Title
Towards an In Vitro Reconstitution of the Alpha-Carboxysome

Author
Chaijarasphong, Thawatchai

Publication Date
2016

Peer reviewed|Thesis/dissertation

Towards an In Vitro Reconstitution of the α-Carboxysome

By

Thawatchai Chaijarasphong

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Chemistry

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor David F. Savage, Chair
Professor John E. Dueber
Professor Jamie H.D. Cate

Fall 2016

Abstract

Towards an In Vitro Reconstitution of the α-Carboxysome

by

Thawatchai Chaijarasphong

Doctor of Philosophy in Chemistry

University of California, Berkeley

Professor David F. Savage, Chair

Many bacteria employ a protein organelle, the carboxysome, to concentrate carbon dioxide and catalyze the initial fixation reaction. Only 10 genes from *Halothiobacillus neapolitanus* are sufficient for heterologous expression of carboxysomes in *Escherichia coli*, opening the door to mechanistic analysis of the assembly process of this 200 MDa+ complex. One of these genes, *csoS2*, produces two highly repetitive intrinsically-disordered protein isoforms and has been shown to be indispensable in carboxysome assembly. Detailed functional characterization of *csoS2*, however, has been hindered by the lack of understanding of how this single gene yields expression of two gene products. In this work, we set out to develop a deeper understanding of CsoS2's biogenesis and its function in α-carboxysome assembly. Using tandem mass spectrometry and biochemical assays, we have revealed that -1 programmed ribosomal frameshifting (- 1 PRF) is responsible for the generation of a truncated protein with C-terminus translated from the -1 frame, CsoS2A, in addition to the full-length protein, CsoS2B. We show for the first time that CsoS2B can be independently produced by mutations of -1 PRF elements and only CsoS2B is necessary for the assembly of *H. neapolitanus* carboxysomes in native and heterologous hosts. With the knowledge of the identity of CsoS2 isoforms, we next investigate the ability of individual CsoS2 domains to participate in protein-protein interaction with RuBisCO, the primary enzymatic component of the carboxysome. Here, we demonstrate that the 259-aa N-terminal domain of CsoS2 multivalently binds RuBisCO, potentially via its short amphiphatic helices. Finally, based on our findings, we propose a hypothetical model that describes the formation and maturation of the α-carboxysome. This work illustrates, for the first time, the simultaneous involvement of cotranslational regulation and an intrinsically-disordered protein in the assembly of a prokaryotic organelle.

Acknowledgements

The completion of this dissertation is a very important moment for me. It marks the end of my PhD, and also the end of my ten-year long journey as a Thai student in the United States. It was a long, unpredictable, and adventurous process, but I was able to achieve my goal, thanks to many people who helped me along the way.

I would like to thank Professor David Savage for his supervision and support throughout my PhD. His optimism and passion for science have inspired me and made me realize the true joy of doing science. At the same time, his advices always remind me to never lose sight of the life outside of science, to also value my personal happiness in addition to the intellectual pursuit. I am honored to have been his student, and I will follow in his footsteps when I have my own students in the future.

I would like to thank the members of the Savage Lab for their company, lively intellectual discussions, and valuable inputs for my research. In particular, thanks to my bay-mates, Robert Nichols and Luke Oltrogge, who have collaborated with me on much of this project and provided me with helpful insights. Also, many thanks to Rachel Hood for sharing with me her interests in arts, books, coffee, foods, cute animals and microbes, and many more.

I am indebted to my family who have tirelessly supported me in every way they can, even though they live so far away. I would like to thank my parents, Montchai Chaijarasphong and Phanthipha Suwancharoensuk, for their support. They always did everything they could, sometimes pushing their financial limit, to ensure that I received the best education possible. I owe successes in my life to their sacrifices. Many thanks to my sister, Maiyaporn Chaijarasphong, who always entertained me with her humorous and sassy updates on life in Thailand. I also would like to thank my aunt, Napaporn Suwancharoensuk, for her wisdom that helped me through the more difficult times during my PhD.

Lastly, I believe this Acknowledgements section would not be complete without the mention of those who supported me through my first few years in the US. It was a critical moment in my life, being the first time I had to live far away from home, in a place where practically no one spoke my language. Without their support, it would have been incredibly difficult to get through those first few years. With this, I would like to thank: Mrs. Easterling, Mr. Hedberg, and Dr. Geller of The Lawrenceville School who kindly and patiently gave me wonderful advices and assistance while I was still navigating the unfamiliar territory of American culture and education system; the Thai Scholars who kept me company and reminded me that I was not alone; and my friends at Stanford who made my undergraduate years an amazing time of my life, especially Charlene Bong, Max Murialdo, Brian Huh, Kolby Hanson, Andrew Gines, Dan Graham, Mike Ding, and the Stanford Thai community.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

-1 PRF	-1 Programmed Ribosomal Frameshifting
1,6-HD	1,6-Hexanediol
3PG	3-Phosphoglycerate
A	Adenine nucleotide/Alanine amino acid
A_{280}	Absorbance at 280 nm
aa	amino acid
AmpR	Ampicillin Resistance Gene
Arg	Arginine
BLI	Bio-layer Interferometry
BMC	Bacterial Microcompartment
BME	Beta-mercaptoethanol
bp	base pair
BSA	Bovine Serum Albumin
C	Cytosine nucleotide/Cysteine amino acid
CA	Carbonic Anhydrase
CB	Carboxysome
CCM	Carbon Dioxide Concentrating Mechanism
CD	Circular Dichroism
cfu	colony-forming units
CmR	Chloramphenical Resistance Gene
CO_2	Carbon Dioxide
cso	carboxysome operon
CTD	C-terminal Domain
Cys	Cysteine amino acid
D	Aspartic acid
DMSO	Dimethylsulfoxide
DNA	Deoxy-Ribonucleic Acid
DTT	Dithiothreitol
E	Glutamic acid
Ec	*Escherichia coli*
EGFP	Enhanced Green Fluorescent Protein
EUT	Ethanolamine Utilization
FPLC	Fast Protein Liquid Chromatography
FRAP	Fluorescence Recovery After Bleaching
G	Guanine nucleotide/Glycine amino acid
GFP	Green Fluorescent Protein
GG	Golden Gate (cloning)
H	Histidine
HCR	High CO_2 Requiring
His	Histidine
Hn	*Halothiobacillus neapolitanus*
HRP	Horse-radish Peroxidase
IDP	Intrinsically Disordered Protein

IPTG ..... Isopropyl beta-D-1-Thiogalactopyranoside
K ..... Lysine
K ..... Lysine
KanR ..... Kanamycin Resistance Gene
kDa ..... kilo-Dalton
L ..... Leucine
LB ..... Lysogeny Borth
Leu ..... Leucine
LZ ..... Leucine Zipper
m/z ..... mass-to-charge ratio
MALDI-TOF ..... Matrix Assisted Laser Desorption/Ionization - Time of Flight
MoRF ..... Molecular Recognition Feature
MR ..... Middle Region
mRNA ..... Messenger Ribonucleic Acid
MS/MS ..... Tandem Mass-spectrometry
MST ..... MicroScale Thermophoresis
MW ..... Molecular Weight
NA ..... Numerical Aperture
NHS ..... N-Hydroxylsuccinimide
NTA ..... Nitrilotriacetic Acid
NTD ..... N-terminal Domain
OD_{600} ..... Optical Density at 600 nm
ORF ..... Open Reading Frame
P ..... Proline
PAGE ..... Polyacrylamide Gel Electrophoresis
PCR ..... Polymerase Chain Reaction
PDU ..... Propanediol Utilization
PMSF ..... Phenylmethylsulfonyl fluoride
R ..... Arginine
RFP ..... Red Fluorescent Protein
RPM ..... Round Per Minute
RuBisCO ..... Ribulose-1,5-Bisphosphate Carboxylase/Oxygenase
RuBP Ribulose-1,5-Bisphosphate
S ..... Serine
SacB ..... *Bacillus subtilis* levansucrase
SDS ..... Sodium Dodecyl Sulfate
SEC ..... Size-Exclusion Chromatography
sfGFP ..... Super Folder Green Fluorescent Protein
T ..... Threonine
TCEP ..... Tris(2-carboxyethyl)phosphine
TEM ..... Transmission Electron Microscopy
TEV ..... Tobacco Etch Virus protease
Tris ..... Tris-(hydroxymethyl)-aminomethane base
tRNA ..... Transfer Ribonucleic Acid

V..Valine
WT..Wild-Type

Chapter 1 Introduction

1.1 Biological Compartments as Catalytic Nanoreactors

With the global population predicted to increase by over 35% by the end of 2050 (Agapakis et al., 2012; Axen et al., 2014; Long et al., 2015), achieving sufficient production of fuels and valuable chemicals to meet the ever increasing global demand has been a collective challenge for humanity in recent years. Rapid dwindling of natural reserves in industrial production of valuable compounds has created an urgent need for alternative manufacturing strategies based on renewable carbon sources (Hoffert et al., 2002; Ragauskas et al., 2006; Yeates et al., 2013). Production in microbial hosts have received increasing attention because of their ability to use enzymatic pathways to produce variety of compounds from cheap carbon sources such as sugar, starch, and lignocellulose (Bonacci et al., 2012; Hassan and Kalam, 2013; Held et al., 2016; Keasling, 2010). While significant cost is involved in development of these microbial factories, it is justified in the long run by high product values, low cost of starting materials, and minimal pollution relative to the use of chemical catalysts (Cameron et al., 2013; Chen et al., 2013; Rude and Schirmer, 2009; Sauer et al., 2008).

