, 6(6), 640–655.

[h] Data are available at http://enve-omics.ce.gatech.edu/data/MTR

[i] Inferred using CheckM and the set of 204 *Actinobacteria*-specific marker genes or set to 100% and 0% because genome is listed as complete in the original publication.

TABLE 2.02 Log2 RPKM values for pooled carotenoid-related gene transcripts within each clade.

	Log2 RPKM Values		
Gene Name	**acI-A**	**acI-B**	**acI-C**
crtE	3.69	4.93	6.48
crtB	2.74	3.45	6.62
crtI	4.33	5.93	6.90
crtYc	1.57	5.62	7.07
crtYd	ND[a]	3.29	4.02
blh	0.07	2.06	NA[b]
actR	14.36	13.86	12.52
recA	7.85	9.16	10.09

[a] ND, not determined; gene is present in clade member assemblies but no transcript mapped.

[b] NA, not applicable; gene is not present in any clade member assembly.

TABLE 2.03 DNA sources and utilized plasmids

	Gene Vector	Genes	*E. coli*-codon-optimized?
DNA Sources	Single-cell genome	*acI-crtE, acI-crtB, acI-crtI*	No
		acI-blh	No
	IDT gBlock		
	HeR L06	*acI-heR$_{L06}$*	Yes
	Plasmid		
	pJExpress404	*acI-actR$_{L06}$*	Yes
	Bba_K274210	*Pa-crtE,Pa-crtB,Pa-crtI,Pa-crtY*	Yes
	Backbone	**Cassette Shorthand**	**Cloning Site Contents**
Utilized Plasmids	pCDFDuet1[a]	-/-	NA / NA
		acI-CrtEBI/-	*acI-crtE, acI-crtB, acI-crtI* / NA
		acI-CrtEBI/Pa-CrtY	*acI-crtE, acI-crtB, acI-crtI* / *Pa-crtY*
		Pa-CrtEBI/-	*Pa-crtE, Pa-crtB, Pa-crtI* / NA
		Pa-CrtEBIY/-	*Pa-crtE, Pa-crtB, Pa-crtI, Pa-crtY* / NA
		Pa-CrtEBIY/acI-Blh	*Pa-crtE, Pa-crtB, Pa-crtI, Pa-crtY* / *acI-blh*
	pET21b+[b]	-	NA
		acI-HeR$_{L06}$	*acI-heR$_{L06}$*
		acI-ActR$_{L06}$	*acI-actR$_{L06}$*

[a] Pa, *Pantoea ananatis*

[b] Plasmid contains two cloning sites; contents of each are written, separated by a slash

[c] Plasmid contains one cloning site in-frame with a C-terminal hexahistidine tag

[d] -, empty.

[e] NA, not applicable because the cloning site is unmodified.

TABLE 2.04 Primers used[a]

	Name	Sequence (5'-3')
Genomic	F_EIB	ATGCAGAGCGCATCAGAGGT
	R_EIB	TTAGCCAATTCGTGCTGCAA
	F_YcYdBlh	TTAGTGTCTTTGCAGCCCT
	R_YcYdBlh	CTAACGAAGGTGCGATTTAT
	Name	Sequence (5'-3')
CrtEBI/- CrtEBI/PaY	F_E_NcoI	aaataattttgtttaactttaataaggagatataccatggATGCAGAGCGCATCAGAGGT
	R_E_AflII	tgtacaatacgattactttctgttcgacttaagTTAGTAGCTCCTACGGATTGCTGTCTC
	F_B_Erbs	agcaatccgtaggagctactaacaagaaggagatatacttATGGATGCAGAACTTGCAGC
	R_B_EcoAfl	cgtgtacaatacgattactttctgttcgacttaagaattcttaGCCAATTCGTGCTGCAA
	F_E/-_BEcoAfl	attgcagcacgaattggctaagaattcTTAAGTCGAACAGAAAGTAATCGTATTGTACAC
	R_E/-_ErbsB	caagttctgcatccataagtatatctccttcttGTTAGTAGCTCCTACGGATTGCTGTCT
	F_I_Brbs	gcagcacgaattggctaagaagaaggagatatacttATGACAAGAAAAGTTAAGGGACCC
	R_I_EcoRIAflII	tgtacaatacgattactttctgttcgacttaagaattttaCTTGACCGGACCCACAATAC
	F_EB/-_IEcoAfl	cgtattgtgggtccggtcaagtaagAATTCTTAAGTCGAACAGAAAGTAATCGTATTGTA
	R_EB/-_BrbsI	ggtcccttaacttttcttgtcataagtatatctccttcttCTTAGCCAATTCGTGCTGCA
	F_PaY_NdeI	tcttagtatattagttaagtataagaaggagatatacatATGCAACCGCATTATGATCTG
	R_PaY_AvrII	atgctagttattgctcagcggtggcagcagcctaggTTAACGATGAGTCGTCATAATGGC
	Name	Sequence (5'-3')
PaCrtEBIY/- PaCrtEBIY/Blh PaCrt EBI/-	F_PaEBIY_NcoI	ataattttgtttaactttaagaaggagatataccatggATGACGGTCTGCGCAAAAAAAC
	R_PaEBIY_AflII	acaatacgattactttctgttcgacttaagttatTTAACGATGAGTCGTCATAATGGCTTG
	F_PaI_H246H	GCCAGAGTCAGCCAcATGGAAACGACAGGAAAC
	R_PaI_H246H	GTTTCCTGTCGTTTCCATgTGGCTGACTCTGGC
	F_PaEBIY_PaY	TAATAACTTAAGTCGAACAGAAAGTAATCGTATTGTAC
	R_PaEBIY_PaI	agatatacttaagTTATTATATCAGATCCTCCAGCATCAAAC
	F_blh_NdeI	tattagttaagtataagaaggagatatacatATGGAGATGGCAAAGTTAAAGACATTTTC
	R_blh_AvrII	tattgctcagcggtggcagcagcctaggttaCTAACGAAGGTGCGATTTATTCTTGAGAG
	Name	Sequence (5'-3')
ActR	F_ActRB_NdeI	aaataattttgtttaactttaagaaggagatatacatATGTCGAGCACCATCG
	R_ActRB_XhoI	atctcagtggtggtggtggtgctcgagATGGTCGTCTTTCATGCC

[a] Annealing regions are indicated with capital letters

Chapter III – Complex Carotenoids in acI *Actinobacteria*

Publication Statement

Parts of this chapter have been published as: Dwulit-Smith JR, Hamilton JJ, Stevenson DM, He S, Oyserman BO, Moya-Flores F, Garcia SL, Amador-Noguez D, McMahon KD, Forest KT. (2018). acI *Actinobacteria* Assemble a Functional Actinorhodopsin with Natively Synthesized Retinal*. Applied and Environmental Microbiology. 84 (24): e01678-18. doi: 10.1128/AEM.01678-18*

Scientific Contribution Statements

Dwulit-Smith JR performed all bioinformatics and biochemical experiments.

Hamilton JJ analyzed metagenomic data and calculated RPKM.

He S and Oyserman BO collected environmental acI samples.

Moya-Flores F collected environmental acI samples and performed RNA extraction and sequencing.

Garcia SL helped select and obtain genome AAA278-O22 for gene expression.

McMahon KD and Forest KT advised the experimental plans.

Abstract

Experimental evidence suggests that acI *Actinobacteria* produce simple carotenoids like lycopene. Additionally, genomic evidence suggests that the organisms are also capable of complex carotenoid production. Here, I map the genes for complex carotenoid production and show that they are in the same neighborhood as genes for simple carotenoid synthesis (lycopene, γ-carotene, β-carotene). Also, I show how the enzymes may fit together into a logical pathway that uses γ-carotene as a platform for modification and terminates in an oxygenated, glycosylated, and possibly acylated complex carotenoid. The final set of possible complex carotenoids are based on verified examples from organisms that produce similar chromophores. We also verify that the genes for complex carotenoid enzymes genes are found in transcriptomics data from Lake Mendota. Determining if these gene products can experimentally work in concert to generate a complex carotenoid, and if that carotenoid is the same as a natively derived sample are next steps in defining the secondary metabolism of acI *Actinobacteria*.

Introduction

Carotenoids are a group of colorful polymers and some of the most widespread and diverse natural products synthesized by plants, animals, fungi, algae, and bacteria. Carotenoids are long-chain isoprenoids (terpenoids) formed from five-carbon units of 2-methyl-1,3-butadiene, isoprene (1). The main unit must be interconvertible between isopentenyl diphosphate and its double-bond-containing isomer, dimethylallyl diphosphate; both needed for initial synthesis. Afterwards, only isopentenyl diphosphate is required for additional enzymatic polymerization. In general, the chains reach forty carbons long as they are condensed from eight isoprene units. However, shorter (C30) and longer (C50) carotenoids have been isolated and described (2). Specific examples include, *Staphylococcus aureus* 4,4' diapolycopene (3) and *Pseudomonas rhodos* glucosyl esters (C30) (4) and *Corynebacterium glutamicum* flavuxanthin (C45) and decaprenoxanthin (C50) (5, 6). Nevertheless, the initial phases of carotenoid biosynthesis are the same before modifications occur and the process branches. The secondary steps include synthesis of larger compounds (e.g. geranyl diphosphate, farnesyl diphosphate, geranylgeranyl diphosphate). Once these compounds are synthesized, joined, and desaturated, simple carotenoids, like lycopene, γ-carotene, β-carotene, and various saturated intermediates exist (7). Simple carotenoids are classified by having only minimal modifications like cyclizations and never substituent groups like oxygenations. On the other hand, more complex carotenoids, like myxoxanthophyll (8), salinixanthin (9, 10), and echinenone (11) also exist in organisms.

In addition to being highly varied in structure, carotenoids play important biological roles for the organisms that produce them. They function as photoprotecting antioxidants, light-harvesting components of energy transduction systems, and membrane fluidity regulators (12, 13). Carotenoids are ideal for these functions due to their extensive conjugated bond networks. Reactivity and isomerization at unsaturated bonds provide the necessary protection from harmful reactive oxygen species and excess light (14). For energy transduction, as in photosynthetic apparatus, the conjugated network allows electron delocalization and energy transfer to other

chromophores (15, 16). Of course, the wavelengths of energy absorbed depend on the conjugation length with longer systems leading to red shifted absorbances (7). Conjugation lengths allow for the tuning of cognate systems (i.e. xanthophylls and carotenes for transfer to chlorophylls in photosynthesis). Another cognate transduction system that carotenoids appear in are antenna rhodopsins. Here, special ketocarotenoids, salinixanthin and echinenone, boost the activity of rhodopsins by transferring energy into the main retinal chromophore (9–11). While only two known examples of this phenomenon have been identified, more likely exist because of the prevalence of organisms that have rhodopsins with glycine near the top of α5 (17) and complex carotenoids (18). The ability to produce functional, chromophore-loaded rhodopsins with a cognate chromophore may have a competitive advantage in their native ecosystems (19, 20). Lastly, carotenoids impact membrane fluidity simply by disrupting tighter packing of phospholipids (21). The packing effect mimics a higher proportion mono- and poly-unsaturated lipid tails in a membrane (22).

While the structures and functions of carotenoids are extremely diverse, the enzymes that synthesize them fall into relatively few classes. The major groupings are based on the types of reaction that the enzymes catalyze (e.g. isoprenyl diphosphate synthases, carotenoid synthases, desaturases, cyclases). Even with the small number of basic reactions, pathways greatly diverge on accumulation of mutations in the genes encoding these enzymes (23–25). Altered substrate specificity and protein-protein interaction residues can subtly modify the flux of a given carotenoid pathway. A simple example is the ratio of single cyclizations of lycopene to γ-carotene versus full β-carotene synthesis (two cyclizations) in *Myxococcus xanthus* (26). Even though this organism has similar genes to other β-carotene-producing organisms, 100% reaction completion is not observed?. Because carotenoid synthesis is sequential in nature, it is reasonable to think that large substrate channeling complexes form on and in membranes. Yet, details of interactions are not known. Evidence for the existence of multienzyme assembly lines has been presented (7), especially in the fungus *Phycomyces blakesleeanus* (27–29). However, the fact that carotenoid-

related enzymes from different organisms form functional, predictable pathways in heterologous hosts shows that a specific enzyme complex is not a prerequisite.

Results

A pathway for a complex carotenoid contains at least nine genes. I previously noted that genes for production of a predicted carotenoid glycoside ester may be encoded adjacent to the lycopene gene neighborhood (30) (Fig. 3.01). I used bioinformatics, sequence homology, topology predictions, and known carotenoid structures to assign functions to gene products and assemble two plausible pathways (Fig. 3.02 and Fig. 3.03 to 2.06) (8, 31–38). The crucial branch point compound for initiating the pathway is γ-carotene (step 4), which is predicted to be produced in the retinal pathway after synthesis of phytoene and lycopene. Modifications of γ-carotene are expected to include desaturation of a carbon-carbon bond at the noncyclized end by CrtD, introduction of two hydroxyl groups in separate steps by CrtA and CruF, and transfer of a monosaccharide, like fucose or rhamnose, by CruG onto the CrtA-inserted hydroxyl group. The carotenoid resulting from CrtD and CruG would be similar to the known cyanobacterial carotenoid myxol 2’-fucoside, although lacking a ring hydroxyl (Fig. 3.02B) (8). For the sake of discussion, I will refer to product of CrtA, CruF, CrtD, and CruG as complex carotenoid A, with the caveat that the carotenoid could have a different structure than predicted, especially regarding substituent modifications, and it may be a previously characterized molecule.

Genes encoding enzymes for complex carotenoid synthesis (*crtA, cruF, cruG*, and *crtD*) are found upstream of the lycopene neighborhood as a contiguous group, which most often includes a putative acyltransferase (Fig. 3.01). The proposed acyltransferase chemistry is not well defined, and it is not homologous to the acyltransferase, CruD (39), yet the involvement of the gene in carotenoid synthesis is supported by database annotation as a coenzyme A (CoA) methyl esterase. Because many carotenoid modification proteins are related in protein sequence (i.e., lycopene desaturases and carotene ketolases) (40), functional validation will be required for all of

the carotenoid pathway enzymes. Specifically, the *crtD* gene could encode a ketolase, CrtO, which would introduce a carbonyl onto the β-ionone ring rather than desaturating an additional bond, and the *cruG* gene product may instead be CruC, which attaches a glucose onto a CruF-inserted hydroxyl group (Fig. 3.02A) (39). CrtA, CruF, CrtO, CruC, and a proposed acyltransferase combination may produce a rhodopsin antenna, like salinixanthin or echinenone (Fig. 3.02C) (9–11) This carotenoid will be referred to as complex carotenoid B.

Metatranscriptomic analysis of pathway gene transcripts in environmental acI populations. For complex carotenoid synthesis in acI cells, the appropriate genes must be expressed. To measure gene expression in environmental acI *Actinobacteria*, four metatranscriptome samples were collected across multiple time points from the surface of eutrophic Lake Mendota (Dane County, WI, USA), and RNA was isolated and sequenced. The resulting transcripts were mapped to available acI SAGs and MAGs to quantify relative gene expression levels in acI cells. Notably, transcription is observed for genes in the putative complex carotenoid pathway (Table 3.01). This mapping of Lake Mendota metatranscriptome samples provides direct evidence for a complex carotenoid operon (Fig. 3.07). Specifically, reads that mapped to the most populous genome, ME885, overlap the intergenic regions within the operon. Synteny across acI clades similarly supports assignment as an operon. The reading frame overlap in acI-C *Actinobacteria* of *crtA* with the putative acyltransferase supports a lipid attachment step in complex carotenoid B synthesis.

Discussion

Prior to this work, there was no biochemical evidence to support the hypothesis that acI *Actinobacteria* synthesize complex carotenoids in freshwater. I have provided bioinformatic and experimental verification of an environmental adaptation that may allow the acI lineage to be so abundant.