Although traditional methods for metabolic pathway optimization, such as adjusting relative amounts of enzymes to optimize metabolic fluxes, are generally effective in microbial production (Ajikumar et al., 2010; Fan and Bobik, 2011), they fall short toxic and volatile intermediates are involved. For intermediates such as CO_2, aldehydes, and reactive oxygen species, tightly-controlled substrate channeling and physical barrier against intermediate leakage become essential. While multi-enzyme complexes are often employed in nature to ensure precise substrate channeling, *de novo* design of such as system is difficult in practice and would require very short distance between two active sites (Bauler et al., 2010; Lawrence et al., 2014). In this regard, biological compartments offer an alternative solution through their dense packing of enzymes, which have been shown to improve substrate channeling (Castellana et al., 2014; Quin et al., 2016), and the shell that restricts the escape of substrates and intermediates (Dou et al., 2008; Heinhorst et al., 2006; Shively et al., 1973; Yeates et al., 2010) (Figure 1-1). Consequently, by localizing enzymes of interest in these molecular chassis, it is possible to improve the kinetics of the pathway and protect the host from toxic intermediates (Agapakis et al., 2012; Sampson and Bobik, 2008)

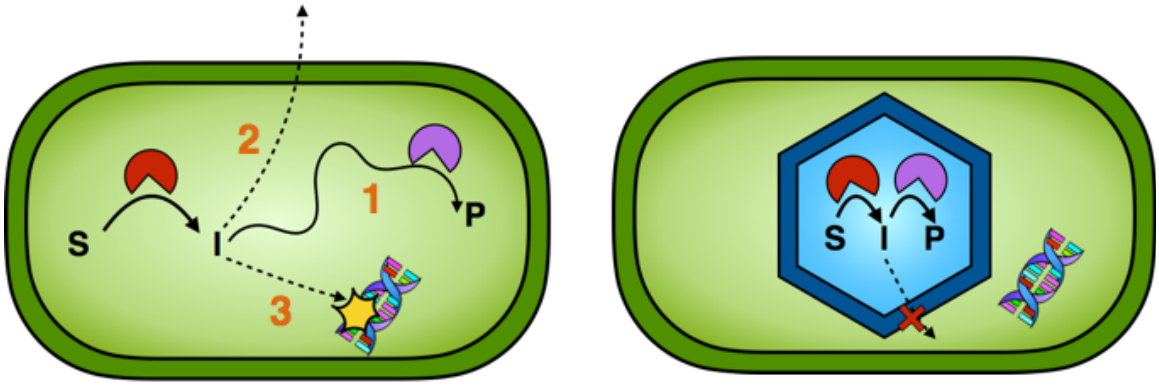

Figure 1-1 Encapsulation of a metabolic pathway by a biological compartment.
A two-step pathway involving two enzymes and one intermediate is used as an example. When the substrate (S), intermediate (I), and enzymes (red and magenta) are located in the cytosol (left), the intermediate from the first step of the pathway is faced with 3 possibilities: (1) the intermediate may need to diffuse for a considerable distance before it reaches the second enzyme and gets processed into the product (P); (2) a gaseous or membrane-permeable intermediate may diffuse out of the cell; (3) a toxic intermediate may encounter and damage an essential macromolecule like DNA. By sequestering the pathway inside a biological compartment (right, blue hexagon), proper substrate channeling can be achieved, and the escape of toxic or volatile intermediates is minimized.

Formerly, it was believed that only eukaryotic organisms could possess intracellular compartments in the form of lipid-bound organelles, but the discovery of a protein-bound compartment in cyanobacteria 50 years ago challenged this notion (Cheng et al., 2008; Ragauskas et al., 2006; Yeates et al., 2013). Since then, various protein-based biological compartments, including very small nanocompartments (less than 20 nm) and medium-sized bacterial microcompartments (BMCs, 80-500 nm), have continuously been brought into the spotlight (Agapakis et al., 2012; Held et al., 2016; Keasling, 2010). Among these, the majority of engineering applications were developed from nanocompartments such as viral-like particles (Cameron et al., 2013; Long et al., 2015; Rude and Schirmer, 2009; Strable and Finn, 2009), encapsulins (Giessen, 2016; Hoffert et al., 2002; Ragauskas et al., 2006), and apo-ferritin (Bauler et al., 2010; Keasling, 2010; Lawrence et al., 2014; Maity et al., 2015). Their robustness to protein modifications, *in vitro* assembly, and chemical functionalization have allowed them to be used in molecular imaging(Castellana et al., 2014; Hassan and Kalam, 2013; Hooker et al., 2007; Quin et al., 2016), drug delivery (Dou et al., 2008; Heinhorst et al., 2006; Ma et al., 2012; Rude and Schirmer, 2009; Sauer et al., 2008; Shively et al., 1973; Yeates et al., 2010), and caged catalysis (Ajikumar et al., 2010; Cannon et al., 2001; Liu et al., 2012). However, due to their small sizes, they are ill-suited for hosting enzymatic pathways, especially when multiple large enzymes are involved.

1.2 Bacterial Microcompartments

Due to their larger sizes, bacterial microcompartments offer advantages that would complement the limitations of the small nanocompartments. These protein-based compartments perform crucial metabolic functions in bacteria such as alcohol metabolism and carbon dioxide fixation, analogous to organelles in eukaryotic organisms (Bauler et al., 2010; Spreitzer and Salvucci, 2002; Yeates et al., 2010; 2008). However, unlike eukaryotic organelles, BMCs are enveloped by a protein shell instead of lipid membrane, and thus they are not organelles by the conventional definition. A typical BMC is a polyhedral body with the diameter of 80-500 nm, filled with enzymes that belong to a short metabolic pathway (Castellana et al., 2014; Kaplan and Reinhold, 1999; Yeates et al., 2010). A recent bioinformatic study identified 23 types of BMCs in 23 phyla of bacteria (Axen et al., 2014; Dou et al., 2008; Shih et al., 2014). This prevalence of BMCs suggests that the metabolic advantages they provide may help enhance the fitness of the organisms.

Based on the types of reactions they catalyze, BMCs can be categorized into 2 main groups: anabolic and catabolic microcompartments (Axen et al., 2014; Sampson and Bobik, 2008). The only known member of the anabolic group is the carboxysome, which performs carbon dioxide fixation in photoautrotophic and chemotrophic bacteria (Cheng et al., 2008; Espie and Kimber, 2011). Catabolic BMCs, on the other hand, are significantly more functionally diverse and can be further subdivided into at least 7 functional clusters (Agapakis et al., 2012; Axen et al., 2014). Overall, they catalyze the oxidation of small molecules to generate compounds that can provide energy for the cell. However, despite the variety, only a small fraction of catabolic BMCs have been extensively characterized, most notably the propanediol-utilizing (Pdu) microcompartment and ethanolamine-utilizing (Eut) microcompartment (Yeates et al., 2013).

Bacterial microcompartments have recently gained increasing attention from the community of synthetic biologists and metabolic engineering scientists due to their modularity and ability to self-assemble. It has been shown that formation of functional BMCs in heterologous hosts requires no other proteins beyond their building blocks (Bonacci et al., 2012; Held et al., 2016), and their assembly is robust towards changes in composition. These properties allows heterologous cargos to be encapsulated in BMCs simply by fusion to a native cargo (Cameron et al., 2013; Chen et al., 2013) or, in certain cases, by inclusion of a short peptide tag (Fan and Bobik, 2011). For instance, recent studies have demonstrated that it is possible to localize non-native metabolic pathways into Pdu (Lawrence et al., 2014) and Eut (Quin et al., 2016). However, despite these successes, our ability to engineer BMCs has relied heavily on trials and errors due to lack of knowledge about their properties, including but not limited to: mechanism of self-assembly, substrate permeability, and internal chemical environment. In addition, a strategy to reconstitute a BMC from purified protomers *in vitro* is desirable, as it will enable the control of the amount of individual components without the laborious process of fine-tuning their

expression levels *in vivo*. This will not be possible in the absence of the knowledge of how a cell assemble individual pieces in the cytosolic milieu. Therefore, alongside the efforts to develop new engineering applications, it is equally important to deepen our fundamental knowledge of factors that determine structure and function of BMCs in their biological context.

1.3 Carboxysome as a Model for Understanding BMC assembly

In this study, we aim to develop our understanding of general principles that govern self-assembly of BMCs by using the α-carboxysome from *H. neapolitanus* as a model system. Found exclusively in photoautotrophic and chemolithotrophic bacteria, a carboxysome is an icosahedral protein complex of 80–500 nm composed of thousands of protomers (Heinhorst et al., 2006; Shively et al., 1973; Yeates et al., 2010). Two defining features of carboxysomes are (i) a protein shell composed of thousands of hexameric and 60 pentameric capsomers and (ii) a lumen possessing numerous copies of the cargo enzymes Ribulose 1,5-Bisphosphate Carboxylase/Oxygenase (RuBisCO) and Carbonic Anhydrase (CA) (Cannon et al., 2001). Overall, the carboxysome catalyzes a segment of the Calvin-Benson-Bessham cycle where ribulose-1,5-bisphosphate (RuBP) forms an unstable adduct with CO_2, followed by its scission into two molecules of 3-phosphoglycerate (3PG) which then enters central metabolism (Spreitzer and Salvucci, 2002) (Figure 1-2A). The intricate structural arrangement of the carboxysome emerged as a uniquely prokaryotic version of Carbon Dioxide Concentrating Mechanism (CCM) to overcome the fundamental problem of inefficient CO_2 fixation (Kaplan and Reinhold, 1999). Specifically, RuBisCO has low catalytic rate of approximately 3-10 CO_2 per second and low affinity towards CO_2 in addition to its off-target activity towards O_2. The adduct formation between RuBP and O_2 results in only one molecule of 3PG and one molecule of 2-phosphoglycolate, the latter of which requires an energetic process known as photorespiration to be recovered as 3PG (Shih et al., 2014). In a carboxysome, carbonic anhydrase converts bicarbonate ion from the cytosol to CO_2 on the lumenal side (Baker et al., 2000; Hoffert et al., 2002; Ragauskas et al., 2006). Since the shell of a carboxysome acts as a permeability barrier to CO_2 (Dou et al., 2008; Hassan and Kalam, 2013; Keasling, 2010), the level of CO_2 eventually reaches a saturating level, and thus enables RuBisCO to catalyze with high rate and specificity. This dramatic improvement of CO_2 has sparked efforts to increase crop yield by expressing carboxysomes in C3 plants, which do not possess CCM (Long et al., 2015; Rude and Schirmer, 2009; Sauer et al., 2008).

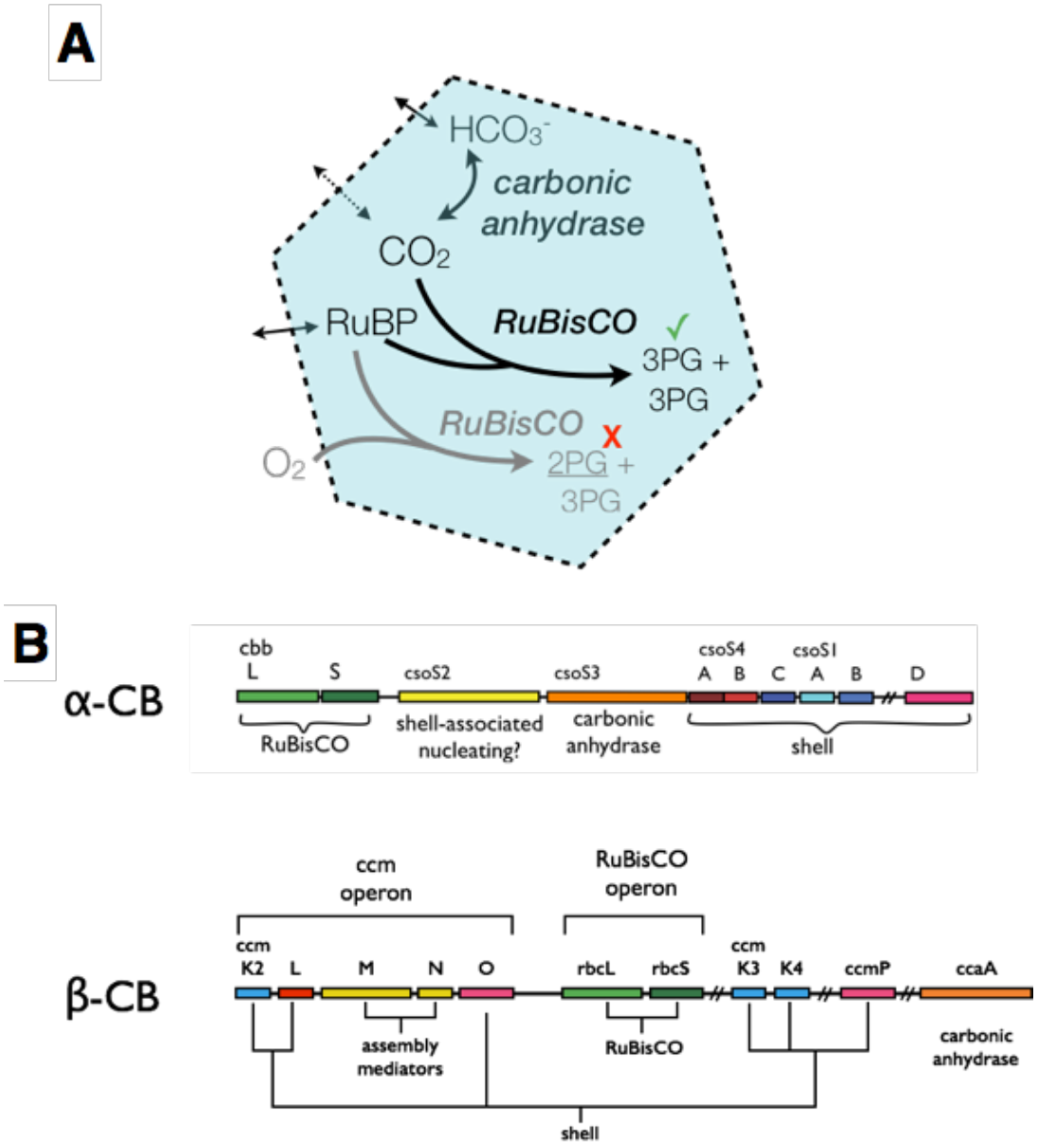

Figure 1-2: Carboxysome as a CO_2-fixing machinery.
(A) Outline of the carbon-concentrating mechanism carried out within a carboxysome. Bicarbonate anion (HCO_3^-) first crosses the shell into the lumen, where it is dehydrated by carbonic anhydrase to CO_2, which cannot leave the carboxysome. RuBisCO then catalyzes the formation of two molecules of 3-phosphoglycerate from 1,5-ribulose-bisphosphate (RuBP) and CO_2. In the presence of O_2, RuBisCO can additionally oxygenate RuBP into one molecule of 2-phosphoglycolate (2PG) and 3PG. Since 2PG cannot be used by the cell, it needs to be converted to 3PG in the cytosol via an energetic process. (B) Distribution of carboxysomal genes in α- and β- carboxysomes from *H. neapolitanus* and *Synechococcus elongatus* PCC7942, respectively. Double-slash between two ORFs indicates that they do not belong to the same operon. Proteins expressed by the genes shown in the α-carboxysome's diagram are described in details in the text. Products of the genes shown in the β-carboxysome's diagram are as follows: hexameric shell protein with single BMC domain (CcmK2, CcmK3, CcmK4), hexameric shell protein with

tandem BMC domains (CcmO, CcmP), pentameric shell protein (CcmL), RuBisCO large and small subunits (RbcL, RbcS), carbonic anhydrase (CcaA), assembly-mediating proteins (CcmM, CcmN).

Although all carboxysomes perform similar CO_2 fixation reaction, two distinct evolutionary lineages of the carboxysome, α and β forms, exist (Rae et al., 2013; Whitehead et al., 2014). The α-carboxysomes are found in α-cyanobacteria in *Prochlorococcus* and *Synechococcus* species, and chemoautotrophs belonging to the phyla Actinobacteria, Nitropirae and Proteobacteria. This family of carboxysomes is characterized by the presence of Form IA RuBisCO as well as the clustering of carboxysome-associate genes into a cso operon (Figure 1-2B). This segregated genetic organization, presumably derived from horizontal gene transfer (Ajikumar et al., 2010; Badger et al., 2002), allows functional α-carboxysome to be successfully produced in *E. coli* simply by expressing the cso operon and a nearby gene from *H. neapolitanus* (Bauler et al., 2010; Bonacci et al., 2012). The β-carboxysomes belong exclusively to cyanobacteria and contain the plant-like Form IB RuBisCO. The genes encoding components of β-carboxysomes are spread across multiple genetic loci in the genome (Figure 1-2B), and it is unclear how their expression is coordinated. Unlike the α counterparts, β-carboxysomes have never been successfully purified from the native organism or expressed in a heterologous host. Therefore, α-carboxysomes, which can potentially be expressed in more genetically-tractable and fast-growing industrial microbes, hold promise for engineering applications. Unfortunately, while the assembly mechanism of β-carboxysomes have been well-characterized, very little is known about α-carboxysomes biogenesis apart from a number of indirect evidence that suggests that they follow a distinct assembly paradigm from the β-lineage (Cameron et al., 2013; Castellana et al., 2014; Iancu et al., 2010; Menon et al., 2008; Rae et al., 2013). The knowledge about the assembly of α-carboxysomes, therefore, will help broaden our understanding of different strategy which nature employs to generate bacterial microcompartments.