The complex carotenoid pathway is predicted to start from a carotenoid intermediate, γ-carotene, to synthesize a chromophore, like complex carotenoid A or B, with a yet unconfirmed structure and function. The platform carotenoid likely originates from retinal and ActR synthesis that was described in Chapter II. Members of all acI clades contain the operon of other carotenoid biosynthetic machinery (a putative acyltransferase, *crtA*, *cruF*, *cruG*, and *crtD*) (Fig. 3.01 and 2.07). The gene products can be predicted to produce complex carotenoids with glycosyl or glycosyl and acyl modifications, complex carotenoid A or B, respectively (Fig. 3.02A). Although the structure and presence of the carotenoid remain to be experimentally validated, transcriptomics analysis detected all pathway genes from acI (Table 3.01). The transcription levels of these genes are much lower than actino-opsin transcripts shown in Chapter II. The discrepancy is similarly explained as for retinal synthesis pathway gene levels; the enzymes for chomophore production are needed in lower, catalytic amounts rather than structural platform roles like the opsin.

Actino-opsins present in acI genomes contain the proper attributes for actinorhodopsin formation as described in Chapter II. In addition the presence of a glycine void near the top of α5 that is a necessary (although not sufficient) condition for binding of an antenna ketocarotenoid (11, 41). Given that actinorhodopsins and complex carotenoids are produced in acI *Actinobacteria*, the proteins could may also bind cognate complex carotenoids in the native environment. The system would be another example of a special class of rhodopsins that I call antenna rhodopsins (9–11), but others are predicted (42). We are left with the following questions: can acI synthesize a complex carotenoid using the genes found upstream from the lycopene cluster, and does it interact with acI actinorhodopsin as an antenna?

Methods

acI gene identification and pathway assembly. Annotations for multiple acI SAGs (30) and MAGs were analyzed for genes relating to carotenoids using the Joint Genome Institute's

Integrated Microbial Genomes Viewer (43). Translated candidate protein sequences were used to identify homologs in other acI genomes, and those with consistent gene neighborhoods and known carotenoid-related functions were prioritized. After function assignments, two pathways for carotenoid biosynthesis and use in acI *Actinobacteria* were assembled.

acI biomass collection and transcriptomics. Environmental sampling, metatranscriptome sequencing, and gene expression calculations were all performed as previously described (44). Briefly, four samples were collected from within the top 12 m of the Lake Mendota (Dane County, WI) water column and filtered through cheesecloth and onto 0.2µm mixed cellulose filters (Whatman). Filters were immediately frozen in liquid nitrogen and stored at 80°C. Samples were subjected to TRIzol-based RNA extraction, phenol-chloroform separation, and RNA precipitation. RNA was further purified using the RNeasy minikit (Qiagen) with on-column digestion of DNA via the RNase-free DNase set (Qiagen). RNA was then sent to the University of Wisconsin—Madison Biotechnology center for rRNA depletion, cDNA synthesis, and sequencing. rRNA was depleted using the RiboZero rRNA removal kit (bacteria) (Illumina). Samples were then prepared for sequencing using the TruSeq RNA library prep kit v2 (Illumina), pooled in an equimolar ratio, and sequenced on an Illumina HiSeq 2500 platform using 2 100-bp paired-end sequencing. After sequencing, metatranscriptomic reads were trimmed, merged, subjected to in silico rRNA removal, mapped to carotenoid-related genes, and counted. Raw paired-end reads were trimmed using Sickle (45), merged using FLASH (46), and subjected to in silico rRNA removal using SortMeRNA. Sickle was run using default parameters; FLASH was run with a maximum overlap of 100 nucleotides; and SortMeRNA (47) was run using databases for bacterial, archaeal, and eukaryotic rRNA, derived from SILVA v119 and RFAM v12.0 (48). Trimmed and merged reads from all four samples were then pooled and mapped to a single reference FASTA file containing 36 high-quality acI genomes from a larger freshwater genome collection (11). Reads were competitively mapped to genes using BBMap (https://sourceforge .net/projects/bbmap/) with the ambig random and minid 0.95 options. Next, reads mapping to each carotenoid-related gene were

counted using hts-count (49) with the intersection_strict option. Within each clade, reads mapping to each carotenoid-related gene were pooled, and gene expression was computed on a reads per kilobase per million (RPKM) basis (50), while also accounting for different gene lengths and total mapped reads for each acI genome. All scripts are found on GitHub (https://github .com/joshamilton/Hamilton_acI_2017/tree/actR).

Data availability. All genomic and metatranscriptomic sequences are available through the Integrated Microbial Genomes (IMG) and National Center for Biotechnology Information (NCBI) databases, respectively. Genome sequences can be accessed using IMG taxon ID numbers provided in Table 1.01. The raw RNA sequences can be found in the Sequence Read Archive (SRA) of the National Center for Biotechnology Information under BioProject accession no. PRJNA362825.

References

1. Ruzicka L. 1953. The isoprene rule and the biogenesis of terpenic compounds. Experientia 9:357–367.

2. Armstrong GA, Hearst JE. 1996. Carotenoids 2: Genetics and molecular biology of carotenoid pigment biosynthesis. FASEB J 10:228–237.

3. Umeno D, Tobias AV, Arnold FH. 2002. Evolution of the C30 Carotenoid Synthase CrtM for Function in a C40 Pathway. J Bacteriol 184:6690–6699.

4. Kleinig H, Schmitt R, Meister W, Englert G, Thommen H. 1979. New C30-Carotenoic Acid Glucosyl Esters from Pseudomonas rhodos. Z Für Naturforschung C 34:181–185.

5. Krubasik P, Takaichi S, Maoka T, Kobayashi M, Masamoto K, Sandmann G. 2001. Detailed biosynthetic pathway to decaprenoxanthin diglucoside in Corynebacterium glutamicum and identification of novel intermediates. Arch Microbiol 176:217–223.

6. Krubasik P, Kobayashi M, Sandmann G. 2001. Expression and functional analysis of a gene cluster involved in the synthesis of decaprenoxanthin reveals the mechanisms for C50 carotenoid formation: Decaprenoxanthin formation. Eur J Biochem 268:3702–3708.

7. Britton G, Liaaen-Jensen S, Pfander H (ed). 2004. Carotenoids. Springer Basel AG, Basel, Switzerland.

8. Graham JE, Bryant DA. 2009. The Biosynthetic Pathway for Myxol-2’ Fucoside (Myxoxanthophyll) in the Cyanobacterium Synechococcus sp. Strain PCC 7002. J Bacteriol 191:3292–3300.

9. Luecke H, Schobert B, Stagno J, Imasheva ES, Wang JM, Balashov SP, Lanyi JK. 2008. Crystallographic structure of xanthorhodopsin, the light-driven proton pump with a dual chromophore. Proc Natl Acad Sci 105:16561–16565.

10. Balashov SP. 2005. Xanthorhodopsin: A Proton Pump with a Light-Harvesting Carotenoid Antenna. Science 309:2061–2064.

11. Balashov SP, Imasheva ES, Choi AR, Jung K-H, Liaaen-Jensen S, Lanyi JK. 2010. Reconstitution of Gloeobacter Rhodopsin with Echinenone: Role of the 4-Keto Group. Biochemistry 49:9792–9799.

12. Olson JA, Krinsky NI. 1995. Introduction: the colorful, fascinating world of the carotenoids: important physiologic modulators. FASEB J 9:1547–1550.

13. Gruszecki WI, Strzałka K. 2005. Carotenoids as modulators of lipid membrane physical properties. Biochim Biophys Acta BBA - Mol Basis Dis 1740:108–115.

14. Tian B, Hua Y. 2010. Carotenoid biosynthesis in extremophilic Deinococcus–Thermus bacteria. Trends Microbiol 18:512–520.

15. Grotjohann I, Fromme P. 2005. Structure of cyanobacterial Photosystem I. Photosynth Res 85:51–72.

16. Zakar T, Laczko-Dobos H, Toth TN, Gombos Z. 2016. Carotenoids Assist in Cyanobacterial Photosystem II Assembly and Function. Front Plant Sci 7:295.

17. Finkel OM, Béjà O, Belkin S. 2013. Global abundance of microbial rhodopsins. ISME J 7:448–451.

18. Kirti K, Amita S, Priti S, Mukesh Kumar A, Jyoti S. 2014. Colorful World of Microbes: Carotenoids and Their Applications. Adv Biol 2014:1–13.

19. Steindler L, Schwalbach MS, Smith DP, Chan F, Giovannoni SJ. 2011. Energy Starved Candidatus Pelagibacter Ubique Substitutes Light-Mediated ATP Production for Endogenous Carbon Respiration. PLoS ONE 6:e19725.

20. Casey JR, Ferrón S, Karl DM. 2017. Light-Enhanced Microbial Organic Carbon Yield. Front Microbiol 8:2157.

21. Domonkos I, Kis M, Gombos Z, Ughy B. 2013. Carotenoids, versatile components of oxygenic photosynthesis. Prog Lipid Res 52:539–561.

22. Widomska J, Kostecka-Gugała A, Latowski D, Gruszecki WI, Strzałka K. 2009. Calorimetric studies of the effect of cis-carotenoids on the thermotropic phase behavior of phosphatidylcholine bilayers. Biophys Chem 140:108–114.

23. Raisig A, Sandmann G. 2001. Functional properties of diapophytoene and related desaturases of C30 and C40 carotenoid biosynthetic pathways. Biochim Biophys Acta BBA - Mol Cell Biol Lipids 1533:164–170.

24. Krubasik P, Sandmann G. 2000. Molecular evolution of lycopene cyclases involved in the formation of carotenoids with ionone end groups. Biochem Soc Trans 28:806-10.

25. Sandmann G. 2002. Molecular evolution of carotenoid biosynthesis from bacteria to plants. Physiol Plant 116:431–440.

26. Iniesta AA, Cervantes M, Murillo FJ. 2008. Conversion of the lycopene monocyclase of *Myxococcus xanthus* into a bicyclase. Appl Microbiol Biotechnol 79:793–802.

27. Aragon CMG, Murillo FJ, Guardia MD, Cerdae-Olmedo E. 1976. An Enzyme Complex for the Dehydrogenation of Phytoene in Phycomyces. Eur J Biochem 63:71–75.

28. Candau R, Bejarano ER, Cerda-Olmedo E. *In vivo* channeling of substrates in an enzyme aggregate for β-carotene biosynthesis. Proc Natl Acad Sci USA. 88: 4936-40.

29. Murillo FJ, Torres-Martinez S, Aragon CMG, Cerda-Olmedo E. 1981. Substrate Transfer in Carotene Biosynthesis in phycomyces. Eur J Biochem 119:511–516.

30. Ghylin TW, Garcia SL, Moya F, Oyserman BO, Schwientek P, Forest KT, Mutschler J, Dwulit-Smith J, Chan L-K, Martinez-Garcia M, others. 2014. Comparative single-cell genomics reveals potential ecological niches for the freshwater acI *Actinobacteria* lineage. ISME J 8:2503-16.

31. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403–410.

32. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genome (ed. F. Cohen). J Mol Biol 305:567–580.

33. Lee PC, Holtzapple E, Schmidt-Dannert C. 2010. Novel Activity of Rhodobacter sphaeroides Spheroidene Monooxygenase CrtA Expressed in Escherichia coli. Appl Environ Microbiol 76:7328–7331.

34. Rählert N, Fraser PD, Sandmann G. 2009. A *crtA* -related gene from *Flavobacterium* P99-3 encodes a novel carotenoid 2-hydroxylase involved in myxol biosynthesis. FEBS Lett 583:1605–1610.

35. Sun Z, Shen S, Wang C, Wang H, Hu Y, Jiao J, Ma T, Tian B, Hua Y. 2009. A novel carotenoid 1,2-hydratase (CruF) from two species of the non-photosynthetic bacterium Deinococcus. Microbiology 155:2775–2783.

36. Takaichi S, Maoka T, Takasaki K, Hanada S. 2010. Carotenoids of *Gemmatimonas aurantiaca* (Gemmatimonadetes): identification of a novel carotenoid, deoxyoscillol 2-rhamnoside, and proposed biosynthetic pathway of oscillol 2,2'-dirhamnoside. Microbiology 156:757–763.

37. Ahn J-W, Kim K-J. 2015. Crystal structure of 1′-OH-carotenoid 3,4-desaturase from *Nonlabens dokdonensis* DSW-6. Enzyme Microb Technol 77:29–37.

38. Teramoto M, Rählert N, Misawa N, Sandmann G. 2004. 1-Hydroxy monocyclic carotenoid 3,4-dehydrogenase from a marine bacterium that produces myxol. FEBS Lett 570:184–188.

39. Maresca JA, Bryant DA. 2006. Two Genes Encoding New Carotenoid-Modifying Enzymes in the Green Sulfur Bacterium *Chlorobium tepidum*. J Bacteriol 188:6217–6223.

40. Klassen JL. 2010. Phylogenetic and Evolutionary Patterns in Microbial Carotenoid Biosynthesis Are Revealed by Comparative Genomics. PLoS ONE 5:e11257.

41. Bertsova YV, Arutyunyan AM, Bogachev AV. 2016. Na+-translocating rhodopsin from *Dokdonia* sp. PRO95 does not contain carotenoid antenna. Biochem Mosc 81:414–419.

42. Imasheva ES, Balashov SP, Choi AR, Jung K-H, Lanyi JK. 2009. Reconstitution of *Gloeobacter violaceus* rhodopsin with a light-harvesting carotenoid antenna. Biochemistry 48:10948–10955.

43. Arrach N, Fernández-Martín R, Cerdá-Olmedo E, Avalos J. 2001. A single gene for lycopene cyclase, phytoene synthase, and regulation of carotene biosynthesis in Phycomyces. Proc Natl Acad Sci 98:1687–1692.

44. Hamilton JJ, Garcia SL, Brown BS, Oyserman BO, Moya-Flores F, Bertilsson S, Malmstrom RR, Forest KT, McMahon KD. 2017. Metabolic Network Analysis and Metatranscriptomics Reveal Auxotrophies and Nutrient Sources of the Cosmopolitan Freshwater Microbial Lineage acI. MSystems 2:e00091–17.

45. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files (Version 1.33).

46. Magoc T, Salzberg SL. 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27:2957–2963.

47. Kopylova E, Noé L, Touzet H. 2012. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics 28:3211–3217.

48. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2012. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res 41:D590–D596.

49. Anders S, Pyl PT, Huber W. 2015. HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics 31:166–169.

50. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5:621–628.

Figures and Tables

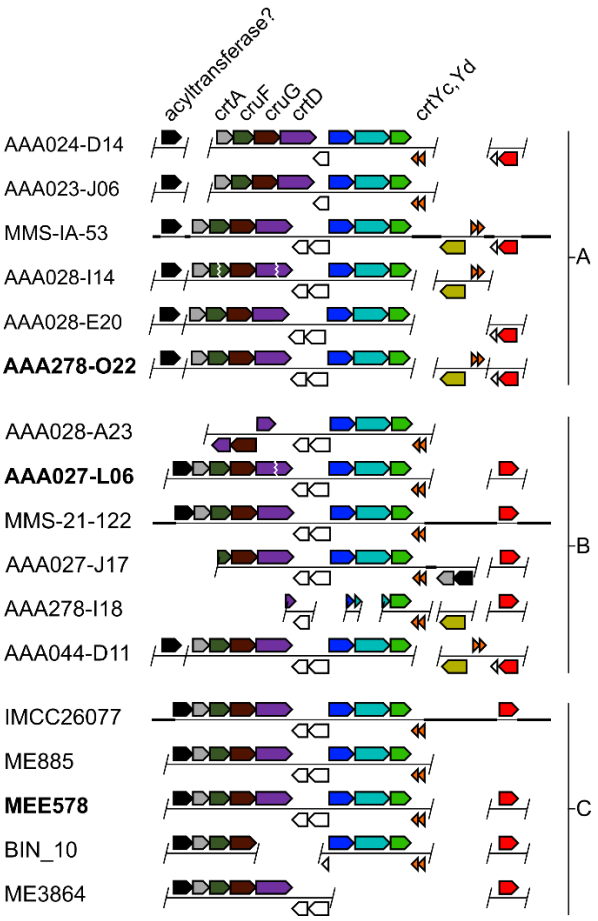


FIG 3.01 acI complex carotenoid-related genes in the genomic context. acI genomic contigs or genomes from clades A, B, and C are primarily labeled by shorthand notation (i.e., designation after "*Actinobacterium* SCGC" or "int.metabat."). Genes are arrows pointing in the direction of transcription. A slash indicates a contig boundary, which may or may not end immediately after the pictured gene, a vertical zig-zag indicates contig ends that have been manually paired, and a thick horizontal bar indicates a longer region of DNA not represented. Uncolored genes are neighboring genes that may or may not be functionally associated with the carotenoid-related genes. Boldface labels indicate a gene source for this study. Relevant Integrated Microbial Genomes (IMG) database locus tags for this study from AAA278-O22 are A278O22DRAFT_00003540 (proposed acyltransferase) and A278O22DRAFT_00010590-10560 (crtA to crtD).