1.4 The α-Carboxysome from *H. neapolitanus*

We chose the α-carboxysome from *Halothiobacillus neapolitanus* as the model BMC for our study due to its simple genomic organization, good expression in *E. coli*, and ease of preparation of monodispersed particles. The first BMC to be visualized and biochemically studied (Dou et al., 2008; Shively et al., 1973), a purified *H. neapolitanus* carboxysome contains at least ten different protomers, nine of which are expressed from the cso operon (Figure 1-3B). The icosahedral shell of the carboxysome is made of the products of six genes: *csoS1ABCD* and *csoS4AB*. The most abundant components of the shell and the carboxysome are four paralogous BMC domain-containing proteins (PF00936) called CsoS1A, CsoS1B, CsoS1C, and CsoS1D. These proteins exist as hexamers (Sampson and Bobik, 2008; Tsai et al., 2007) and together they account for roughly 800 copies in a carboxysome (Cheng et al., 2008; Heinhorst et al., 2006). CsoS1A and CsoS1C are highly homologous to each other, with 98% identity, while CsoS1B has extra 12 amino acids on the C-terminus. Considering their near-perfect homology, it is unclear why different paralogs are

present; in fact, certain organisms possess only 1 type of a CsoS1 paralog. In contrast, CsoS1D possesses several properties that set it apart from the other three paralogs. While *csoS1ABC* are a part of the cso operon, *csoS1D* is located in a different gene cluster downstream of the operon and encodes a fusion protein of two BMC domains that assemble into a trimer (pseudohexamer). The crystal structure of CsoS1D reveals that its functional unit may consist of two pseudohexameric rings stacking on top of each other. Depending on the relative angles of the rings, CsoS1D can assume either an open or close configuration, suggesting its involvement in a gating mechanism (Agapakis et al., 2012; Klein et al., 2009)., The vertices of the carboxysome are occupied by pentameric shell proteins (PF03319), CsoS4A and CsoS4B. Although only 60 copies of CsoS4AB are present per carboxysome, they are critical for the morphology and function of the carboxysome, as a *H. neapolitanus* knockout mutant that did not produce CsoS4AB abnormally elongated carboxysomes and acquired a high CO_2 requiring phenotype (Cai et al., 2009; Long et al., 2015; Strable and Finn, 2009).

The lumen of a carboxysome is densely packed with its main enzymatic components, RuBisCO and carbonic anhydrase, encoded by *cbbLS* and *csoS3*, respectively. A molecule of RuBisCO holoenzyme consists of eight large subunits (CbbL) and eight small subunits (CbbS) that assemble into a 560-kDa hexadecamer. There are approximately 250 copies of RuBisCO hexadecamer in the interior of the carboxysome, together serving as highly concentrated active sites for CO_2 fixation (Figure 1-3A). Carbonic anhydrase (CsoS3 or CA) exists as 114-kDa dimers that total roughly 50 copies per carboxysome. CA dimers were found to be strongly associated to the shell and possibly are localized underneath the shell (Figure 1-3A) (Baker et al., 2000; Dou et al., 2008; Giessen, 2016; Hoffert et al., 2002; Ragauskas et al., 2006). Together, dense colocalization of RuBisCO and CA may allow for tight substrate channeling and consequently highly effective CO_2-fixation in the manner previously described.

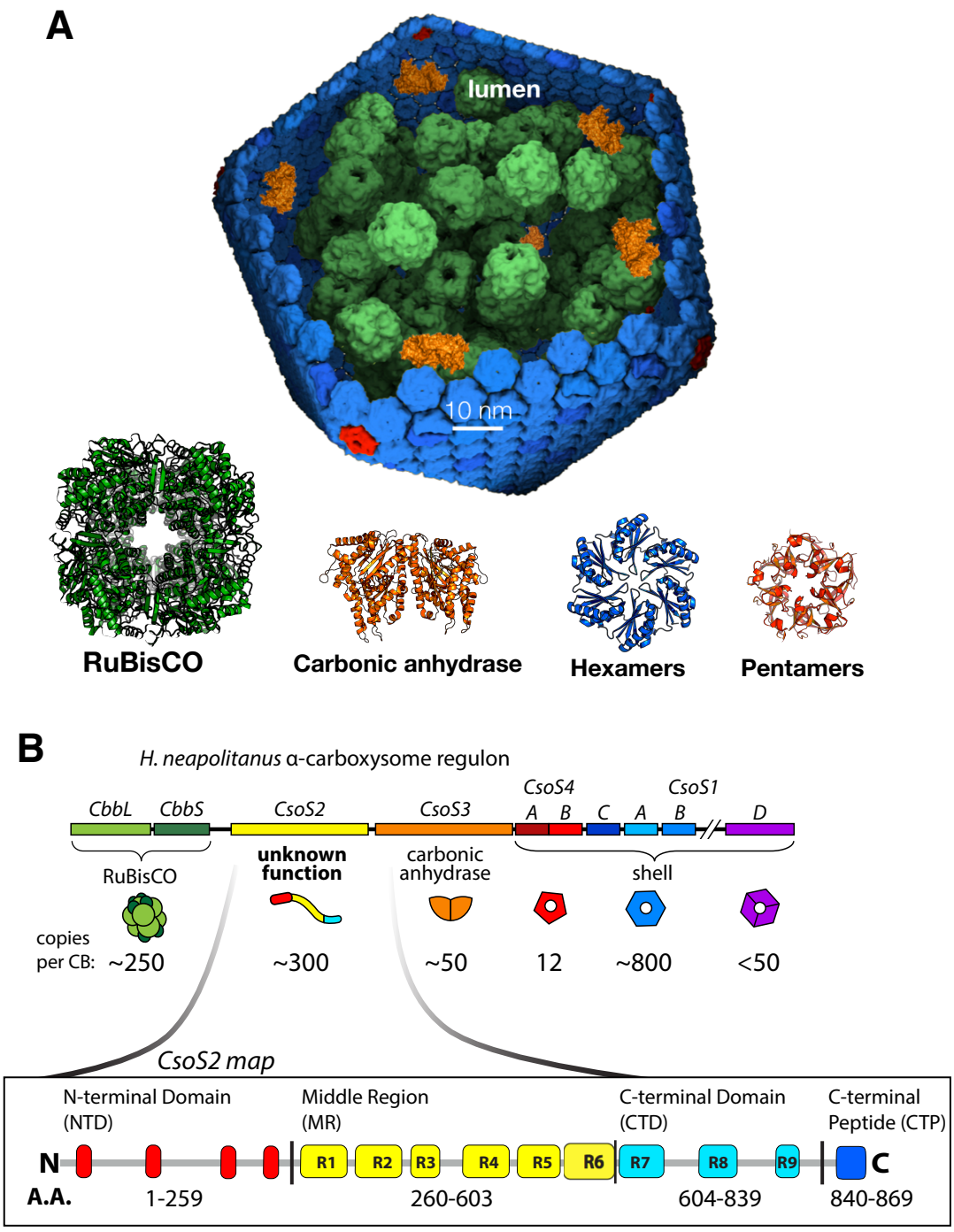

Figure 1-3: Protein subunits in a *H. neapolitanus* α-carboxysome
(A) Partial molecular surface representation of *H. neapolitanus* carboxysome. RuBisCO (green), hexameric (blue) and pentameric (red) shell protomers, and carbonic anhydrase (orange) are shown (This figure is reproduced with permission from David F. Savage). 3D structures of individual protomers from Protein Data Bank (PDB) are displayed below. (B) Copy numbers of protomers in a *H. neapolitanus* carboxysome. A cartoon diagram showing three domains of CsoS2 and individual repeats is shown at the bottom.

1.5 CsoS2: a Critical Factor in α-Carboxysome Assembly

In addition to the shell and enzymatic components described above, the α-carboxysome is also home to an enigmatic but crucial protein called CsoS2. It was first discovered as an unknown band on the SDS-PAGE analysis of purified *H. neapolitanus* carboxysomes (Cannon and Shively, 1983; Keasling, 2010; Maity et al., 2015), before Baker *et al.* confirmed that it was encoded by *csoS2* gene in the cso operon (Baker et al., 1999; Hassan and Kalam, 2013; Hooker et al., 2007). According to the primary sequence, *H. neapolitanus* CsoS2 is a 869-aa protein with multiple amino acid repeats and can be subdivided into 3 domains based on sequence similarity (Figure 1-3B). CsoS2 was presumed to play essential role in organizing the assembly of the carboxysome for several reasons. First, inspection of primary sequence revealed that CsoS2 had exceptionally high pKa compared to other carboxysomal components (Heinhorst et al., 2006; Ma et al., 2012; Rude and Schirmer, 2009; Sauer et al., 2008), and thus it was likely to be positively charged while the other proteins were negatively charged at the physiological pH. In addition, the densitometric analysis performed on SDS-PAGE of purified carboxysomes showed that CsoS2 and RuBisCO were present in almost equimolar amounts (Ajikumar et al., 2010; Heinhorst et al., 2006; Liu et al., 2012), suggesting that these proteins may interact in a 1:1 stoichiometry. Collectively, these findings hinted at a strong possibility that CsoS2 may participate in electrostatic interaction with carboxysomal proteins and associate with RuBisCO and carbonic anhydrase. The significance of CsoS2 in α-carboxysome biogenesis was confirmed by a recent finding that CsoS2-knockout mutant of *H. neapolitanus* was unable to produce carboxysomes (Bauler et al., 2010; Cai et al., 2015; Yeates et al., 2008; 2010).

Although we hypothesized CsoS2 could be essential for directing α-carboxysome assembly, biochemical characterization of CsoS2 has been surprisingly scarce. From the time of the original characterization to the beginning of this work, no other CsoS2 characterization was published. The dearth of CsoS2 characterization was understandably due to the two most confounding properties of CsoS2. First, a single ORF of *csoS2* was found to encode two polypeptides, CsoS2A and CsoS2B, in *H. neapolitanus* and when overexpressed in *E. coli* (Baker et al., 1999; Castellana et al., 2014; Yeates et al., 2010). Since the mechanism that generates these isoforms is not known, there has not been a way to express each isoform separately in order to tease apart their individual functions *in vivo*. Secondly, CsoS2 belongs to a large class of proteins called intrinsically-disordered proteins (IDPs) which generally display exceptional ability in organizing large protein assemblies, but are also plagued with undesirable properties (Axen et al., 2014; Dou et al., 2008; Uversky and Dunker, 2010). True to form, CsoS2 is highly sensitive to *in vivo* proteolysis, prone to self-aggregation, and intractable to many quantitative biochemical techniques that were designed with globular proteins in mind (Axen et al., 2014; Jerabek-Willemsen et al., 2014; Sampson and Bobik, 2008; Velazquez-Campoy et al., 2004). Therefore, dealing

with CsoS2 requires exploration of biochemical strategies that could tame its recalcitrant behaviors.

1.6 Objective of This Study

This dissertation is organized into two main parts, with the overarching goal of shedding light on the role of the CsoS2 in the assembly of the α-carboxysomes. In Chapter 2, we devoted our effort in tackling the long-standing conundrum of CsoS2 isoforms using combination of techniques including point mutagenesis, mass-spectrometry, and bioinformatics. Once the mechanism was unraveled, we examined the necessity of each isoform by evaluating the morphology of carboxysomes when only one isoform was present. In Chapter 3, we dived deeper into the issue of protein-protein interaction between CsoS2 and RuBisCO using the combination of biochemical assays and imaging techniques. In our effort to procure well-behaved CsoS2 and its derivatives for *in vitro* assays, we developed strategies that, to a certain extent, helped circumvent the inherent difficulty of working with an IDP. While the assembly of a massive particle like the carboxysome would likely involve a complex interaction network far beyond the scope of this dissertation, the insights acquired from this work should serve as an important first step towards building the complete picture of the carboxysome assembly and help guide future engineering and *in vitro* reconstitution efforts.

1.7 References

Agapakis, C.M., Boyle, P.M., and Silver, P.A. (2012). Natural strategies for the spatial optimization of metabolism in synthetic biology. Nat Chem Biol *8*, 527–535.

Ajikumar, P.K., Xiao, W.-H., Tyo, K.E.J., Wang, Y., Simeon, F., Leonard, E., Mucha, O., Phon, T.H., Pfeifer, B., and Stephanopoulos, G. (2010). Isoprenoid pathway optimization for Taxol precursor overproduction in *Escherichia coli*. Science *330*, 70–74.

Axen, S.D., Erbilgin, O., and Kerfeld, C.A. (2014). A taxonomy of bacterial microcompartment loci constructed by a novel scoring method. PLoS Comp Biol *10*, e1003898.

Badger, M.R., Hanson, D., and Price, G.D. (2002). Evolution and diversity of CO_2 concentrating mechanisms in cyanobacteria. Functional Plant Biology.

Baker, S.H., Williams, D.S., Aldrich, H.C., Gambrell, A.C., and Shively, J.M. (2000). Identification and localization of the carboxysome peptide CsoS3 and its corresponding gene in *Thiobacillus neapolitanus.* Arch. Microbiol. *173*, 278–283.

Baker, S.H., Lorbach, S.C., Rodriguez-Buey, M., Williams, D.S., Aldrich, H.C., and Shively, J.M. (1999). The correlation of the gene *csoS2* of the carboxysome operon

with two polypeptides of the carboxysome in *Thiobacillus neapolitanus*. Arch. Microbiol. *172*, 233–239.

Bauler, P., Huber, G., Leyh, T., and McCammon, J.A. (2010). Channeling by proximity: the catalytic advantages of active site colocalization using Brownian dynamics. J. Phys. Chem. Lett. *1*, 1332–1335.

Bonacci, W., Teng, P.K., Afonso, B., Niederholtmeyer, H., Grob, P., Silver, P.A., and Savage, D.F. (2012). Modularity of a carbon-fixing protein organelle. Proc. Natl. Acad. Sci. U.S.A *109*, 478–483.

Cai, F., Dou, Z., Bernstein, S., Leverenz, R., Williams, E., Heinhorst, S., Shively, J., Cannon, G., and Kerfeld, C. (2015). Advances in understanding carboxysome assembly in *Prochlorococcus* and *Synechococcus* implicate CsoS2 as a critical component. Life *5*, 1141–1171.

Cai, F., Menon, B.B., Cannon, G.C., Curry, K.J., Shively, J.M., and Heinhorst, S. (2009). The pentameric vertex proteins are necessary for the icosahedral carboxysome shell to function as a CO_2 leakage barrier. PLoS ONE *4*, e7521.

Cameron, J.C., Wilson, S.C., Bernstein, S.L., and Kerfeld, C.A. (2013). Biogenesis of a bacterial organelle: the carboxysome assembly pathway. Cell *155*, 1131–1140.

Cannon, G.C., and Shively, J.M. (1983). Characterization of a homogenous preparation of carboxysomes from *Thiobacillus neapolitanus*. Arch. Microbiol. *134*, 52–59.

Cannon, G.C., Bradburne, C.E., Aldrich, H.C., Baker, S.H., Heinhorst, S., and Shively, J.M. (2001). Microcompartments in prokaryotes: carboxysomes and related polyhedra. Appl. Environ. Microbiol. *67*, 5351–5361.

Castellana, M., Wilson, M.Z., Xu, Y., Joshi, P., Cristea, I.M., Rabinowitz, J.D., Gitai, Z., and Wingreen, N.S. (2014). Enzyme clustering accelerates processing of intermediates through metabolic channeling. Nat Biotechnol *32*, 1011–1018.

Chen, A.H., Robinson-Mosher, A., Savage, D.F., Silver, P.A., and Polka, J.K. (2013). The bacterial carbon-fixing organelle is formed by shell envelopment of preassembled cargo. PLoS ONE *8*, e76127.

Cheng, S., Liu, Y., Crowley, C.S., Yeates, T.O., and Bobik, T.A. (2008). Bacterial microcompartments: their properties and paradoxes. Bioessays *30*, 1084–1095.

Dou, Z., Heinhorst, S., Williams, E.B., Murin, C.D., Shively, J.M., and Cannon, G.C. (2008). CO_2 fixation kinetics of *Halothiobacillus neapolitanus* mutant carboxysomes lacking carbonic anhydrase suggest the shell acts as a diffusional barrier for CO_2. J. Biol. Chem. *283*, 10377–10384.

Espie, G.S., and Kimber, M.S. (2011). Carboxysomes: cyanobacterial RubisCO comes in small packages. Photosyn. Res. *109*, 7–20.

Fan, C., and Bobik, T.A. (2011). The N-terminal region of the medium subunit (PduD) packages adenosylcobalamin-dependent diol dehydratase (PduCDE) into the Pdu microcompartment. J. Bacteriol. *193*, 5623–5628.

Giessen, T.W. (2016). Encapsulins: microbial nanocompartments with applications in biomedicine, nanobiotechnology and materials science. Curr. Opin. Chem. Biol. *34*, 1–10.

Hassan, M.H., and Kalam, M.A. (2013). An overview of biofuel as a renewable energy source: development and challenges. Procedia Engineering *56*, 39–53.

Heinhorst, S., Cannon, G.C., and Shively, J.M. (2006). Carboxysomes and carboxysome-like inclusions. Complex Intracellular Structures in Prokaryotes 142–165.

Held, M., Kolb, A., Perdue, S., Hsu, S., Bloch, S.E., Quin, M.B., and Schmidt-Dannert, C. (2016). Engineering formation of multiple recombinant Eut protein nanocompartments in *E. coli*. Sci Rep *6*, 24359.

Hoffert, M.I., Caldeira, K., Benford, G., Criswell, D.R., Green, C., Herzog, H., Jain, A.K., Kheshgi, H.S., Lackner, K.S., Lewis, J.S., et al. (2002). Advanced technology paths to global climate stability: energy for a greenhouse planet. Science *298*, 981–987.

Hooker, J.M., Datta, A., Botta, M., Raymond, K.N., and Francis, M.B. (2007). Magnetic resonance contrast agents from viral capsid shells: a comparison of exterior and interior cargo strategies. Nano Lett. *7*, 2207–2210.

Iancu, C.V., Morris, D.M., Dou, Z., Heinhorst, S., Cannon, G.C., and Jensen, G.J. (2010). Organization, structure, and assembly of alpha-carboxysomes determined by electron cryotomography of intact cells. J. Mol. Biol. *396*, 105–117.

Jerabek-Willemsen, M., André, T., Wanner, R., Roth, H.M., Duhr, S., Baaske, P., and Breitsprecher, D. (2014). MicroScale Thermophoresis: Interaction analysis and beyond. J. Mol. Struct *1077*, 101–113.

Kaplan, A., and Reinhold, L. (1999). CO_2-concentrating mechanisms in photosynthetic microorganisms. Annu. Rev. Plant Biol. *50*, 539–570.

Keasling, J.D. (2010). Manufacturing molecules through metabolic engineering. Science *330*, 1355–1358.

Klein, M.G., Zwart, P., Bagby, S.C., Cai, F., Chisholm, S.W., Heinhorst, S., Cannon, G.C., and Kerfeld, C.A. (2009). Identification and structural analysis of a novel carboxysome shell protein with implications for metabolite transport. J. Mol. Biol.

392, 319–333.

Lawrence, A.D., Frank, S., Newnham, S., Lee, M.J., Brown, I.R., Xue, W., Rowe, M.L., Mulvihill, D.P., Prentice, M.B., Howard, M.J., et al. (2014). Solution structure of a bacterial microcompartment targeting peptide and its application in the construction of an ethanol bioreactor. ACS Synth. Biol.