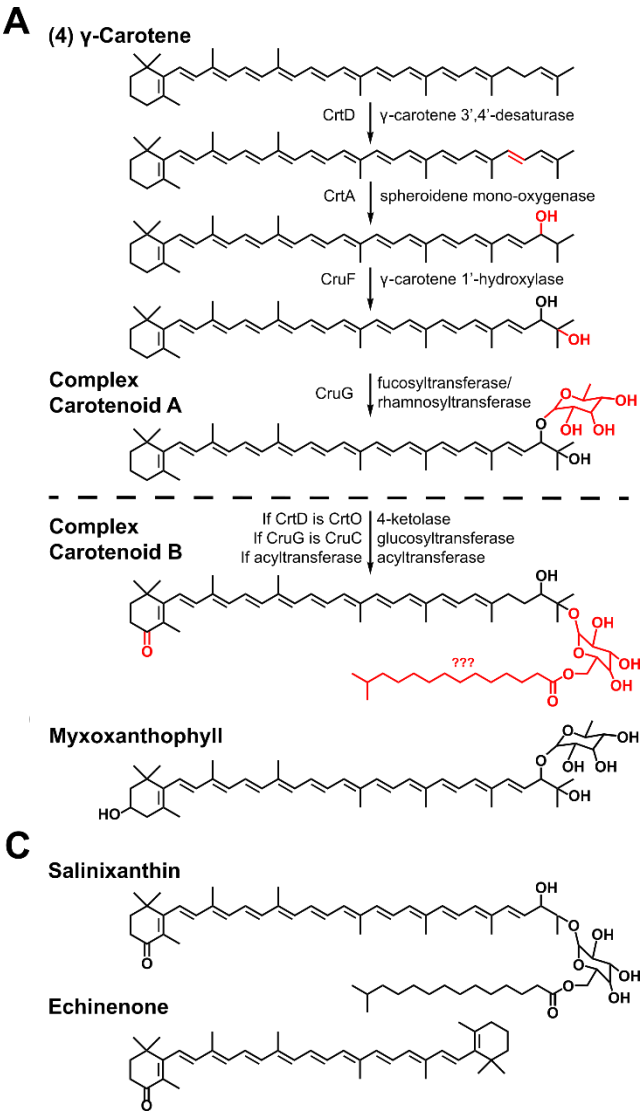


FIG 3.02 Predicted complex carotenoid-related pathways in acI. (A) A complex carotenoid synthesis pathway is predicted to require γ-carotene produced in the actinorhodopsin pathway, with further desaturation, hydroxylation, and glycosylation for a myxoxanthophyll-like carotenoid. A carotenoid with different structure can result if enzyme identities are different and a proposed acyltransferase is involved in the pathway. Question marks indicate uncertainty in naming and/or chemical structure. (B) A complex carotenoid used by many cyanobacteria for photosynthesis. (C) Complex carotenoids from other organisms that function as carotenoid antennae on proton-pumping rhodopsins.

```
Rs              --MQTVTLSIFRFNEFEKRLWVLGQMTANKLGMHYLPKAKFWKMFGSGTGQGFTP-KPNW   57
MbP99-3         MSKQITTLTFFKYPKLKDKIWAFWMMQFAHSALRRQSGLQSYKLLGSGKEG-FSP-WPDW   58
AAA278-O22      --MQLTVVYLFSVE-RKSIPFALISMAIDRLRTRMFTGISFSKLLGTGTGRTFTPSDADL   57
                   * ..: :*     :.  :.:  *   :   :     .  *::*:*.   *:*   :
Rs              HVWSILAVWPDEETARREVAESPIYQRWTKMADESYTVLLQPTSAWGKWDGKEPFEPVKP  117
MbP99-3         SVYGLLQVWDSHKEAQEFFDSSSLYKKYLNHSEQQLTFYMKNIKAYGQWSKKNPFEQHSD  118
AAA278-O22      TRWGMVVVID--KDRLDAFDQSSIVTNWRKRSTSEFRALLSPLSSHGLWAKKNPFDFVAP  115
                  :.:: *    :     . .* :  .: : : ..    :.  .: * *  *:**:
Rs              ASDVRP-IAALTRATVKFWKAERFWAREPAISHMIGRNKDVVFKIGVGEVPFVQQVTFSI  176
MbP99-3         MDTNNPYLTVITRATIKPHMLKKFWDYVPISQKGLKENPSLLFTAGVGERPFTHMATFSL  178
AAA278-O22      FSNSEAQIAAITRARIKWNKNFIFWKSVPPVVLDLHSNPGLIAAIGIGEAPIGLQGTFSL  175
                      .  .  ::.:*** :*      **   *:  * .::   *:** *:    ***:
Rs              WPDASSMEEFARGAGGPHGEAIKAVRAENWFKEELYARFQILGTIGKWEGKDPVGEALTA  236
MbP99-3         WDDARALKKFAYR-GNNHKQAIQQTQALQWYKEEMFSRFQPYLITGSWQGFTIPELL-SF  236
AAA278-O22      WQSASALRDFAYK-SKAHQVAIAQTESIGWYSEELFARFEVLELRGEITAHAGK------  228
                * .* ::..**   .  *  **  ..:  *:.**:::**:     *.  .
Rs              RPSEAPKPAPAPAAAQPAPAVEAPKPAPAPVAEKPALAVEMPKSAEPPKPVVEAPKPASA  296
MbP99-3         KET-----------------------------------------------------------  239
AAA278-O22      -----------------------------------------------------------  228

Rs              PVASKPMPQGGKPNFKGKPGKGGRKENA                                  324
MbP99-3         ----------------------------                                  239
AAA278-O22      ----------------------------                                  228
```

FIG 3.03 Alignment for predicted spheroidene mono-oxygenase (acI-CrtA). The protein from acI genome AAA278-O22 is aligned to functionally characterized CrtA proteins from *Rhodobacter sphaeroides* (Rs) and *Marine bacterium* P99-3 (MbP99-3).

```
AAA278-O22      ---------------------------------------------------MSIRHYN    7
Dr              MISPPLLRWGLAAAGLGLAFLGALLALAEQSAGWALIALGLPLSGVFALAGDALGSAFGA   60
Dg              -MSPALLRLTLSVAALGLAFLGAVLVRRGLGPGWLLIALGLPLTAVLALAGDALGPALRG   59
                                                                        :.
AAA278-O22      PRRRRRGGISPKARNFLFLLVGVTIALQISYPLITGETLRI--VTIATVYVGAMSMVIHG   65
Dr              TLRRRWQRLLAETPPWL-GWLALSVALKIPVPLWPQGFALLGLLSTGALFVAGLSYAA--  117
Dg              TLRRRTGLLVRQMRPWL-WLTGLCAALKIPVPLWPEGFPLLALLSTGALGLAALAYLE--  116
                  ***   :  :   :*    .:  **:*  **       :  :: .:: :..::
AAA278-O22      HLSYGAKYSTRYLPITALFGLGIEVLGVHTGWPFGIYEYDASLGAQLFGVPIVVPFAWVM  125
Dr              -QRAGWGQAAQLAALACLAGLGVELLGSRTGVPFGQYSYATAPAPTVLTVPLIVPLGWFA  176
Dg              -ERVGAWRALGLAALGFGVGLGVELLGSQTGWPFGVYSYATTPAPALLGVPLIVPLGWFA  175
                    *   :     :    ***:*:** :** *** *.* :: .  :: **::**:.*.
AAA278-O22      MVHPALIAARRIAGHWVFLYGGLLLAAWDLFLDPQMVAAGRWKWEVPGAHVPFTPDIPLS  185
Dr              FTLAGLQ----LAGGRP-WLAGLLITCWDVGLEPLMTAQNYWRWSDPH---PLWAGAPVQ  228
Dg              LTLCAAL----LAGGRA-WLAGLLLVAWDVGLEPLMTAAGYWHWTDPR---PLWAGAPLQ  227
                :.  .      :**      .***:..**: *:* *.* . *:*  *    *:  . *:.
AAA278-O22      NALGWLLAGIVIIGALNKILPRERRKEAASLAAVDALLLWTLFAGFIGNLFFFDRPGLAF  245
Dr              NFLGWWAVGTAIAWGMKRLAPGLFDAKEKQKAEG--LS-PSFSLAYFTEAF-FLPGGLVL  284
Dg              NFVGWWAVGAGLAWAFVRLAPGLVGPRSAR------PR-LTFAVAYLVETF-FLPGGLVL  279
                * :**  .*  : .: :: *     .             ::  .:: : * *   **.:
AAA278-O22      FGTFIMGALLTPY----FFSSWLGNKD-                                  268
Dr              VGRLAEAGVTLAVMGLGALLARGVRRAS                                  312
Dg              VGRVREAAVTLLVMLGALALAWALRGDR                                  307
                .* .  ..:           :   .
```

FIG 3.04 Alignment for predicted g-carotene 1'-hydroxylase (acI-CruF). The protein from acI genome AAA278-O22 is aligned to functionally characterized CruF proteins from *Deinococcus radiodurans* R1 (Dr) and *Deinococcus geothermalis* DSM11300 (Ds).

```
AAA278-O22      ----------------MAQSLLTLSILIFIISLINFVTIRIPK----RDEEIKKSVTVL   39
Ga              MIDAVPWSSLGVASF-LVGQAL---CLMVLVSRLAPGRSRRPPVSPRLAPRDD-TTVSVI   55
SP              -MDFSPVWIGGLCLFGLLIQGS---GALVVLSRLMKGAVRRSPLTPQASNSDNLAAVTVV   56
                                : *.      ::.:  *      * *        :   :*:*:
AAA278-O22      APMRNEAENVPEFISALSSQMGVKNLNFVIINDGSTDKTAELLTSVIDGDPRFSFIDSPI   99
Ga              VATLNEAHRIGPCLDGLLQQ-PAPLLEVLVVDSRSRDGTPELVQTYADRDPRIRLITDDP  114
SP              VPSLNEVERIQPCLDGLSQQ-SYEVREILVVDSNSTDGTREKVLAKAATDPRFRLLTDDP  115
                .   **...:   :..* .*      :.:::::. * * * * : :    ***: :: .
AAA278-O22      QRDGWLGKVSALQSGYES--ARSEFIITLDADVRLQPNAIMRAISQLERLKLDFVSPYPR  157
Ga              LPPGWVGKVWALQTGLQA--ARGAWVLGIDADTVPAAGLVGGVIDAAERDRFDVVSFSPR  172
SP              LPQGWVGRPWALNWGFEHSSPDSEWILGVDADTQPQRGMIASVLQAAEEEGFDLVSLSPQ  175
                   **:*:  **: * :     . ::: :***.    . :  .:.  *.  :*.**  *:
AAA278-O22      QIAQTFAEKLIQPLLHWSWMSTVILRLAEKFPRRSTAVANGQFFVARKNALDAINGFESV  217
Ga              FAGQTAAERLVQPAMLVTLVYRTG-AAGAEQ-QPDRVLANGQCFLARRAVLEQHGGYAVA  230
SP              FILRSPGEWWLQPALLMTLLFRFD-VAGIRQPDQDSVMANGQCFLCRRKVLENVGGYRSA  234
                   :: .*  :** :  : :       . .    . .:**** *:.*: .*:  .*:  .
AAA278-O22      STQILDDIELARSLISAGYRGVVTEGSGIASTRMYSSFDEIRQGYGKSLWKAFGGSIGT-  276
Ga              RGSFSDDVTLARHLAMHGARVGFLDGSRIIEVRAYATLREMWREWGRSFDLKDSASRVQG  290
SP              AGSFCDDVTLARNIAKAGYRVGFLDGAKIIKVRMYEGMRETWDEWGRSLDLKDATSRTEL  294
                  .: **: *** :   * *  . :*: * ..* *  : *    :*:*:    . *
AAA278-O22      VIAIAFLFATGILPVLMILN----------------GYLIGWLIYLYIV-FSR--EISA  316
Ga              WLDVVLVWLVQALPLPVLLGGVLLWFGAPAALSPYPQWLLVALIATNSFAVLLRILMLWV  350
SP              WGQLWLLTMVQGLPIPLTM---LLFGGIEEGLS---NPFLSGWLDLNVFLLLVRFGMLLA  348
                   : ::  .  **: : :                        * . :    :  : *   : .
AAA278-O22      IRS--------RSNPLFAFLHPLSSALLIYLIIYSWRNRGTIQWKGRTV-------      357
Ga              LRTSYHE---RGVTYWLSWLSD---AAAAWRLTMS-MARTPRSWRGRRYASERAPV      399
SP              ISGSYDRQFSVGKSAWLFWLSPFADPFAVARIFLS-ARQKPKAWRGRVYS------      397
                :            .  : :*           :  *   :    *:**
```

FIG 3.05 Alignment for glycosyltransferase (acI-CruG). The protein from acI genome AAA278-O22 is aligned to functionally characterized CruG proteins from *Gemmatimonas aurantiaca* (Ga) and *Synechococcus* sp. PCC7002 (SP).

```
AAA278-O22      MTGQVGKVKNPEKIAVIGAGIGGLCTAARLAKAGHSVTIFEASSRTGGKCRTEWIGRYAF   60
MbP99-3         ----------MKKAIIIGSGIAGLAAALRLKKKGYQVSVFEKNDYAGGKLHAIELGGYRF   50
Nd              ----------MKNAIVIGAGIGGLAAALRLRHQGYSVTIFEKNDYAGGKLHAIEKDGYRF   50
                          ::  :**:**.**.:* ** : *:.*::** .. :*** ::   . * *
AAA278-O22      DTGPSLLTLPAVYRDFFQRTGDVMGRVLELEEVNPSFDYRFHDGTSVKFANLSRKQTLAS  120
MbP99-3         DLGPSLFTLPHLIDELHQLFPDVEID-FNYIKKKTACHYFWEDGTSFEAPADLENFAVKA  109
Nd              DLGPSLFTLPHLVENLFALFPEEIID-FGYKSKAISFHYFWDDGTLFKASTDSSQFIEDA  109
                * ****:*** :  ::.    :     : .   : .* :.*** .:      :    :
AAA278-O22      IEQTLGAAAASEWNRVMLRAEAMWDVSREPFIESELKSPLSLLKRPRLIRDLKIISPWK-  179
MbP99-3         --AEIFDEKQNTLSKYLQNSKMKYESTKSLFLEKSLHKSNTYFSKQTLKAI--LKIPFLG  165
Nd              --SKVFKEEKSTIKKYLAKSKSKYELTKSLFLEKSLHKATTYFSLDTVKAI--VHAPFLG  165
                     :     .  .: : .::  :: ::.*:*..*:.  : :.   :     :  *:
AAA278-O22      -----SLREIGITNPYLAKIMDRYATYSGSDPRYAPAVLSTIAFVEEAFGAWHIKGGIGT  234
MbP99-3         INNTLDQENKKFSNPKLNQLFNRYATYNGSSPYLTPGIMSMIPYLELGLGTYYPQGGMHR  225
Nd              LNNTLNDENSKFKNPKLTQLFNRYATYNGSSPYQTPGIMTMIPHLELGLGTYYPDGGMHR  225
                     . .:  :.** * ::::*****.**.*  :*.::: * .:* .:*::: .**:
AAA278-O22      LAEKITERCEKLGVDIRLNSRVNEIVLNKSRVTGVVVNDATLTFARVVANADAQFVYEKL  294
MbP99-3         ISQSLFELAQKVGVEFRFRKNVKKINHSNNKVTGVTTEKGTHDADIVLCNMDVFPTYRQL  285
Nd              ISQSLFELAQKVGVKFRFRESVTNITTSKNKVTGVETKNGSYLSDLVVSNMDIVPTYRNL  285
                :::.: * .:*:**.:*:.. *.:*  .:.:**** .:..:     *:.* *   .*.:*
AAA278-O22      LAPTKKVVNIRKKLAKQEPSLAGFSLLLGLKPSDSQPLEHHTILFPENYDLEFESIFTKR  354
MbP99-3         LQDIK----APEKTLKQERSSSALIFYWGIKKS-FPQLDLHNILFSENYKAEFEAIFNNK  340
Nd              MKDVP----APEKTLSQERSSSALIFYWGIDRE-FPELDLHNILFSEDYKTEFEHIFEHK  340
                :          :*  .** * :.: :  *:. .    *: *.*** *:*. *** ** ::
AAA278-O22      TPVEKPTIYICAPRD-PLMVKDKDHESWFVLVNAPRHSTSGNGFDWSNQDFNRQYANSII  413
MbP99-3         SLYEDPTVYINITSKQSPQDAPKGCENWFVMINTPGD----YGQNWENLVIKA--KKNIL  394
Nd              TLAQDPTVYINITSKESSNDAPAGHENWFVMINAPGD----YGQDWEQLVEES--KKQII  394
                :  :.**:**   .        . *.***::*:* .     * :*.:   :    :.*:
AAA278-O22      NQI-ETAGISIRERLEVLEIRTPLDLQESVNAPGGSIYGTSSNGARSAFARAKN-RSPIK  471
MbP99-3         SKIKRCLNIDVEELIDVEYVLTPQGIEKNTSSYRGALYGAASNNKFAAFLRHPNFNKTIG  454
Nd              AKIKKCLHVDISKHITTEYILTPQGIEKNTSSYRGALYGAASNNKFAAFLRHPNFNGKIK  454
                 :* .   :.: : : .  : ** .:::...:  *::**::**.  :** *  * .  *
AAA278-O22      GLYLVGGSAHPGGGLPLVGLSAEMVAKAILE---                            502
MbP99-3         NLYHVGGSVHPGGGIPLCLLSAKITADLIPNTNA                            488
Nd              NLYHVGGSVHPGGGIPLCLLSAQITADLIQKEQ-                            487
                .** ****.*****:**  ***::.*. * :
```

FIG 3.06 Alignment for predicted g carotene 3',4' desaturase (acl-CrtD). The protein from acl genome AAA278-O22 is aligned to functionally characterized CrtD proteins from Marine bacterium P99-3 (MbP99- 3) and *Nonlabens dokdonensis* DSW-6 (Nd).