Liu, X., Wei, W., Yuan, Q., Zhang, X., Li, N., Du, Y., Ma, G., Yan, C., and Ma, D. (2012). Apoferritin-CeO_2 nano-truffle that has excellent artificial redox enzyme activity. Chem. Commun. *48*, 3155–3157.

Long, S.P., Marshall-Colon, A., and Zhu, X. (2015). Meeting the global food demand of the future by engineering crop photosynthesis and yield potential. Cell *161*, 56–66.

Ma, Y., Nolte, R.J.M., and Cornelissen, J.J.L.M. (2012). Virus-based nanocarriers for drug delivery. Adv. Drug Deliv. Rev. *64*, 811–825.

Maity, B., Fujita, K., and Ueno, T. (2015). Use of the confined spaces of apo-ferritin and virus capsids as nanoreactors for catalytic reactions. Curr. Opin. Chem. Biol. 88–97.

Menon, B.B., Dou, Z., Heinhorst, S., Shively, J.M., and Cannon, G.C. (2008). *Halothiobacillus neapolitanus* carboxysomes sequester heterologous and chimeric RuBisCO Species. PLoS ONE *3*, e3570.

Quin, M.B., Perdue, S.A., Hsu, S., and Schmidt-Dannert, C. (2016). Encapsulation of multiple cargo proteins within recombinant Eut nanocompartments. Appl Microbiol Biotechnol.

Rae, B.D., Long, B.M., Badger, M.R., and Price, G.D. (2013). Functions, compositions, and evolution of the two types of carboxysomes: polyhedral microcompartments that facilitate CO_2 fixation in cyanobacteria and some proteobacteria. Microbiol. Mol. Biol. Rev. *77*, 357–379.

Ragauskas, A.J., Williams, C.K., Davison, B.H., Britovsek, G., Cairney, J., Eckert, C.A., Frederick, W.J., Hallett, J.P., Leak, D.J., Liotta, C.L., et al. (2006). The path forward for biofuels and biomaterials. Science *311*, 484–489.

Rude, M.A., and Schirmer, A. (2009). New microbial fuels: a biotech perspective. Curr. Opin. Microbiol. *12*, 274–281.

Sampson, E.M., and Bobik, T.A. (2008). Microcompartments for B_{12}-dependent 1,2-propanediol degradation provide protection from DNA and cellular damage by a reactive metabolic intermediate. J. Bacteriol. *190*, 2966–2971.

Sauer, M., Porro, D., Mattanovich, D., and Branduardi, P. (2008). Microbial

production of organic acids: expanding the markets. Trends Biotechnol. *26*, 100–108.

Shih, P.M., Zarzycki, J., Niyogi, K.K., and Kerfeld, C.A. (2014). Introduction of a synthetic CO_2-fixing photorespiratory bypass into a cyanobacterium. J. Biol. Chem.

Shively, J.M., Ball, F., Brown, D.H., and Saunders, R.E. (1973). Functional organelles in prokaryotes: Polyhedral inclusions (carboxysomes) of *Thiobacillus neapolitanus*. Science *182*, 584–586.

Spreitzer, R.J., and Salvucci, M.E. (2002). Rubisco: structure, regulatory interactions, and possibilities for a better enzyme. Annu. Rev. Plant Biol. *53*, 449–475.

Strable, E., and Finn, M.G. (2009). Chemical modification of viruses and virus-like particles. In Viruses and Nanotechnology, (Berlin, Heidelberg: Springer), pp. 1–21.

Tsai, Y., Sawaya, M.R., Cannon, G.C., Cai, F., Williams, E.B., Heinhorst, S., Kerfeld, C.A., and Yeates, T.O. (2007). Structural analysis of CsoS1A and the protein shell of the *Halothiobacillus neapolitanus* carboxysome. PLoS Biol. *5*, e144.

Uversky, V.N., and Dunker, A.K. (2010). Understanding protein non-folding. Biochim. Biophys. Acta *1804*, 1231–1264.

Velazquez-Campoy, A., Leavitt, S.A., and Freire, E. (2004). Characterization of protein-protein interactions by isothermal titration calorimetry. Methods Mol. Biol. *261*, 35–54.

Whitehead, L., Long, B.M., Price, G.D., and Badger, M.R. (2014). Comparing the *in vivo* function of α-carboxysomes and β-carboxysomes in two model cyanobacteria. Plant Physiol. *165*, 398–411.

Yeates, T.O., Crowley, C.S., and Tanaka, S. (2010). Bacterial microcompartment organelles: Protein shell structure and evolution. Annu. Rev. Biophys. *39*, 185–205.

Yeates, T.O., Jorda, J., and Bobik, T.A. (2013). The shells of BMC-type microcompartment organelles in bacteria. J Mol Microbiol Biotechnol *23*, 290–299.

Yeates, T.O., Kerfeld, C.A., Heinhorst, S., Cannon, G.C., and Shively, J.M. (2008). Protein-based organelles in bacteria: carboxysomes and related microcompartments. Nat Rev Micro *6*, 681–691.

Chapter 2 Programmed Ribosomal Frameshifting Mediates Expression of the α-Carboxysome

Adapted from *Journal of Molecular Biology*, Volume 428, Chaijarasphong, T., Nichols, R. J., Kortright, K. E., Nixon, C. F., Teng, P. K., Oltrogge, L. M., Savage, D. F., Programmed Ribosomal Frameshifting Mediates Expression of the α-Carboxysome, Pages 153-164, Copyright 2015, with permission from Elsevier.

2.1 Introduction

Much of the uncertainty in CsoS2 function lies in an unexpected observation that carboxysomes from *H. neapolitanus* possess two different forms of CsoS2 (Baker et al., 1999). The predicted molecular mass of the *csos2* gene product is 92 kDa but purified carboxysomes possess a long isoform, CsoS2B, and a short isoform, CsoS2A, with apparent molecular masses of ~ 120 and ~ 70 kDa, respectively, as analyzed by denaturing polyacrylamide gel electrophoresis (PAGE) analysis. Previous work has shown that CsoS2A and CsoS2B share the same N-terminus but may be post-translationally modified via glycosylation(Baker et al., 1999), thus offering an explanation for the discrepancy between predicted and observed MWs. However, other evidence suggests that the proteins may not be modified (Espie and Kimber, 2011; Gonzales et al., 2005), and our attempt to reproduce the glycosylation assay employed by the authors also did not yield a positive result (Figure 2-2A). Therefore, the true origin of the CsoS2 isoforms remained an unsolved puzzle.

Elucidating a clear functional role for CsoS2 requires understanding its specific molecular interactions with other carboxysome components and, therefore, the precise definition of primary amino acid sequence structure for both CsoS2B and CsoS2A. Our investigation of the possible mechanisms that would explain the isoform formation led us to a cotranslational process known as programmed ribosomal frameshifting (PRF). PRF is a process whereby an mRNA encodes a specific signal to shift the ribosome from the usual translated 0 frame to either the +1 or the −1 frame (Caliskan et al., 2015). This can result in the production of two or more different polypeptides from a single mRNA. While pervasive in viruses, PRF is much less common in prokaryotes and eukaryotes. Only a handful of proteins in *E. coli*, most notably the gamma subunit of DNA polymerase III, are known to be generated by PRF (Tsuchihashi and Kornberg, 1990).

The model system for −1 PRF is the gene *dnaX* from *E. coli*, which generates the gene products tau and gamma of the DNA polymerase III complex (Flower and McHenry, 1990; Tsuchihashi and Kornberg, 1990). A large body of biochemical and biophysical data on this system have defined the signal for frameshifting: (i) a 5′ ribosomal binding site upstream of the frameshifting site that stalls the ribosome, (ii) a slippery sequence of A-AAA-AAG where the ribosome changes frame, and (iii) a significant secondary structure downstream of the slippery sequence to further impede ribosome translation (Caliskan et al., 2015). These signals cause the

ribosome slipping from a 0 frame codon back to the −1 frame codon. This process is enabled by the degeneracy of the genetic code and allowance of mismatches at the wobble codon position. As a consequence, Kornberg *et al*. have shown that a sequence matching the nucleotide motif X-XXY-YYZ has the potential to encode PRF (Tsuchihashi and Kornberg, 1990).

Here, using a heterologous *E. coli* expression system, we show that differential expression of the two isoforms is encoded by ribosomal frameshifting *cis* elements present on the *csoS2* mRNA that are capable of encoding roughly 1:1 stoichiometric expression of CsoS2A and CsoS2B. CsoS2B is 869 amino acids long and is produced from full-length expression of the 0 frame of the gene. Frameshifting specifically generates the CsoS2A protein possessing a C-terminus from the −1 frame and results in a truncated protein that is 570 amino acids in length. Surprisingly, in carboxysome formation assays, CsoS2B by itself leads to the formation of assemblies structurally similar to carboxysomes, while CsoS2A alone is incapable of forming carboxysomes. Thus, we have defined the sequence of each CsoS2 isoform, a mechanism of their formation, and we provide initial evidence toward their functional importance.

2.2 Materials and Methods

2.2.1 Preparation of Plasmids

Plasmids were constructed using Golden-Gate assembly and Phusion polymerase (New England Biolabs). Primers were designed to contain BsaI restriction sites and Golden-Gate recombination sites appropriate for the assembly of the desired construct. In those constructs with an affinity tag, the sequence expressing the tag was also included in a primer. Expression vectors were modified in our laboratory to be compatible with Golden-Gate cloning (Table 2-1).

The open reading frame for *H. neapolitanus* CsoS2 was amplified from the pHnCB10 plasmid containing nine genes of the *cso* operon and *csoS1D*. For the construct containing *csoS2* without stop codons in alternative reading frames, gBlocks containing synonymous mutations that remove stop codons in alternative frames, starting at the 1278th nucleotide of *csoS2*, were ordered from Integrated DNA Technologies. The gBlocks were amplified by PCR to generate BsaI restriction sites and then incorporated into the pET14-based destination vector using Golden-Gate assembly.

For the construct expressing only CsoS2B, a set of primers was designed to either (1) change the slippery sequence from 5′-C-CCA-AAG-3′ to 5′-A-CCT-AAG-3′ or (2) synonymously mutate several nucleotides in the downstream hairpin to reduce hairpin formation. The sequence of the mutated hairpin region is shown in Figure 2-4B. To make the frameshift reporter construct, we employed PCR to introduce the BsmBI sites to the EGFP gene, the slippery sequence and stem–loop region in *csoS2*,

and the mCherry gene. A TEV protease cleavage site was also introduced between the stem–loop and mCherry. These amplified products were assembled simultaneously into the pET14-based destination vector via Golden-Gate assembly. The control construct, with the slippery sequence and stem-loop replaced by 5′-GCGGCGGGCGGTGCAGGAGCT-3′ (encoding Ala-Ala-Gly-Gly-Ala-Gly-Ala linker), was prepared in a similar manner.

2.2.2 Heterologous Expression and Purification of His-tagged H. neapolitanus *CsoS2*
Chemically competent *E. coli* BL21(AI) transformed with pBz13 were grown in 1-l LB media. At OD_{600} of 0.3–0.5, the culture was induced with 0.1% arabinose and grown at 22 °C, 175 RPM for 16 h. Cells were subsequently harvested by centrifugation at 4000g for 20 min. The pellets were resuspended in buffer A [20 mM Tris and 300 mM NaCl (pH 7.5)] with cOmplete protease inhibitor (Roche), 0.1 mg/mL lysozyme (Sigma-Aldrich), and 25 U/μL benzonase (EMD Millipore). Cell lysis was performed by passing the cell suspension through an Avestin EmulsiFlex-C3 homogenizer three times. Crude lysate was centrifuged at 15,000g for 20 min to remove cell debris. Clarified lysate was incubated with Ni-NTA resin for 1 h and transferred to a gravity-flow column. The resin was washed with buffer A plus 30 mM imidazole and 1 mM phenylmethylsulfonyl fluoride. The protein was eluted with buffer A plus 300 mM imidazole and 1 mM PMSF. Eluted protein was passed through 10DG desalting column (Bio-Rad) to exchange into buffer B [50 mM Tris and 150 mM NaCl (pH 7.4)] plus cOmplete ethylenediaminetetraacetic acid (EDTA)-free protease inhibitor cocktail (Roche) and stored at -80 °C until further use.

2.2.3 Expression and Visualization of EGFP-mCherry Frameshift Reporter
BL21(AI) transformed with pRJN-FS or pRJN-control were grown in LB media (10 mL). At OD_{600} of 0.3–0.5, the cultures were induced with 0.1% arabinose and further grown at 22 °C, 175 RPM for 16 h. Cells were harvested by centrifugation at 4000g for 20 min and subsequently chemically lysed with SoluLyse-Tris (Genlantis). The lysate was clarified by centrifuging at 15,000g for 20 min. The clarified lysate was mixed with SDS-PAGE sample loading buffer. Without heating, we loaded gel samples on a 4–20% TGX Criterion gel (Bio-Rad). SDS-PAGE was performed at 150 V with Laemmli electrophoresis running buffer (25 mM Tris, 192 mM glycine, and 1% SDS). The gel was subsequently imaged with ChemiDoc imager (Bio-Rad), using GFP and RFP filters to detect EGFP and mCherry signals, respectively.

2.2.4 Western Blotting
Clarified lysates of BL21(AI) expressing the desired protein were mixed with SDS-PAGE sample loading buffer and heated at 95 °C for 10 min. Gel samples were electrophoresed on a 4–20% TGX Criterion gel (Bio-Rad) at 200 V with Laemmli electrophoresis running buffer and blotted to polyvinylidene fluoride membrane using a Trans-Blot Turbo Transfer System (Bio-Rad). Proteins were subsequently probed with His-probe Antibody (H-3) HRP (Santa Cruz Biotechnology) and detected through chemiluminescence using SuperSignal West Pico Chemi-luminescent Substrate (Thermo Scientific).

2.2.5 Purification of Carboxysomes from E. coli *and* H. neapolitanus
Tuner-competent cells (Novagen) were transformed with a plasmid (pHnCB10LC) expressing carboxysomes. Cells were grown in 1 L of Terrific Broth at 37 °C to an OD_{600} of 0.3–0.5 at which point isopropyl-β-D-1-thiogalactopyranoside (IPTG) was added to a final concentration of 1 mM. The culture was induced for 16 h at 20 °C and harvested by centrifugation at 5000g for 20 min.

The cell pellet was resuspended in 50 mL buffer TEMB [10 mM Tris, 10 mM $MgCl_2$, 1 mM EDTA, and 20 mM $NaHCO_3$ (pH 8.4)] with 1 mM PMSF, 0.1 mg/mL lysozyme (Sigma-Aldrich), and 25 U/μL benzonase (EMD Millipore). Cell lysis was performed by passing the cell suspension through an Avestin EmulsiFlex-C3 homogenizer three times. The crude cell extract was centrifuged at 12,000g for 30 min to remove cell debris. The supernatant was subsequently centrifuged at 40,000g for 30 min to pellet the carboxysomes. The pellet was then resuspended in 20 mL of 33% cellLytic B (Sigma-Aldrich) in TEMB buffer in order to remove lipid contaminants and centrifuged again at 40,000g for 30 min. The pellet was resuspended in 3 mL TEMB and centrifuged at 3000g. The supernatant was then applied to 25 mL sucrose gradient made from 10%, 20%, 30%, 40%, and 50% sucrose. The gradient was centrifuged at 105,000g for 30 min. Fractions of 1 mL were collected and analyzed for the presence of carboxysomes via SDS-PAGE gel. The fractions that contained carboxysomes were pooled and centrifuged at 105,000g for 90 min to pellet the carboxysomes. Carboxysomes were then resuspended in 1 mL TEMB and stored at 4 °C.

To prepare carboxysomes from *H. neapolitanus*, the media for culturing *H. neapolitanus* (DSMZ-68 media) was first prepared by mixing the following reagents in 900 mL distilled water: 4 g KH_2PO_4, 4 g K_2HPO_4, 0.8 g $MgSO_4.7H_2O$, 0.4 g NH_4Cl, 5 mL trace element solution (see below), 0.5 mL saturated bromocresol purple, 15 g bacto agar (for solid media). Trace element solution was prepared following the recipe reported by Santer and Vishniac (Santer and Vishniac, 1957). After autoclaving the mixture, 100 mL of sterile 10% $Na_2S_2O_3$ solution was added. To make DSMZ-68 solid media, 25 mL of the autoclaved agar-media mixture was poured to a petri dish and allowed to dry overnight.

To start the growth of *H. neapolitanus*, cells from the frozen stock were streaked on solid medium and allowed to grow at 30 °C in air for 3-5 days. A single colony was then picked and cultured in 10 mL DSMZ-68 media. Once the stationary phase was reached, the culture was diluted in 100 mL media and allowed to grow to stationary phase again. The culture was back-diluted once more in 1 L media and upon the stationary phase was transferred to 10 L DSMZ-68 media in a bioreactor. Cells were allowed to grow in air for 2 more days, with air bubbled into the bioreactor, before they were centrifuged at 6000g for 20 min. Lysis, sucrose gradient fractionation, and SDS-PAGE analysis of the purified carboxysomes were performed as described for *E. coli*

2.2.6 MALDI-TOF Analysis

The matrix was prepared by dissolving 15 mg of sinapic acid (Sigma-Aldrich) in 1 mL of 70:30 water:acetonitrile with 1% trifluoroacetic acid (TFA). Purified EcS2 was mixed with the matrix in 1:10 v/v ratio. A volume of 2 µl of protein matrix mixture was spotted on a well in 384-well plate (DE1580TA; Axima) and desiccated. Another 2 µl of protein matrix was spotted on the same well and desiccated once more. MALDI-TOF was performed in Axima Performance (Shimadzu Technology) with 180 U of power, accumulating 400 profiles.

2.2.7 Tandem Mass Spectrometry

Two forms of CsoS2 in purified EcS2 were separated via SDS-PAGE (4–20% TGX gel; Bio-Rad) and stained with Coomassie stain. Bands containing CsoS2A and CsoS2B were excised and submitted to the University of California Davis Proteomics Core Facility. Analysis was performed on a Q ExactiveTM Hybrid Quadrupole Orbitrap. Acquired spectra were searched in Scaffold against the WT CsoS2 sequence and the frameshifted sequence library was generated as described in the bioinformatic methods section below. The list of peptides was used to deduce the location of translational frameshifting.

2.2.8 Preparation of Samples for Negative-staining EM

A total of 5 µL of a protein sample were added to a freshly glow discharged, carbon-coated copper grid (EMS) and allowed to incubate for 2 min. The grids were washed with deionized water three times, before incubation with 5 µL of 1% uranyl acetate for 1 min. Excess liquid was gently wicked off and the grid was allowed to air-dry for 2 min. Negative-stain EM was performed with a Tecnai 12 TEM (FEI) operated at 120 kV. Size measurements of the visualized particles were performed on FIJI image processing software and size distribution was plotted using matplotlib package in Jupyter Notebook.