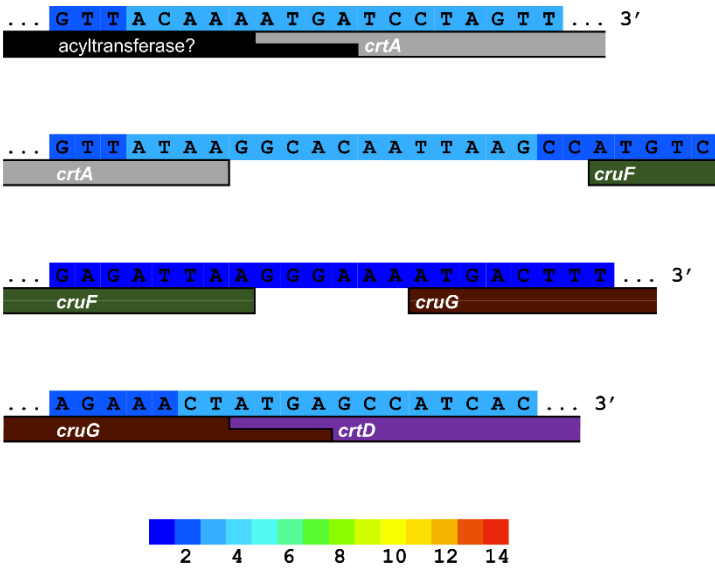


FIG 3.07 Intergenic transcripts that map to carotenoid-related genes in genome ME885. An operon is evident for complex carotenoid (putative acyltransferase through *crtD*) biosynthesis in an acI-C member. The color key in the center indicates of the number of times a base was covered. Transcripts and genes may continue beyond the edges of the DNA window.

TABLE 3.01 Log2 RPKM values for pooled carotenoid-related gene transcripts within each clade.

	Log2 RPKM Values		
Gene Name	**acI-A**	**acI-B**	**acI-C**
crtA	2.11	3.39	5.24
cruF	2.63	3.85	5.59
cruG	2.90	3.62	4.51
Putative acyltransferase	2.64	4.04	5.37
recA	7.85	9.16	10.09

Chapter IV – The Structure and Function of a Peptidase from an acI-B1 *Actinobacterium*

Publication Statement

To be published as Dwulit-Smith JR, Satyshur K, McMahon KD, Forest KT. The Structure and Function of an acI *Actinobacterial* Dipeptidase E, PepE.

Scientific Contribution Statements

Dwulit-Smith JR performed all bioinformatics and biochemical experiments and contributed to crystallographic structure determination and analysis.

Satyshur K contributed to crystallographic data collection, phasing, and refinement.

McMahon KD and Forest KT advised the experimental plans.

Steinüchel A and Weifel L provided the generously provided the plasmids for CphA and CphE overexpression.

Abstract

acI *Actinobacteria* are abundant in lakes likely because the organisms use a range of dissolved molecules from the freshwater in which they live. On top of this, genomes from acI *Actinobacteria* contain a gene that has been annotated as a cyanophycinase (CphB) or as a peptidase E (PepE) by varying databases. If a cyanophycinase, the enzyme would be able to degrade cyanophycin, an Asp-Arg storage polymer made by cyanobacteria in the lake. This would have major implications for the nutrient flow between cyanobacteria and acI *Actinobacteria* in the lake environment. If the enzyme is a peptidase E, it would allow acI to degrade Asp-X dipeptides. Regardless of function, determining the role of this protein would enable a better model of intracellular nitrogen metabolism in acI *Actinobacteria*. Here, I map the doubly annotated gene in the acI genomic context and go on to determine its function and structure. To serve as a trial substrate, I heterologously produce cyanophycin in *Escherichia coli* using a cyanobacterial cyanophycin synthetase and show that the acI gene product is unable to break down the cyanophycin as a cyanophycinase would. However, the purified acI enzyme can degrade an aspartate-leading dipeptide substrate, aspartate-p-nitroaniline. Having determined the function to be PepE-like, I crystallize this representative PepE and refine the X-ray crystal structure to 1.93 Å. This near atomic resolution model reveals the standard S-H-E catalytic triad and substrate binding pocket of other peptidase E enzymes. Determining the range of peptide substrates is a next step in defining how acI *Actinobacteria* additionally adapt to their ecosystem.

Introduction

Serine peptidases are prevalent and found in all kingdoms of cellular life as well as many viral genomes. In fact, over one third of all known proteolytic enzymes are serine peptidases. The name is apt because the active site contains a nucleophilic serine that attacks the peptide bond in the substrate at the carbonyl carbon (1). Nucleophilicity in many serine enzymes depends on a triad charge relay system: the main serine, a general base that abstracts the serine sidechain proton, and a third residue that stabilizes the activating sidechain (2). Serine triad enzymes can be classified into five groups: the trypsin-like proteases (3), the subtilisin-like proteases (4), the serine carboxypeptidases (5), ClpP protease (6), and cytomegalovirus protease (7). The triad consists of S-H-D, except for cytomegalovirus protease, where it is S-H-H. However, the three-dimensional fold and the order that the catalytic triad residues appear in the primary sequence are characteristic for each group.

Serine peptidases have a generalized reaction pathway involving two tetrahedral intermediates (1). As stated above, the main base-activated serine sidechain attacks the carbonyl carbon of the peptide substrate. The tetrahedral oxyanion is stabilized by interactions with the protein backbone because of a positively charged pocket known as the oxyanion hole. Collapse of the tetrahedral intermediate then generates a stable acyl-enzyme complex. Finally, a water displaces the freed substrate fragment and generates a new tetrahedral intermediate with the enzyme-linked substrate. Collapse of this intermediate leads to a regenerated enzyme. There are two key examples of this reactivity: the cyanophycinases, CphB and CphE (8), and the dipeptidase, PepE (9).

Most cyanobacteria intracellularly deposit granules of the cyanophycin polymer under excess nitrogen conditions (10). Structurally, the cyanophycin unit is L-aspartic acid linked to L-arginine, or sometimes L-lysine. The aspartate sidechain carboxylate is joined to the amino terminus of the arginine, however the main polymer backbone contains normal peptide linkages between the aspartates (11). Cyanophycin is non-ribosomally synthesized by cyanophycin

synthetase, CphA, and generally falls between 25 and 100 kDa in length (12). When nitrogen is limiting, the polymer is broken down to the constituent Asp-Arg unit in a carboxy-terminal-specific fashion and then further processed back to free amino acids (13). The Asp-Arg monomer generating enzyme, cyanophycinase (CphB), is seen in a number of eubacterial strains with a secreted version, CphE, found in *Pseudomonas anguilliseptica* strain BI, for example (14). Cyanophycin is highly resistant to degradation by all proteases, and the only enzyme known to be capable of hydrolyzing it are these specialized cyanophycinases. Additionally, the enzymes show noticeable sequence similarity to aspartyl dipeptidase, PepE, from *Salmonella enterica* serovar Typhimurium (15).

PepE is an enzyme that differs in specificity from all other known peptidases in that it hydrolyzes Asp-X dipeptides and only one aspartate-leading tripeptide, Asp-Gly-Gly (15). This is compared to the many enzymes capable of hydrolyzing other small peptides (16, 17). It does not hydrolyze Glu-X, Asn-X, Gly-X, or any other nonaspartyl peptide (18). The only other Asp-X-specific peptidases are caspases, which are endoproteases that hydrolyze an internal Asp-X peptide bond (18). The PepE enzyme from *S. enterica* serovar Typhimurium has been heavily studied. It is hoped to be a target for decreasing organism virulence as it is believed to play a role in the salvage of nutrients from intracellular proteins and imported peptides, even though there are redundant aspartate-peptide-metabolizing systems (19). Additionally, a tested homolog is encoded in the genome of the eukaryote *Xenopus laevis* (15). In *X. laevis*, the enzyme is tied to development of the embryo in response to thyroid hormone, and it is thought to aid in apoptosis during tail resorption rather than simple dipeptide salvage (20).

Serine proteases are among the best-studied enzyme classes but determining their substrates can be difficult because of sequence divergence even though certain enzyme features are nearly universal. Protein similarity is likely the reason that databases annotate an acI CphB-homolog as both cyanophycinase and dipeptidase. Both substrate and structure experiments can be helpful in determining the role of CphB-like/PepE-like serine proteins. This chapter does this

by performing functional assays of an acI protein with cyanophycin and an aspartate dipeptide. Additionally, the structure of the protein is determined to confirm a serine-based active site and visualize the overall protein topology.

Results

All acI CphB/PepE contain key amino acids for function. I first determined whether acI protein sequences were consistent with serine peptidase activity either as a cyanophycinase (Fig. 4.01) or a peptidase E (Fig. 4.02). Indeed, all identified sequences contain the features for proper catalytic function. The sequence from *Actinobacterial* L06 was compared to a CphB from *Synechocystis* 6803 (21) and PepE from *S. enterica* serovar Typhimurium (9) and *Listeria monocytogenes* (PDB 3L4E unpublished). The catalytic serine in the S-X-G-X motif is present. Additionally, the general base histidine (H-F) and base stabilizing acidic residue (G-E) are found in the sequence in sequential order. The overall alignment statistics are: 67% coverage with and E-value of 8E-13 and 31% identity for CphB. For PepE sequences, statistics are: 34% coverage with E-value of 0.29 and 22% identity for *S. enterica*, and 76% coverage with E-value of 5E-08 and 27% identity for *L. monocytogenes*.

Overall, the acI sequence comparison is inconclusive in terms of functional identification, which fits previously reported similarities of cyanophycinases and peptidases. As a result, I mapped out the possible chemical breakdown pathways for the protein as either a peptidase E or a cyanophycinase (Fig. 4.03). For activity as a peptidase E, aspartate-leading dipeptides (e.g. Asp-Asp, Asp-Phe) are metabolized to two free amino acids (Fig. 4.03A). For activity as a cyanophycinase, the large cyanophycin polymer must be cleaved along the main chain peptide bonds to monomers (Fig. 4.03B). Both functions represent canonical peptide bond cleavages.

In addition to universal serine protein attributes, the acI sequences also contained a novel feature. A cysteine preceding the catalytic serine was present in all acI *Actinobacterial* sequences. This may be a secondary nucleophile/general base for the catalytic serine

Two genes cluster into a nitrogen related neighborhood. The acI gene is almost certainly related to nitrogen metabolism in acI, whether a cyanophycinase or a peptidase. I find that in addition to encoding this gene, acI single-cell amplified genomes (SAGs) and metagenome-assembled genomes (MAGs) likely encode the gene for Global Nitrogen Regulator, *ntcA* (Fig. 4.04). The gene is annotated as a general Cpr/Fnr transcription factor, and the gene product is homologous to cyanobacterial NtcA. The gene is also encoded directly upstream of the putative nitrogen-related enzyme, CphB/PepE. NtcA in cyanobacteria regulates the global nitrogen cascade and is required for expression of many ammonia repressible genes. When compared to NtcA from *Synechocystis* sp. PCC 6803 and *Anabaena* sp. PCC 7120, the overall alignment statistics are: 79% coverage with an E-value of 8E-22 and 26% identity, and 79% coverage with an E-value 6E-22 and 25% identity, respectively (Fig. 4.06). The residues for binding oxo-glutarate, leucine 78 and arginine 88 (7120), are also conserved across the protein sequences (22, 23). Basic analysis of NtcA DNA-binding motifs was performed, but a direct match for the GTANNNNNNNNTAC consensus sequence was not found upstream of the starts of the NtcA-like gene or the CphB/PepE-like gene. This does not rule out regulation but may require a modified search sequence GTNNNNNNNNNAC.

Purification of the acI CphB/PepE-like protein. In order to characterize the biochemical activity of acI CphB/PepE, the target protein first needed to be purified. After IPTG-based induction, nickel affinity and gel filtration chromatography, and final concentration, over 3 ml of pure protein at 32 mg/ml was obtained. Total amounts of protein (8 µg and 2 µg) loaded on an SDS-PAGE gel confirmed purity and yield (Fig. 4.05A). PepE proteins can be monomers (PDB 3L4E, unpublished) or dimers (24). Significant dimer and monomer populations were observed for the acI protease during size exclusion as judged from a broad elution peak with a shoulder and the column calibration (Fig. 4.05B). Fractions pooled for further concentration aimed at capturing the entire rightmost segment of the peak with knowledge that some dimer population may exist.

The acI CphB/PepE-like protein does not degrade cyanophycin. Cyanophycinase activity is canonically measured by digestion of the full-length polymer suspended in agar (25). The plate appears transparent white due to precipitated cyanophycin and if cyanophycinase activity is observed, a clearing effect is seen. To test this clearing, 5 µL of protein samples and cell lysate samples were dropped onto a dried cyanophycin-overlay plate (Fig. 4.07). Purified protein and protein buffer samples did not display noticeable clearing zones when compared to an unspotted blank region. A fresh cell lysate not expressing CphE from *Pseudomonas alcaligenes* (a secreted cyanophycinase (25)) and lysate sample buffer also showed no clearing. A large zone of clearing was observed for CphE-containing cell lysate indicating that cyanophycin polymers were effectively broken down.

acI contain a PepE that degrades an aspartate-leading dipeptide. Aspartate dipeptidase E activity is canonically measured by digestion of aspartate-leading dipeptides and analysis by HPLC (15) and/or measurement of absorbance at 405 nm as aspartate-p-nitroaniline is broken down (26). The dipeptide analog, aspartate-p-nitroaniline, is cleaved to aspartate and p-nitroaniline; the second molecule is responsible for the increase in absorbance. To test the ability of the acI protein to degrade peptides in a cost-effective manner, the chromophore-based assay was selected. Purified protein was suspended in aspartate-p-nitroaniline-containing buffer and incubated at room temperature. Absorbance for the protein-containing samples displayed increasing absorbance over time compared to protein buffer in the assay solution (Fig. 4.08).

The structure of acI PepE displays a monomeric enzyme with a canonical serine-histidine-glutamate catalytic triad. In order to confirm a serine-based active site and visualize the overall protein topology, the X-ray crystal structure of PepE from AAA027-L06 was solved (PDB 6NQC). The crystal used for the solution was mounted in a loop directly from the crystal drop and plunged into liquid nitrogen (Fig. 4.09A). Data were collected to 1.8 Å resolution (Fig. 4.09B). The structure was solved by molecular replacement using a previously solved structure of a putative dipeptidase (PDB 3L4E) and refined to 1.93 Å resolution. Model geometry and

Rwork/Rfree values of 17.6/21.8 % (Table 4.01 Table 4.02) were obtained as of 2018.12.21. The asymmetric unit has one protein molecule and one sulfate ion (Fig. 4.10). Crystal packing suggests acI PepE is a monomer because no substantial hydrophobic interactions are seen between symmetry mates. However, the N-terminal twenty-one amino acid tag, a hexahistidine, TEV recognition site, and various linkers (MGSSHHHHHHSSGENLYFQGH), extends towards the active site of an adjacent protein chain with possible unresolved contacts. All main chain atoms of the native protein sequence are clearly resolved with under 5% of sidechains not fully resolved. Unresolved sidechains are highly solvent exposed. Density is also observed for the C-terminal residues of the tag, but the certainty quickly drops for TEV recognition residues. No density for any of the hexahistidine region is seen.

Topologically, the protein follows an alternating motif of strand-helix for most of the sequence. A large eight-stranded β-sheet forms the core of the monomeric protein with a smaller four-strand sheet abutting one side. The smaller sheet is relatively flat while the large sheet displays a ~120 degree right handed twist when looking down the rotation axis. The tip of a ridge formed along the junction of these two β-sheets contains a serine133-histidine168-glutamate198 catalytic triad (Fig. 4.10). The active site triad is not contained in the strands of either sheet but rather in the loops at the crest of the ridge. The positioning allows ample access to the aqueous environment, unlike the closed dimeric interface seen in the *S. enterica* serovar Typhimurium structure (24).

The conformations of key active site residues remain unchanged from canonical serine triad peptidases without ligand (Fig. 4.11) (9). Specifically, the active site serine 133 sits in a G-X-S-X-G sequence motif (Fig. 4.02, Fig. 4.11). A sharp turn in this hairpin region is required for proper placement of the catalytic residue in a fully solvent exposed pocket. The two glycines are found in favored regions of the Ramachandran plot, and help complete the turn. Side chains larger than hydrogen at either of these positions would sterically disrupt the serine position. The active site histidine 168 coordinates the catalytic serine. A general base, like histidine, must abstract the

proton from the serine hydroxyl to generate a potent nucleophile. Additionally, the glutamate 198 of the triad stabilizes the histidine on the side opposite to the serine. Both the serine and glutamate oxygen atoms rest within the plane of the imidazole ring.

The catalytic serine is positioned such that histidine 168 acts as a general base for the sidechain proton. Glutamate 198 positions histidine 168 for proton abstraction. Alanine 132 and glycine 100 form the oxyanion hole important for tetrahedral intermediate stabilization during peptide cleavage and enzyme regeneration. Other locations of interest include the C-terminal residue, sulfate trap region, and aromatic residues in general (Fig. 4.12). The terminal phenylalanine is buried into a hydrophobic core of the protein rather than towards the solvent. The motif seems to provide a catch hook, so the C-terminus of the protein is stably positioned (Fig. 4.12A). Additionally, a sulfate ion appears in the crystal structure (Fig. 4.12B). The oxygens appear to interact with three positively charged residues: lysine 56, lysine 71, and arginine 60. While not fully resolved likely from flexibility, the nitrogen-rich sidechains appear to act as a tripronged pincer to trap the polyatomic ion. The role of the sulfate is unknown and may be a result of over 1M sulfate present in the crystallization liquor. An additional area of interest is the center of most aromatic residues (Fig. 4.12C). Because of high data quality and accurate phases, electron density puckering is seen towards the middle of conjugated rings in tyrosine and phenylalanine. The forming annulus represents the lack of electrons in the center of the sidechain and agrees with chemical understanding of aromatic substituents.

Discussion

Prior to this work, there was no biochemical evidence to support the hypothesis that acI *Actinobacteria* use cyanophycin or any peptide-based substrate as a source of nitrogen. I have now provided experimental verification of an advantage that allows the acI lineage to metabolize nitrogen-containing substrates. The process may be one reason why acI are so abundant. Specifically, I show that an acI protein annotated as both a cyanophycinase and a dipeptidase

can break down an aspartate-leading dipeptide, homologous to aspartate-phenylalanine, to free amino acids. The purified enzyme has no activity against high molecular weight cyanophycin. Additionally, the structure of the now classified PepE was determined to better understand the structural contribution to dipeptide cleavage.

The acI aspartyl dipeptidase is homologous to both characterized cyanophycinases and other aspartyl dipeptidases (Fig. 4.01, Fig. 4.02). Thus, purely from a protein sequence argument, it was equally likely for the protein to breakdown cyanophycin, polymers composed of aspartate backbones with arginine sidechain modifications, or aspartate-leading dipeptides, short molecules composed of two amino acids (FIG 4.03). Either way, the gene product was interesting because the gene was on a list of 10-20 genes thought to be horizontally transferred to acI genomes (Weizhou Zhao, Siv Anderson lab, personal communication). The list was populated by genes likely found in most acI genomes but not found in related *Actinobacteria*. The list also included carotenoid biosynthetic genes discussed in Chapter III.

Regardless of function, the protein was likely to be involved with nitrogen metabolism because of its genomic placement downstream of the acI homolog to the global nitrogen-regulating transcription factor, NtcA (Fig. 4.04, Fig. 4.06). NtcA has been shown to be the master regulator of nitrogen-related gene expression in various cyanobacteria like *Anabaena* sp. PCC 7120 and *Synechocystis* sp. PCC 6803 in response to oxo-glutarate (22, 27). Specifically, *gln* (regulatory and glutamate related genes), *nrt* (nitrate transport genes), *amt* (ammonium permease genes), and *urt* (urea transport genes) genes have all shown changes in expression during the NtcA response. Logically, the nitrogen transcription system regulates the genes related to nitrogen-containing compounds and their transport. However, work in *Nostoc* 7120 has also shown that ketocarotenoid biosynthetic genes are regulated by the protein (28). Two different carotenoid ketolases appear in the *Nostoc* genome, and both contain upstream binding sequences for NtcA, GTANNNNNNNNTAC. In response to binding and the presence of light, extra echinenone and/or canthaxanthin are produced. Therefore, a process whereby acI respond

to lack of nitrogen by triggering carotenoid and perhaps actinorhodopsin production could be hypothesized. Light would then serve to create a proton gradient to continue basic metabolic function until more molecules for growth are encountered. Nitrogen nutrients are especially important for growth as they help synthesize amino acids and nitrogenous bases.

Even though the regulation of *pepE* expression by NtcA was not demonstrated, two major potential catalytic activities of the enzyme were conclusively tested. The PepE from acI *Actinobacteria* was not able to breakdown cyanophycin polymers when compared to a known cyanophycinase, CphE (Fig. 4.07). This lack of activity is not entirely unexpected due to the enzyme and substrate likely being on either side of the acI cell wall; on one hand, acI PepE proteins do not contain an identifiable secretion signal sequence, and on the other, acI *Actinobacteria* would not likely be able to import the 25-100 kDa cyanophycin polymer. Peptide and oligopeptide transporters are annotated in acI genomes, but their substrates would likely be limited in length to a few amino acids (29). Cyanobacteria can only mobilize the polymer because it precipitates as granules inside of the cells for use later as a carbon and nitrogen source (30). PepE was confirmed to break down an aspartate-leading dipeptide, aspartate-p-nitroaniline (Fig. 4.08), an assay which has been used a gold-standard measure of PepE activity in *Salmonella enterica* serovar Typhimurium and *Xenopus laevis* (15). With the two assays in mind, the acI protein can certainly be classified as a PepE.

In addition to determining a PepE function, the structure of the acI protein was determined to 1.93 Å resolution by crystallization in high ammonium sulfate and X-diffraction analysis (PDB 6NQC) (Fig. 4.09, Table 4.01, Table 4.02). The catalytic triad is composed of serine 133, histidine 168, and glutamate 198, and the residue positions and geometry match known and putative peptidase E structures (9). While the roles for serine and histidine as nucleophile and general base are clear, the glutamate is too distant for a direct role. The acidic residue is necessary for robust catalysis in well-studied serine proteases, and it is conserved among available aspartyl dipeptidases (15). The contribution of the glutamate to hydrolysis was tested by generation of a

Salmonella Typhimurium PepE alanine variant. Activity on an aspartate-leucine substrate was approximately 1% of the wild type (0.021 nmol/minμg *vs* 2.13 nmol/minμg) (9).

Even though a broad range of aspartate dipeptides are cleaved by canonical PepE enzymes, substrates larger then dipeptides (and one highly flexible tripeptide) are not. Discrimination against larger peptides and even proteins results from restricted steric access (9, 24). In the case of the acI enzyme, however, it should be noted that it has not been tested with larger substrates, like dimers and trimers of cyanophycin Asp-Arg blocks.

Another unresolved area for the peptidase E from acI is the role of a small region present after the general base histidine. In AAA027-L06, the region has a sequence FNKFFKNIPDSAA with most of the sequence forming a loop (Fig 4.02). In *S. enterica* serovar Typhimurium the sequence is FTNALPEGHKGET (Fig 4.02), and the entire structure is a loop. However, the loop is not unstructured; it provides dimerization interactions and the glutamate and arginine make substrate pocket contacts with the other protomer (24). This lid severely limits solvent accessibility. In another structure of the same protein, the loop is unresolved and no ligand is seen (9). These regions have extraordinarily different compositions. The acI sequence has many aromatic residues that are missing in the in *S. enterica* serovar Typhimurium sequence. It should be noted that the N-terminal tag of the protein in this study extends towards this aromatic region, possibly leading to its clearly resolved structure. Additionally, tag interactions with other protein chains may help explain the unusual monomer-dimer equilibrium seen in size exclusion (Fig. 4.05B) and dynamic light scattering (data not shown).

Overall, the acI aspartyl dipeptidase, PepE represents another example of serine hydrolases and it displays a typical strand-helix overall polypeptide fold and canonical catalytic residues. The protein cannot degrade cyanophycin in contrast to predictions by database annotations and bioinformatic studies of acI genomes (31–33). However, smaller versions of cyanophycin should be tested for breakdown. It is feasible to imagine high amounts of free-floating degraded cyanophycin in a lake environment. As a result, acI may still be able to use cyanophycin

peptides marking a new class of peptidase in the literature. Similarly, the role of the small region after the general base histidine should be determined. Perhaps the region is a good signal of whether a given PepE will form a preferential monomer or dimer with ligand. Either way further study of acI PepE is warranted and may help uncover unknown characteristics of PepE mechanisms.

Methods

acI gene identification and pathway assembly. Annotations for multiple acI SAGs (31) were analyzed for cyanophycinase genes using the Joint Genome Institute's Integrated Microbial Genomes Viewer (32). Translated candidate protein sequences were used to identify homologs in other acI genomes, and those with consistent gene neighborhoods. After narrowing possible functions to peptidase E or cyanophycinase, two pathways for dipeptide and cyanophycin cleavage were mapped. The Basic Local Alignment Search Tool (BLAST) and Clustal Omega were used for identity percentage calculation and to align proposed acI proteins to known examples of both enzymes (34, 35) (36–38).

Plasmid construction. All cloning was performed using Phusion high-fidelity (HF) GC master mix (Thermo Fisher) and the listed primers (Table 1.04). The L06 protein sequence was obtained as an *E. coli*-optimized gBlock from IDT with an N-terminal hexahistag and TEV protease cleavage site and a C-terminal stop codon. Genes were placed into the pET21b+ (Novagen) site between NdeI and NotI restriction sites by ligation-free recombination in *E. cloni* 10G cells (Lucigen). The insert molarity was up to 10-fold more abundant than that of appropriately digested and purified backbone. After selection on 100 µg/ml disodium carbenicillin LB-Miller agar plates, insert presence was validated by colony PCR using GoTaq Green master mix (Thermo Fisher). Restriction enzyme digestion in lab and Sanger sequencing at the UW Biotechnology Center further confirmed plasmid correctness.

Protein purification. The protein from genome L06 was selected for further characterization because it originated from a highly studied genome and highly complete genome (33). BL21(DE3) Tuner *E. coli* cells transformed with a plasmid expressing acI-PepE were grown in half full 2-liter flasks of LB-Miller broth plus 100 µg/ml disodium carbenicillin and shaken in 37°C at 250 rpm until an optical density at 600 nm of 0.6. Cells were induced with 100µM IPTG and shaken in 16°C at 250 rpm for 20 h. Cells were centrifuged at 3,300 g, washed in 100 mM Tris (pH 6.8), centrifuged again, and frozen in liquid nitrogen. Around 8g of cells was suspended in 5 g/ml lysis buffer (50 mM Tris-HCl [pH 8.0], 300 mM NaCl, 1 mg/ml lysozyme, 20 g/ml DNase I, 5 mM MgCl2, and 130 µM CaCl2. Cells were lysed at 16,000 lb/in2 by four passes through a French pressure cell. Centrifugation at 25,000 g for 20 min cleared debris. Lysate was filtered through 0.22 µm syringe filters and loaded at 2.5 ml/min onto an equilibrated 5 ml Ni-nitrilotriacetic acid (NTA) column. The processing profile in column volumes (CV) at 5ml/min was as follows: 20 lysis buffer, 40 gradient to 100% elution buffer (50 mM Tris-HCl [pH 8.0], 300 mM NaCl, and 500 mM imidazole). The elution fractions were pooled and dialyzed against 4 liters 4°C final buffer (30 mM maleic acid [pH 6.7], 200 mM Na_2SO_4) for 20 h. The resulting solution was centrifuged for 15 min at 25,000 g and concentrated in 15 ml 10kDa MWCO centrifugal concentrators. Once at 6 mL, two rounds of 3mL injections on a Superdex 200 HiLoad column were performed in the same buffer (Fig. 4.05B). Peak fractions were pooled, concentrated to 32 mg/ml, and vitrified as small aliquots. SDS-PAGE was performed on calculated amounts of protein, and the gel was stained with Coommassie blue (Fig. 4.05A).

Cyanophycin purification and degradation assay. CphA from Synechocystis PCC 6803 was overexpressed in BL21 (DE3) cells transformed with pCOLADuet1::cphAC595S-dapLSs (39). Expression was as for acI protein except that all incubations were performed in Terrific Broth, induction occurred at 1mM IPTG and 30°C, and selection was achieved via 100 µg/mL kanamycin sulfate. After collection and freezing of 40g biomass, cyanophycin purification was started. Suspension by vortexing in 80ml 0.1M HCl at 2 ml/g was followed by ~70 drops of

concentrated HCl to bring the pH to 1.00. The suspension was stirred by magnetic bar in a beaker on high at 4°C for 20 h. After centrifugation at 9000 g for 30min at 4°C, the liquid was decanted and brought to pH 7.00 using 10M NaOH. After centrifugation, the insoluble cyanophycin was suspended in 0.1M HCl to 75mL and processed through two more rounds of neutralization and harvest. The final pellet was dried at 65°C overnight.

Isopropanol was used to sterilize cyanophycin for 15 min. The suspension was centrifuged, excess alcohol was aspirated, and cyanophycin was dried for 5 min at room temperature. The sterilized cyanophycin was suspended in 0.1M HCl at 0.1g/3mL. This suspension was dripped into 50 ml 0.5% m/v sterile warm liquid agar at 0.2% m/v while stirring on extra high. Carbenicillin to 100 µg/mL and an equivalent cyanophycin solution volume of 0.1M NaOH was also dripped into the vigorously stirring suspension. Upon base addition, cyanophycin precipitated, creating an even suspension rather than solution. 5 ml was overlaid by serological pipette onto a normal 1.5% m/v LB-agar plate. 10 ml was also overlaid if a thicker layer was desired.

Purified acI protein and lysates from control, acI protein, and CphE expressing cells (39) (cells were produced in the same manner as for acI protein) were plated onto the cooled dry cyanophycin overlay plate along with buffer controls. All samples were spotted as 5 µl aliquots. Plates were incubated at room temperature for 1 day and then checked for zones of clearing. Plates were dyed with Coomassie Blue if desired for better visualization.

Aspartate-p-nitroaniline degradation assay. Purified acI protein (20 µl at 32 mg/mL) was suspended in 980 µl of 50mM imidazole, 5mM aspartate-p-nitroaniline (Chem Impex) at room temperature in the dark. The sample was incubated in a Beckman Coulter DU640B spectrophotometer and the absorbance at 405 nm was taken at various time points. The control solution was subjected to the same assay except the 20 µl addition was protein buffer only. The assay was repeated in sets of two three times with the same ligand stock solution (i.e. a protein sample and control sample were started seconds apart).

Protein crystallization and structure determination. Crystals of acI enzyme were grown by hanging drop vapor diffusion after mixing 2 µL 16 mg/ml protein in final purification buffer with an equal volume of condition 7 (100 mM HEPES-NaOH [pH 7.5], 2% v/v PEG 400, 2 M $(NH_4)_2SO_4$) from the TOP96 screen (Anatrace) over 500 µl of the screen solution. The crystals were dyed blue with 0.5 µl Izit crystal dye (Hampton Research) after 4 weeks of growth (Fig. 4.09A). Crystals used for data collection were mounted in a loop directly from the crystal drop and flash frozen in liquid nitrogen. Two data sets were collected at the Life Sciences Collaborative Access Team beamline 21-ID-F at the Advanced Photon Source, Argonne National Laboratory at a detector distance of 260 mm: 1) 0.5 s exposure with rotation range between frames of 0.5° for 360° and 2) 1 s exposure with rotation range between frames of 1° for 180° (Fig. 4.09A, bleached region and crosshairs region). Crystals diffracted to 1.8 Å (Fig. 4.09B) and belonged to the monoclinic space group, $C222_1$, with cell dimensions a=44.69 Å, b=88.92 Å and c=131.31 Å. Only the first 180° of the first data set was used for final scaling and structure determination. There was no discernible difference in the solutions for either set. Data were integrated using HKL2000 and scaled with SCALA (40). The structure was solved using molecular replacement and an ITASSER-generated model based on a putative dipeptidase (PDB 3L4E, unpublished) (41-43). The model had active site geometry modifications completed in Sybyl (44). Rfree were generated after an acceptable Phaser solution was found (5%). Rounds of building were carried out in real space (COOT) (45) with alternating reciprocal space refinement (PHENIX) (46-49). PyMOL was used to visualize protein structures (50). The model was assigned Protein Data Bank Code, 6NQC.

This research used resources of the Advanced Photon Source, a U.S. Department of Energy (DOE) Office of Science User Facility operated for the DOE Office of Science by Argonne National Laboratory under Contract No. DE-AC02-06CH11357. Use of the LS-CAT Sector 21 was supported by the Michigan Economic Development Corporation and the Michigan Technology Tri-Corridor (Grant 085P1000817).

Figures and Tables

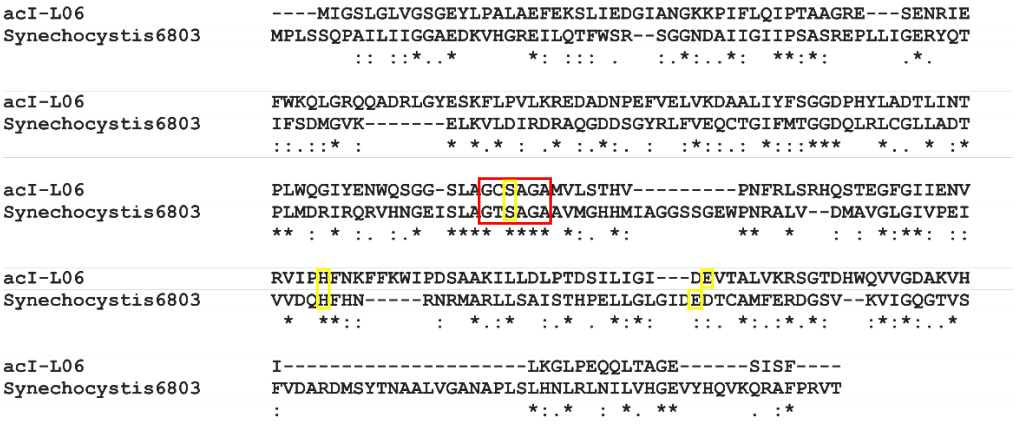


FIG 4.01 Alignment to a cyanophycinase (CphB). The protein from acI genome AAA027-L06 is aligned to a functionally characterized cyanophycinase from *Synechocystsis* sp. PCC 6803. In this and following alignment figures, identical (*), highly similar (:), and slightly similar (.) residues are indicated. The catalytic serine (yellow box) is part of the G-X-S-X-G-X motif (red box). Other catalytic triad residues are also boxed in yellow (H, E).

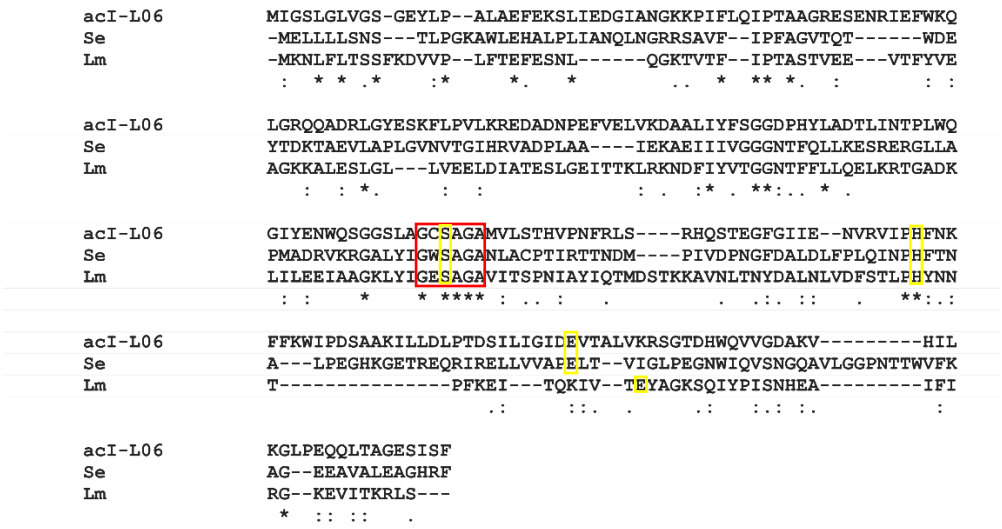


FIG 4.02 Alignment to two peptidase E (PepE). The protein from acI genome AAA027-L06 is aligned to a functionally characterized peptidase E from *Salmonella enterica* serovar Typhimurium (Se) and *Listeria monocytogenes*. The catalytic serine is part of the G-X-S-X-G-X motif (red box). Other catalytic triad residues are also boxed in yellow (H, E).

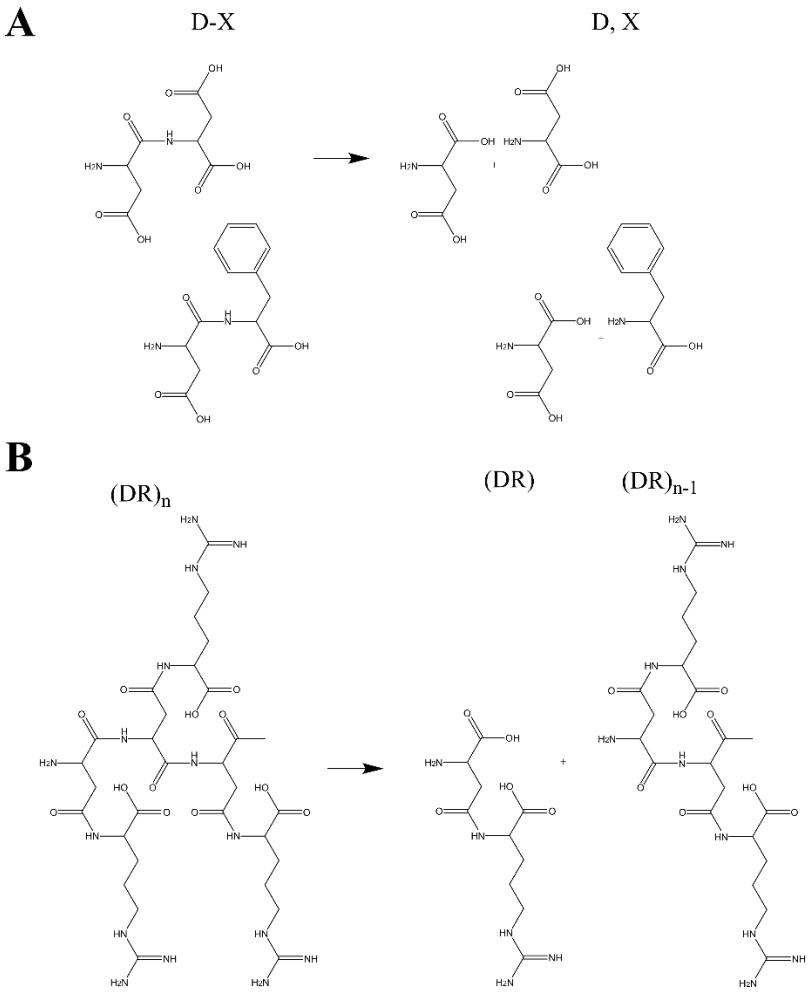


FIG 4.03 Possible reactions for an acI protein based on alignments. (A) The protein from acI genome AAA027-L06 could catalyze a reaction that takes aspartate leading dipeptides and cleaves them to free amino acids. Aspartate-aspartate and aspartate-phenylalanine are shown as represented examples of this reaction. The canonical peptide bond is cleaved in the reaction. (B) The protein from acI genome AAA027-L06 could catalyze a reaction that cyanophycin polymer and cleaves off one unit of the aspartate-arginine chain. A tripeptide of the polymer represents the entire much longer polymer. The canonical peptide bond is cleaved in the reaction while the linkage between the aspartate sidechain and arginine amino group is maintained.

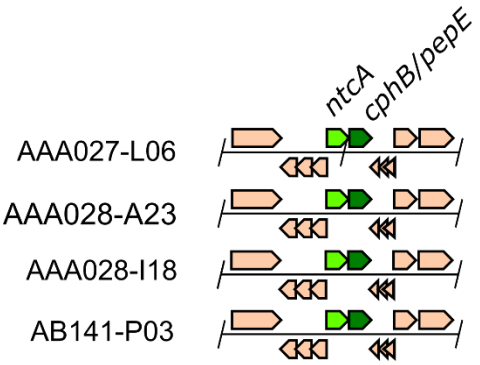


FIG 4.04 acI *cphB/pepE* in the genomic context. acI genomic contigs are primarily labeled by shorthand notation (i.e., designation after "*Actinobacterium* SCGC"). Genes are arrows pointing in the direction of transcription. A slash indicates a contig boundary, which may or may not end immediately after the pictured gene. Tan genes are neighboring genes that may or may not be functionally associated with the nitrogen-related genes. A gene source for this study is AAA027-L06. Relevant Integrated Microbial Genomes (IMG) database locus tags for this study from AAA027-L06 are A27L6_0024.00000190 (*ntcA)* and A27L6_0025.00000020 (*cphB/pepE*). All of these genomes are from acI-B1. The L06 assembly has partial reads for genes at the contig boundary, and the genes match the assembly as pictured.

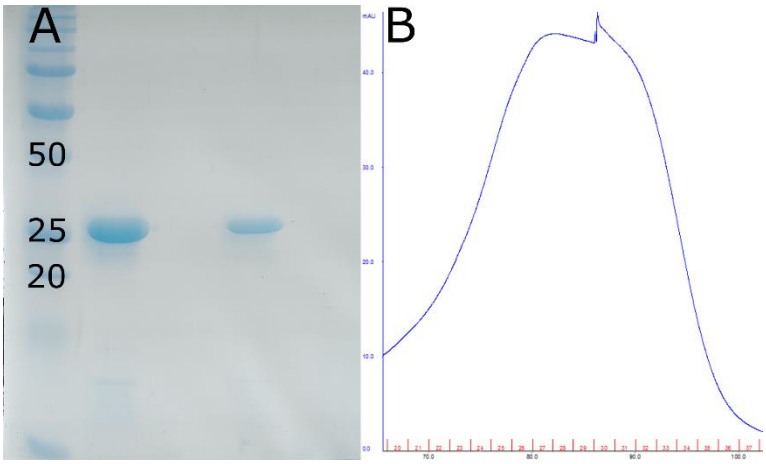


FIG 4.05 Purification of AAA027-L06 CphB/PepE. (A) Target protein purified by nickel affinity and gel filtration chromatography was concentrated to 32 mg/ml and run on an SDS-PAGE gel at 8µg and 2µg total protein. Numbers indicate the molecular weight in kDa of protein standards. The purified protein has a sequence-based size of 28.6 kDa (B) An example size exclusion run depicts superimposition of a dimer (left side) and monomer (right side) populations. Oligomer estimates are based on column calibration with protein standards, conalbumin, ovalbumin, and RNAse A.

```
acI       MPNKNDDEVVRR---AALFTALDDASAATLRASMSGIKISKGQILFKEGDPGDRLFVVVE      57
7120      ------MIVTQDKALANVFRQMATGA---FPP--VVETFERNKTIFFPGDPAERVYFLLK      49
6803      ----MDQSLTQDRPLAAVFRRLGSEL---MPP--VVETFDRSKTIFFPGDPAERVYFLLK      51
                 :.:    * :*  :       :       .:.:.: :*  ***.:*::.:::

acI       GKLKLGTSSGDGRENLLSILGPGDMFGELSLFDPGP--RTATATAVVDSRLLTLANDQVI      115
7120      GAVKLSRVYEAGEEITVALLRENSVFGVLSLLTGNKSDRFYHAVAFTPVELLSAPIEQVE      109
6803      GAVKLSRVYEAGEEITVALLRENSVFGVLSLVTGQRSDRFYHAVAFTPVELLSAPIEQVE      111
          * :**.     *.*  :::*  ..:** ***.      *   *.*..  .**:   :**

acI       GWVTAHPEVSLQLLGRLAQRLRKANDVLSDLVFADVPGRVAKAIIELGERFGTKKDDGLH      175
7120      QALKENPELSMLMLRGLSSRILQTEMMIETLAHRDMGSRLVSFLLILCRDFGVPCADGIT      169
6803      QALKEHPDLSLLMLQGLSSRILQTEMMIETLAHRDMGSRLVSFLLILCRDFGVPAPDGIR      171
            :. :*::*: :*  *:.*: ::: ::. *.. *: .*:.. :: * . **.   **:

acI       VNHELTQEELAQLVGASRETVNKALADFATRGWVKLEPRAVIVLDYERLVKRGR             229
7120      IDLKLSHQAIAEAIGSTRVTVTRLLGDLREKKMISIHKKKITVHKPVTLSRQFT             223
6803      IDLKLSHQAIAEAIGSTRVTVTRLLGDLREGNMISITKKKITVHNPVALSQQFT             225
          :: :*::: :*: :*::* **.: *.*:     :.:  : : * .   * ::
```

FIG 4.06 Alignment to a nitrogen-control transcription factor (NtcA). The protein from acI genome AAA027-L06 is aligned to a functionally characterized NtcA from *Synechocystsis* sp. PCC 6803 (6803) and Anabaena sp. PCC 7120 (7120). The residues for oxo-glutarate binding in 7120 are also conserved, leucine 78 and arginine 88 (red box).

FIG 4.07 Lack of cyanophycin degradation by acI CphB/PepE. An LB agar plate with an overlay of heterologously-produced cyanophycin. The overlay is 5 ml and contains 0.5% m/v agar and 0.2% m/v cyanophycin. Samples tested include: acI protein buffer, purified acI protein, lysate buffer, control cell lysate, CphE containing cell lysate, and acI protein containing lysate. Spottings are 5 µl drops that were incubated at room temperature for one day before staining with Coomassie Blue and imaging. The blank designation denotes a control area of the plate where nothing is spotted.

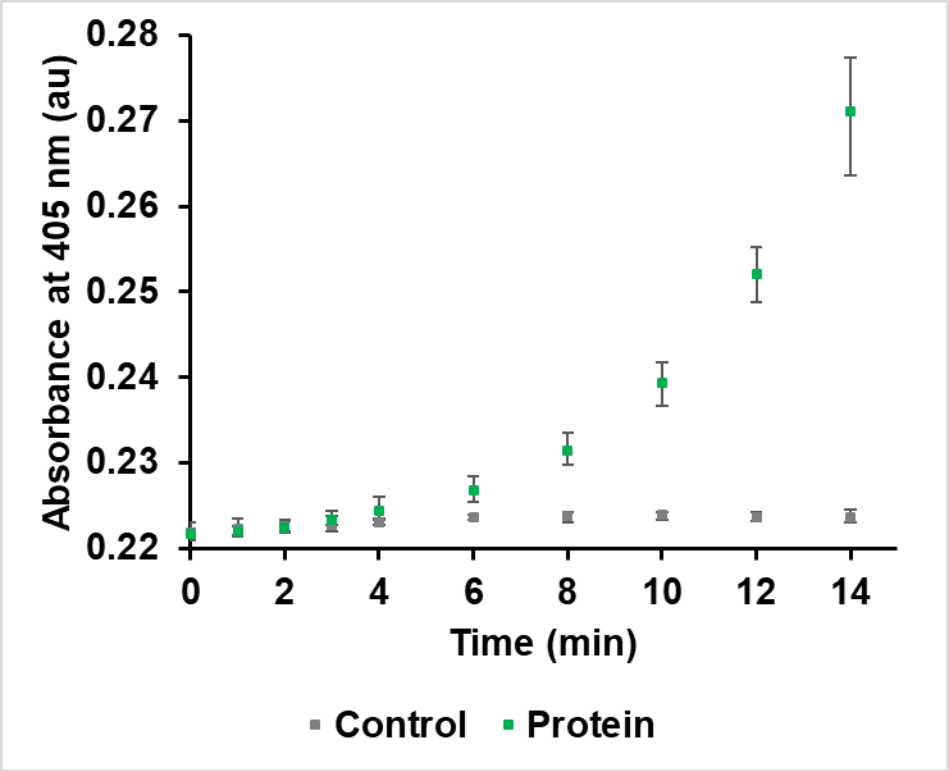
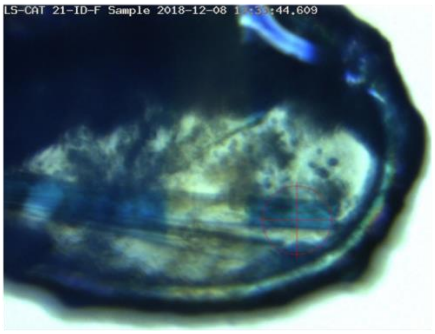


FIG 4.08 Cleavage of aspartate-p-nitroaniline by acI CphB/PepE. Three control assays of protein buffer in aspartate-p-nitroaniline buffer were performed. Similarly, three purified proteins samples in the same assay solution were performed. Absorbance was monitored at 405 nm because this is the maximal absorbance of p-nitroaniline when free in solution.

A



B

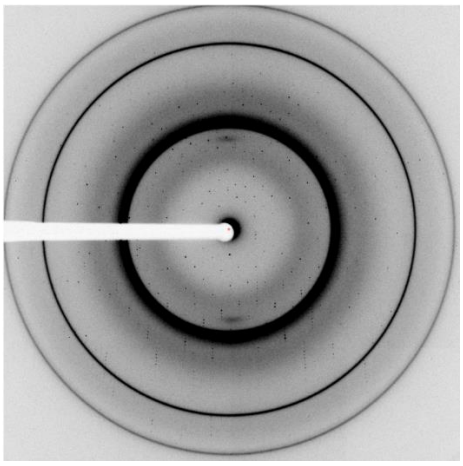


FIG 4.09 Crystal mounting and diffraction of acI PepE. (A) A protein crystal of acI PepE that was mounted from the crystallization condition and flash frozen. The blue color is from Izit crystal dye that was introduced to detect protein crystals. The non-blue stripe down the center of the image and crystal is bleached dye from a round of data collection before the image was taken. (B) X-ray diffraction of the looped sample in A. Water diffraction rings at three locations (black circles) are seen because the sample was not cryoprotected before vitrification in liquid nitrogen. The white cutout extending from the left is the beam stop.

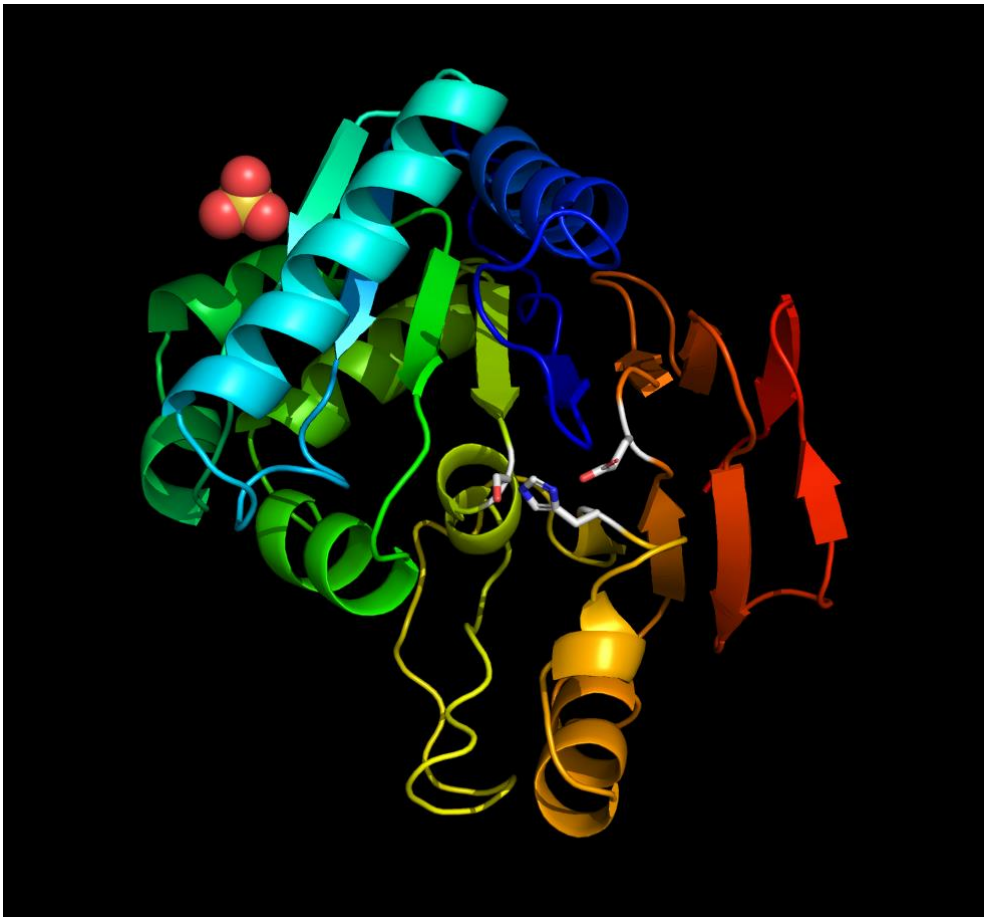


FIG 4.10 Structure of acI-PepE (PDB 6NQC). The structure of PepE from AAA027-L06 is a monomer shown as chainbow cartoon. The N-terminus is dark blue, and the C-terminus is red. A sulfate ion found in the structure is shown as spheres with a yellow sulfur atom and red oxygen atoms. Residues of catalytic triad (serine 133, histidine 168, and glutamate 198) are shown as sticks with atom coloring scheme (carbon gray, nitrogen blue, oxygen red). The model is current as of 12.21.2018.

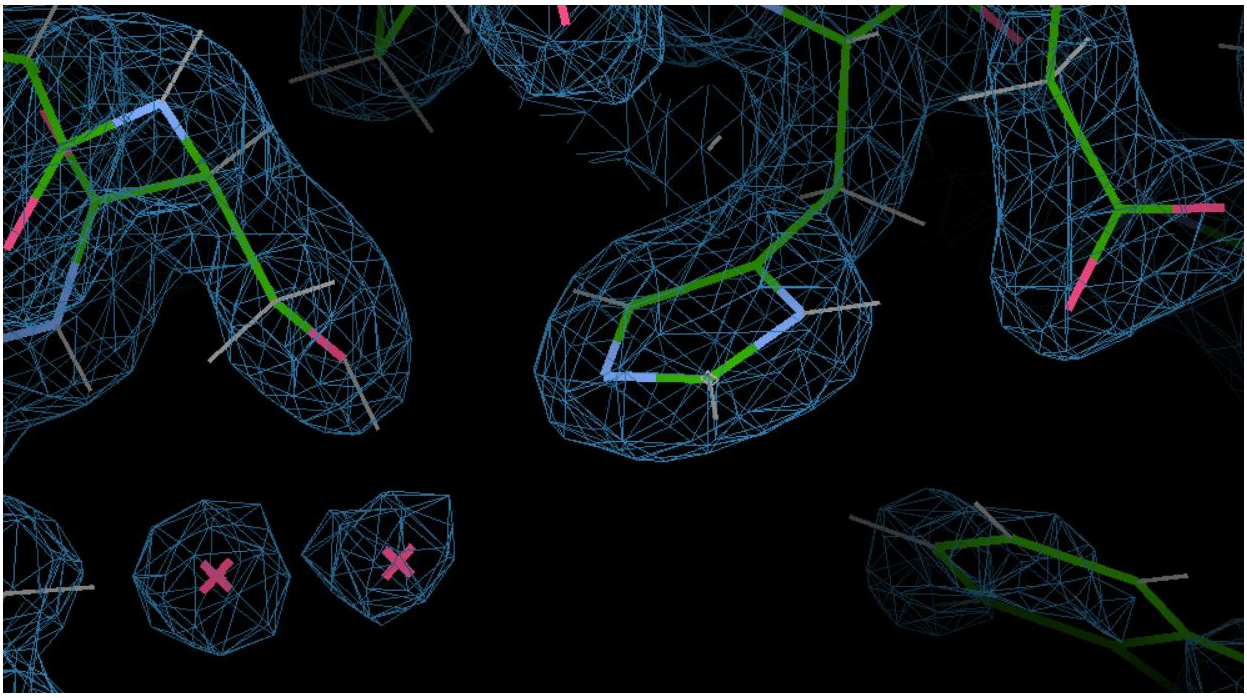


FIG 4.11 Electron density for the active site of the acI-PepE structure (PDB 6NQC). In this and following figures, images depict the protein model (sticks) with a mesh overlay ($2F_{obs} - F_{calc}$ electron density contoured at 1.50 standard deviations above the map mean). Carbons, nitrogens, oxygens, and hydrogens are shown in green, blue, red, and gray, respectively. Hydrogens are not refined and are shown for a more realistic interpretation. Water molecules are indicated by red jacks and indicate the position of the oxygen atom. The catalytic triad is made of serine 133, histidine 168, and glutamate 198 (left to right). The model is current as of 12.21.2018.

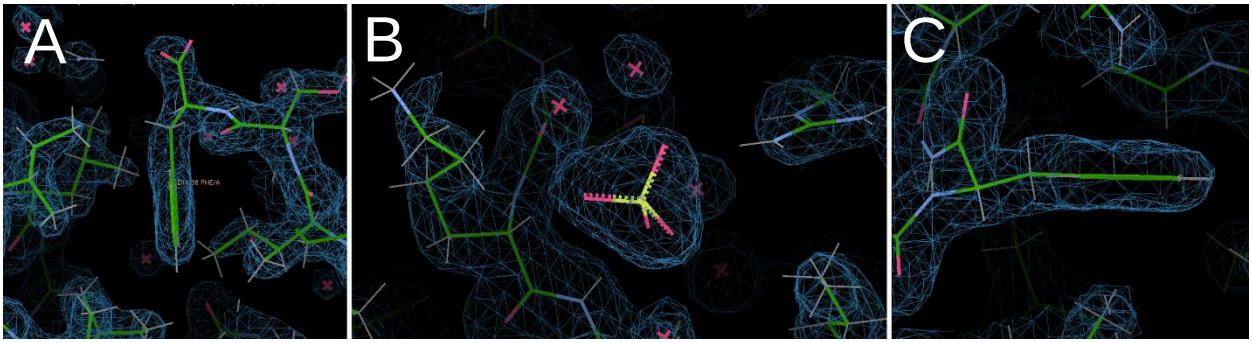


FIG 4.12 Electron density for three interesting attributes of the acI-PepE structure (PDB 6NQC). (A) The terminal residue of the protein chain is phenylalanine 238; it displays tight packing with the internal core of the protein. (B) The only bound sulfate ion is loosely trapped by positively charged lysines and arginine. The sulfur atom is shown in yellow. (C) Many aromatic residues display center puckering. Phenylalanine 20 is displayed as an example.

Table 4.01 Data collection and processing statistics for AAA027-L06 PepE crystal structure (PDB 6NQC)

Diffraction Source	Advanced Photon Source, LS CAT 21-ID-F
Wavelength (Å)	0.97872
Temperature (K)	100 K
Detector	Rayonic MX-300
Crystal-detector distance (mm)	260
Rotation range per image (°)	0.5
Total rotation range (°)	180
Exposure time per image (s)	0.5
Space Group	C 2 2 21
a, b, c (Å)	44.7, 88.9, 131.3
α, β, γ (°)	90.0, 90.0, 90.0
Mosaicity (°)	0.34 - 0.50
Resolution Range (Å)*	26.5 - 1.94 (1.97 - 1.94)
Unique reflections	18380 (953)
Completeness (%)	91.6 (98.8)
Redundancy	6.8 (6.8)
I/σ(I)	26.4 (3.14)
R_{pim}	0.034 (0.303)
Overall B factor from Wilson plot ($Å^2$)	21.30

Table 4.02 Structure refinement statistics for AAA027-L06 PepE crystal structure (PDB 6NQC)

Resolution Range (Å)	26.41 - 1.94 (2.03 - 1.93)
Completeness (%)	91.1 (95.0)
σ cutoff (Å)	1.34
Number of reflections, working set	18310 (2554)
Number of reflections, test set	917 (134)
R_{cryst}	0.164 (0.176)
R_{free}	0.227 (0.224)
Number of non-H atoms	2066
Protein	1898
Ligand	5
Water	163
RMS Deviations	
Bonds (Å)	0.65
Angles (°)	0.57
Average B factors (Å^2)	28
Protein	28
Ligand	40
Water	28
Ramachandran Plot	
Favored regions (%)	96
Additionally allowed (%)	4
Outliers (%)	0.00

References

1. Hedstrom L. 2002. Serine Protease Mechanism and Specificity. Chem Rev 102:4501–4524.

2. Dodson G. 1998. Catalytic triads and their relatives. Trends Biochem Sci 23:347–352.

3. Blow DM. 1969. The Study of a-Chymotrypsin by X-Ray Diffraction 11.Complete reference

4. BottS R, Ultsch M, Kossiakoff A, Graycar T, Katz B, Power S. The Three-dimensional Structure of Bacillus arnyloliquefaciens A Subtilisin at 1.8 and anAnalysis of the StructuralConsequences of Peroxide Inactivation 12. Fix reference

5. Liao DI, Breddam K, Sweet RM, Bullock T, Remington SJ. 1992. Refined atomic model of wheat serine carboxypeptidase II at 2.2 angstrom resolution. Biochemistry 31:9796–9812.

6. Wang J, Hartling JA, Flanagan JM. 1997. The Structure of ClpP at 2.3 Å Resolution Suggests a Model for ATP-Dependent Proteolysis. Cell 91:447–456.

7. Das S, Vasanji A, Pellett PE. 2007. Three-Dimensional Structure of the Human Cytomegalovirus Cytoplasmic Virion Assembly Complex Includes a Reoriented Secretory Apparatus. J Virol 81:11861–11869.

8. Richter R, Hejazi M, Kraft R, Ziegler K, Lockau W. 1999. Cyanophycinase, a peptidase degrading the cyanobacterial reserve material multi-L-arginyl-poly-L-aspartic acid (cyanophycin). Molecular cloning of the gene of Synechocystis sp. PCC 6803, expression in Escherichia coli, and biochemical characterization of the purified enzyme. Eur J Biochem 263:163–169.

9. Hakansson K, Wang AH-J, Miller CG. 2000. The structure of aspartyl dipeptidase reveals a unique fold with a Ser-His-Glu catalytic triad. Proc Natl Acad Sci 97:14097–14102.

10. Mackerrasn AH, de Chazal NM, Smith G . 1990. Transient accumulations of cyanophycin in *Anabaena* cylindrica and *Synechocystis* 6308. J Gen Micro. 136:2057-65.

11. Simon RD, Weathers P. 1976. Determination of the structure of the novel polypeptide containing aspartic acid and arginine which is found in cyanobacteria. Biochim Biophys Acta BBA - Protein Struct 420:165–176.

12. Simon RD. 1971. Cyanophycin Granules from the Blue-Green Alga Anabaena cylindrica: A Reserve Material Consisting of Copolymers of Aspartic Acid and Arginine. Proc Natl Acad Sci 68:265–267.

13. Füser G, Steinbüchel A. 2007. Analysis of Genome Sequences for Genes of Cyanophycin Metabolism: Identifying Putative Cyanophycin Metabolizing Prokaryotes. Macromol Biosci 7:278–296.

14. Obst M, Sallam A, Luftmann H, Steinbüchel A. 2004. Isolation and Characterization of Gram-Positive Cyanophycin-Degrading BacteriaKinetic Studies on Cyanophycin Depolymerase Activity in Aerobic Bacteria. Biomacromolecules 5:153–161.

15. Lassy RAL, Miller CG. 2000. Peptidase E, a Peptidase Specific for N-Terminal Aspartic Dipeptides, Is a Serine Hydrolase. J Bacteriol 182:2536–2543.

16. Yen C, Green L, Miller CG. 1980. Degradation of intracellular protein in Salmonella typhimurium peptidase mutants. J Mol Biol 143:21–33.

17. Yen C, Green L, Miller CG. 1980. Peptide accumulation during growth of peptidase deficient mutants. J Mol Biol 143:35–48.

18. Carter TH, Miller CG. 1984. Aspartate-Specific Peptidases in Salmonella typhimurium: Mutants Deficient in Peptidase E. J Bacteriol 159:453-9.

19. Larsen RA, Knox TM, Miller CG. 2001. Aspartic Peptide Hydrolases in Salmonella enterica Serovar Typhimurium. J Bacteriol 183:3089–3097.

20. Brown DD, Wang Z, Furlow JD, Kanamori A, Schwartzman RA, Remo BF, Pinder A. 1996. The thyroid hormone-induced tail resorption program during Xenopus laevis metamorphosis. Proc Natl Acad Sci 93:1924–1929.

21. Law AM, Lai SWS, Tavares J, Kimber MS. 2009. The Structural Basis of β-Peptide-Specific Cleavage by the Serine Protease Cyanophycinase. J Mol Biol 392:393–404.

22. Olmedo-Verd E, Valladares A, Flores E, Herrero A, Muro-Pastor AM. 2008. Role of Two NtcA-Binding Sites in the Complex ntcA Gene Promoter of the Heterocyst-Forming Cyanobacterium Anabaena sp. Strain PCC 7120. J Bacteriol 190:7584–7590.

23. Llacer JL, Espinosa J, Castells MA, Contreras A, Forchhammer K, Rubio V. 2010. Structural basis for the regulation of NtcA-dependent transcription by proteins PipX and PII. Proc Natl Acad Sci 107:15397–15402.

24. Yadav P, Goyal VD, Gaur NK, Kumar A, Gokhale SM, Makde RD. 2018. Structure of Asp-bound peptidase E from *Salmonella enterica* : Active site at dimer interface illuminates Asp recognition. FEBS Lett 592:3346–3354.

25. Obst M, Oppermann-Sanio FB, Luftmann H, Steinbüchel A. 2002. Isolation of Cyanophycin-degrading Bacteria, Cloning and Characterization of an Extracellular Cyanophycinase Gene (*cphE*) from *Pseudomonas anguilliseptica* Strain BI: THE *cphE* Gene from P. anguillaseptica BI encodes a cynophycin-hydrolizing enzyme. J Biol Chem 277:25096–25105.

26. Conlin CA, Hakensson IK, Liljas A, MILLERl CG. 1994. Cloning and Nucleotide Sequence of the Cyclic AMP Receptor Protein-Regulated Salmonella typhimurium pepE Gene and Crystallization of Its Product, an ax-Aspartyl Dipeptidase 176:166-72.

27. Giner-Lamia J, Robles-Rengel R, Hernández-Prieto MA, Muro-Pastor MI, Florencio FJ, Futschik ME. 2017. Identification of the direct regulon of NtcA during early acclimation to nitrogen starvation in the cyanobacterium *Synechocystis* sp. PCC 6803. Nucleic Acids Res 45:11800–11820.

28. Sandmann G, Mautz J, Breitenbach J. 2016. Control of light-dependent keto carotenoid biosynthesis in *Nostoc* 7120 by the transcription factor NtcA. Z Für Naturforschung C 71:303–311.

29. Hamilton JJ, Garcia SL, Brown BS, Oyserman BO, Moya-Flores F, Bertilsson S, Malmstrom RR, Forest KT, McMahon KD. 2017. Metabolic Network Analysis and Metatranscriptomics Reveal Auxotrophies and Nutrient Sources of the Cosmopolitan Freshwater Microbial Lineage acI. MSystems 2:e00091–17.

30. Liang B, Wu T-D, Sun H-J, Vali H, Guerquin-Kern J-L, Wang C-H, Bosak T. 2014. Cyanophycin Mediates the Accumulation and Storage of Fixed Carbon in Non-Heterocystous Filamentous Cyanobacteria from Coniform Mats. PLoS ONE 9:e88142.

31. Ghylin TW, Garcia SL, Moya F, Oyserman BO, Schwientek P, Forest KT, Mutschler J, Dwulit-Smith J, Chan L-K, Martinez-Garcia M, others. 2014. Comparative single-cell genomics reveals potential ecological niches for the freshwater acI *Actinobacteria* lineage. ISME J 8:2503-16.

32. Markowitz VM, Chen I-MA, Palaniappan K, Chu K, Szeto E, Grechkin Y, Ratner A, Jacob B, Huang J, Williams P, Huntemann M, Anderson I, Mavromatis K, Ivanova NN, Kyrpides

NC. 2012. IMG: the integrated microbial genomes database and comparative analysis system. Nucleic Acids Res 40:D115–D122.

33. Garcia SL, McMahon KD, Martinez-Garcia M, Srivastava A, Sczyrba A, Stepanauskas R, Grossart H-P, Woyke T, Warnecke F. 2013. Metabolic potential of a single cell belonging to one of the most abundant lineages in freshwater bacterioplankton. ISME J 7:137-?.

34. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403–410.

35. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes11Edited by F. Cohen. J Mol Biol 305:567–580.

36. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Soding J, Thompson JD, Higgins DG. 2014. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7:539–539.

37. McWilliam H, Li W, Uludag M, Squizzato S, Park YM, Buso N, Cowley AP, Lopez R. 2013. Analysis Tool Web Services from the EMBL-EBI. Nucleic Acids Res 41:W597–W600.

38. Li W, Cowley A, Uludag M, Gur T, McWilliam H, Squizzato S, Park YM, Buso N, Lopez R. 2015. The EMBL-EBI bioinformatics web and programmatic tools framework. Nucleic Acids Res 43:W580–W584.

39. Wiefel L, Steinbüchel A. 2014. Solubility Behavior of Cyanophycin Depending on Lysine Content. Appl Environ Microbiol 80:1091–1096.

40. Otwinowski Z and Minor W. " Processing of X-ray Diffraction Data Collected in Oscillation Mode ", Methods in Enzymology, Volume 276: Macromolecular Crystallography, part A, p.307-326, 1997,C.W. Carter, Jr. & R. M. Sweet, Eds., Academic Press (New York).

41. Roy A, Kucukural A, Zhang Y. 2010. I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc 5:725–738.

42. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. 2015. The I-TASSER Suite: protein structure and function prediction. Nat Methods 12:7–8.

43. Zhang Y. 2008. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics 9:40.

44. Sybyl-X Molecular Modeling Software Packages, Version 2.0. TRIPOS Associates, Inc; St. Louis, MO, USA: 2012

45. Emsley P, Lohkamp B, Scott WG, Cowtan K. 2010. Features and Development of Coot. Coot, Acta Crystal D Biol Crystal. 66:286-501. Version 0.8.9.

46. Afonine PV et al. 2012. Towards automated crystallographic structure refinement with phenix.refine. Acta Crystal D Biol Crystal 68:352-67 (2012).

47. Headd JJ et al. 2012. Use of knowledge-based restraints in phenix.refine to improve macromolecular refinement at low resolution. Acta Crystal D Biol Crystal 68:381-39.

48. Afonine PV et al. 2013. Bulk-solvent and overall scaling revisited: faster calculations, improved results. Acta Crystal D Biol Crystal 69:625-34.

49. Afonine PV et al. 2009. Automatic multiple-zone rigid-body refinement with a large convergence radius. Appl Crystal 42:607-615.

50. Schrödinger, LLC. The PyMOL Molecular Graphics System, V2.0.0. Schrödinger, LLC, New York, NY.

Chapter V – Future Directions

Future Directions for Chapters II and III

My goal when starting this project was to determine if acI *Actinobacteria* contain the necessary physiology to synthesize a functional actinorhodopsin using native retinal production. In addition to showing actinorhodopsin activity, I demonstrated that many acI have the capacity to synthesize a complex carotenoid and possibly a heliorhodopsin. However, work remains for these systems: 1) determining if heliorhodopsin forms from helio-opsin and retinal, and what the holoprotein function is, 2) determining the exact identity of the final carotenoid or set of carotenoids produced from complex carotenoid-related enzymes, and if a ketocarotenoid, if it interacts with rhodopsins of acI, and 3) determining the carotenoid, retinal, and rhodopsin composition of environmental acI.

The immediate future work should focus on optimization of heliorhodopsin production and purification. Because heliorhodopsins are still relatively unknown, it is an excellent advantage to know that a homologous retinal-binding protein may be present in acI genomes. Heliorhodopsin work is feasible because I have already set up a small-scale pipeline for purification of an actinorhodopsin using intracellular retinal production and exogenous retinal addition. Yet, successful detergent solubilization is not assured because of variable protein behavior. Another limitation of the heliorhodopsin work is that the function, even of the characterized example, is unknown. Ultrafast spectroscopy would be required to get information about the photocycle and creative assay development to determine any activity may be required. This is perhaps a more far-term goal until more information is known about other heliorhodopsins. One possible approach would be to more closely analyze sensory rhodopsins that activate cytoplasmic transducers. A transducer homolog may still be undiscovered in genomes.

Secondary to heliorhodopsin production is heterologous production and/or environmental enrichment of carotenoids found in acI. The groundwork for heterologous production and identification has been laid using a retinal pathway as a feasibility test. Additionally, the identified

carotenoid-related genes and possible reaction pathways have been described. These two bodies of work allow for the rapid testing for gene function. While this is a large undertaking due to possible carotenoid complexity, rational combinations of genes according to the synthesis pathway should be tried. These combinations of modification genes can be paired with both native acI precursor producing genes or *Pantoea ananatis* genes that produce high levels of precursors. However, another look at stable production of monocylic carotenes should first be attempted. While production of carotenes using acI CrtYc and CrtYd was not demonstrated, *Myxococcus xanthus* cyclase genes have been shown to stably produce monocyclic carotenes. I suggest that these enzymes be substituted for acI proteins in future work.

Even if heterologous work encounters difficulty, it is feasible to enrich acI in the laboratory and/or collect acI from their native environment. Extraction and analysis of the sample with my methodologies should be tested. Care must be taken to truly have a highly enriched or preferably pure sample. Any contaminating organisms may produce carotenoids of their own and skew results. Purity was a major practical hurdle for any culture enrichment of acI, in addition to extremely low levels of biomass in the first place. Environmental acI work also suffered from the same limitations. While prefiltering lake water removes larger organisms, it can be difficult to then trap acI efficiently. On top of this, unknown amounts of free-floating membranes from larger organisms could be enriched with the acI sample. These would still contain carotenoids because of their hydrophobic nature.

A tangential line of inquiry for enriched and collected acI is direct visualization. Surprisingly, no work has been published in visualizing acI cells via electron microscopy and /or cryo-electron tomography. Perhaps rhodopsins fill acI membranes and would be clearly resolved with this method. However, the obvious limitation of this is ensuring that actual acI cells are being looked at and, again, not contaminating organisms.

Most of the above-mentioned problems in compound identification result from impure complex cultures of acI. A long-term goal of the project should be to focus on axenically culturing

acI. While attempts at this have been made, no rigorous combinatorial approach has been used. While this seems daunting to try and generate a lake water medium, there may be a subset of nutrients that enrich acI to higher levels than currently obtainable. Part of the problem is that lake water is not readily analyzable because of its extreme complexity and seasonal compositional shifts. On top of this, sterile lake water is not currently obtainable because heat sterilization methods destroy sensitive compounds and leads to salt precipitation. Also, filter sterilization is not 100% effective, even at submicron cutoffs.

Future Directions for Chapter IV

Because membranes and membrane proteins were and still are difficult to manipulate without specialized equipment, I also wanted to determine if acI *Actinobacteria* have soluble proteins of interest in the context of unique physiological adaptations. I determined the structure and function of an acI PepE aspartate dipeptidase enzyme that could be regulated by a possible NtcA homolog. However, work remains for this system: 1) characterizing other possible PepE ligands, like degraded cyanophycin and 2) determining if the NtcA homolog is a DNA binding transcription factor at a specific recognition sequence and what the NtcA homolog has regulatory influence over (i.e. other nitrogen related genes and/or carotenoid/rhodopsin-related genes). This part of the acI project seems the most immediately accessible to researchers unfamiliar with the acI project.

In terms of PepE ligands, simple aspartate-leading dipeptides of canonical amino acids can be tested for cleavage. Although this would require substantial upfront financial resources, methods for peptide derivatization and detection are readily available. Perhaps this is not the most cutting-edge work, but it would further help show the range of substrates for acI PepE. More interestingly, work can be started on determining if small cyanophycin dimers or trimers can be broken down by the enzyme. While cyanophycin is a long bulky polymer, the smaller peptides are more like PepE simple dipeptides. It is not infeasible for these small DR repeats to be in the lake

and imported by acI cells. However, a natural source of these peptides is not likely feasible. As a result, I suggest chemical synthesis with unblocked N- and C- termini. This work would still require more effort in terms of peptide identification by HPLC and/or MS, but should be straight forward in execution like aspartate-p-nitroaniline cleavage assays.

If bioinformatics is preferred to benchtop work, NtcA consensus sequence and target identification is certainly feasible. With the explosion of acI genomes, including complete versions, a model network for NtcA regulation could be determined. I suggest using commonly studied cyanobacteria as a guide for possible targets of global nitrogen regulation in acI. While the consensus sequence may not be identical to other NtcA systems, flexibility in search sequences and restriction by upstream distances from targeted genes should allow progress. If this pipeline is successful, I suggest moving the project into more experimental work. This would include purification and crystallization of the transcription factor. Additional experiments may include: assays of consensus sequences found in acI to determine relative binding affinity and cocrystallization of these DNA sequences with the transcription factor to determine if structural changes are at play. Overall, I see the NtcA consensus and protein work as the overall most feasible next step in better defining large scale acI *Actinobacterial* physiology.

Even more long-term work on acI *Actinobacteria* physiology should focus on defining the large secondary metabolic systems. While certain aspects of the acI project are interesting (i.e. antenna carotenoids and PepE cyanophycin breakdown), detailing the larger systems of metabolites and their regulation would be more beneficial to the field. Given that my results are the first investigations acI experimental physiology, I hope that future work will help push other systems to the forefront. An ideal, personally-interesting system is sulfur metabolism, which can be studied at the bioinformatic, functional, and structural level. I hope that the interesting locations of cysteines in acI actino-opsins and PepE lead to larger conclusions in this system.