2.2.9 Bioinformatic Methods

To construct a library of frameshifted sequence, a Python script was written by Avi F. Flamholz to construct the list of tryptic peptides that would result from digestion of frameshifted Csos2. In brief, we generated all possible fusions of the primary Csos2 reading frame with the − 1 and − 2 reading frames, including only cases in which the fusion event was between 30% and 80% of the way through the Csos2 coding sequence. We then simulated a tryptic digest of these ~900 fusion proteins, generating a comprehensive list of possible fusion peptides. This database was then used to match m/z spectra from gel bands corresponding to CsoS2A and CsoS2B. The piling diagram was constructed from peptide counts derived from comparing mass spectra of CsoS2A and CsoS2B gel bands to the primary coding sequence of Csos2 (i.e., disregarding the potential for frameshifting). We counted the number of times each amino acid was observed in a peptide matched against the primary coding sequence. These counts are plotted in Figure 2-3 in the Results.

The secondary structure of mRNA was computed using the RNAstructure Web server. Nucleotide sequence starting at 50 bases 5′ of the slippery sequence and

ending at 50 bases 3′ of the slippery sequence was used as the input. The temperature parameter was set to 303.15 K. To calculate the stability of the stem–loop alone, we entered the sequence corresponding to the predicted stem–loop without additional nucleotides.

To construct a phylogenetic tree mapped with −1 PRF prediction (Fig. S5), we curated species bearing *csoS2* genes from the Joint Genome Institute Integrated Microbial Genomes database. FASTA sequences for the CsoS2-encoding mRNAs were input into the KnotInFrame −1 PRF prediction tool. Significant −1 PRF candidates as predicted by the KnotInFrame tool were further filtered by setting a threshold such that the candidate slip sequence must be found in the last two-thirds of the gene and that the structured mRNA element immediately downstream of the "XXXYYYZ" slip sequence must have minimal free energy value of 0.025 kcal/mol more stable than the nested mRNA structure. Species with candidate −1 PRF sequences meeting these parameters were marked in green on the phylogeny of *csoS2* nucleotide sequences.

2.2.10 Construction of csoS2::sacB-kan^r *and* csoS2::csoS2B H. neapolitanus *strains*
A gene cassette containing *sacB* and kanamycin-resistance marker was amplified by PCR from pAK31 plasmid obtained from Komelli group. The homology regions, ~1000 bp upstream and downstream of *csoS2* ORF in *cso* operon, were also amplified from pHnCB10 plasmid. The upstream homology region, *sacB-kan^r* cassette, and downstream homology region were assembled into a destination plasmid via Golden-Gate assembly and propagated in *E. coli*, resulting in the plasmid product called pBz89. Linear DNA containing *sacB-kan^r* and flanking homology sequences and then amplified pBz89 and cleaned with gel purification.

Wild-type *H. neapolitanus* from our laboratory frozen stock was streaked on a DSMZ -68 agar plate and allowed to grow in air for 3 days at 30 °C . A colony was picked and cultured in 10 mL liquid DSMZ-68 medium at 30 °C with 300 RPM shaking for 2-3 days or until the media turned yellow. From this starter culture, 250 μL was diluted in 50 mL DSMZ-68 in a 250 mL baffled flask and allowed to grow for ~12 h, until the OD at 600 nm reached 0.15-0.25.

To start a transformation, 10 mL of cells were centrifuged at 6000g for 10 min and resuspended in 1 mL fresh DSMZ-68 media. The linear DNA product amplified from pBz89 earlier was added to the cells to the final concentration of 250-400 ng/mL and incubated with 300 RPM shaking at 30 ^{o}C in 5 % CO_2 overnight. On the next day, 200 μL of the cells was plated on a DSMZ-68 agar plate with 10 μg/mL kanamycin and allowed to grow at 30 ^{o}C in 5% CO_2 for 3-5 days. Colony PCR was then performed to verify the presence of the desired *sacB-kan^r* insertion using two sets of primers specific to 5'- and 3'- regions of the gene cassette. The correct colonies were grown in 10 mL DSMZ-68 at 30 ^{o}C in 5% CO_2 until the stationary phase was reached. A 1-mL aliquot of this culture was then mixed with glycerol to the final concentration of 50% and stored in -80 ^{o}C until needed.

To construct a *csoS2wt::csoS2B* strain, a plasmid harboring *csoS2* with two point mutations in the slippery sequence, flanked by the upstream and downstream ~1000 bp homology regions, was constructed in a similar manner as described for pBz89. The resulting plasmid, called pBz90, was then used as the template to amplify the linear DNA. The *csoSwt2::sacB-kan*^{r} strain acquired previously was transformed with the linear DNA and subsequently plated on DSMZ 68 agar plates containing 0, 5, or 10% sucrose.

As the positive selection by sucrose was not successful, an alternative selection strategy was carried out. Linear DNA amplified from pBz90 was transformed again to *csoS2::sacB-kan*^{r} cells and spread on a DSMZ-68 agar plate with neither sucrose nor antibiotics. The plate was incubated in air at 30 °C for 1 week. A colony was picked and verified by colony PCR. The product from colony PCR was subjected to gel purification and submitted for Sanger sequencing at UC Berkeley Sequencing Facility. After the desired point mutations were verified, a frozen glycerol stock of the strain was prepared as described previously.

Table 2-1. Plasmids used in Chapter 2

Plasmid	Relevant genotype	Resistance	Reference
pNS3	IPTG-inducible expression plasmid derived from pUC57	Cm	(1)
pNS3-noRBS	Derivative of pNS3 with RBS removed from the parent plasmid	Cm	(1)
pHnCB10	pNS3-noRBS-based plasmid with 7.7 kb carboxysome operon from *H.neapolitanus* (Hneap_0922-Hneap_0914), including 33 bp upstream of CbbL containing RBS. csoS1D (Hneap_903) and its native RBS was cloned into the region after the carboxysome operon	Cm	(1)
pHnCB10-LC	pHnCB10 with resistance changed to Amp and origin changed to ColE1	Amp	This work
pBz51	pHnCB10-LC with *csoS2* gene in carboxysome operon replaced with *csoS2A* subcloned from pBz31. This construct expresses carboxysomal proteins without CsoS2B	Amp	This work
pBz52	pHnCB10-LC with *csoS2* gene in carboxysome operon replaced with *csoS2B* subcloned from pBz32. This construct expresses carboxysomal proteins without CsoS2A	Amp	This work
pBz31	pET14-based plasmid carrying *csoS2* gene from *H. neapolitanus*, with 5' removed and 2 last codons altered to express CsoS2A. The protein has hexahistidine tag on the N-terminus and Strep-II tag on the C-terminus	Amp	This work
pBz32	pET14-based plasmid carrying *csoS2* gene from *H. neapolitanus*, with ribosomal slip sequence mutated from 5'-C-CCA-AAG-3' to 5'-A-CCT-AAG-3', to express CsoS2B alone. The protein has a hexahistidine tag on the N-terminus and Strep-II tag on the C-terminus	Amp	This work

Plasmid	Relevant genotype	Resistance	Reference
pBz33	pET14-based plasmid carrying *csoS2* gene from *H. neapolitanus*, with hairpin downstream of the ribosomal slip sequence mutated such that the secondary structure is attenuated	Amp	This work
pBz13	pET14-based plasmid expressing CsoS2 from *H. neapolitanus* with N-terminal hexahistidine tag	Amp	This work
pBz16	pET14-based plasmid expressing CsoS2 from *H. neapolitanus* with C-terminal Strep-tag II	Amp	This work
pBz19	pET14-based plasmid expressing CsoS2 from *H. neapolitanus* with all stop codons in alternative reading frames after the 400th codon mutated. The mutations do not alter amino acids in the original reading frame	Amp	This work
pRJN-FS	pET14-based plasmid expressing slippery sequence and stem-loop from *H. neapolitanus csoS2* gene, fused with EGFP on the N-terminus an mCherry on the C-terminus. TEV cleavage site is included on the N-terminus of mCherry	Amp	This work
pRJN-control	pET14-based plasmid analogous to pRJN-FS, but with the slippery sequence and stem-loop replaced with nucleotides coding for Ala-Ala-Gly-Gly-Ala-Gly-Ala	Amp	This work
pBz89	Plasmid containing *sacB-kan*^{r} cassette flanked by ~1000 bp homology regions taken from the regions immediately upstream and downstream of *csoS2* gene in the cso operon	Kan	This work
pBz90	Plasmid containing *csoS2* gene with NoSlip mutations, flanked by ~1000 bp homology regions identical to those in pBz89	Kan	This work

1. Bonacci W, Teng PK, Afonso B, Niederholtmeyer H, Grob P, Silver PA, et al. Modularity of a carbon-fixing protein organelle. Proceedings of the National Academy of Sciences 2012;109:478–83. doi:10.1073/pnas.1108557109.

Table 2-2. Oligos used in Chapter 2

Oligo name	Description	Product	Restriction site	Sequence
TC26	Primes 3'-end of *csoS2.*	pBz13	BsaI	cacaccaggtctcacgctttacttaatcaaccgcgcg
TC62	Primes 5'-end of *csoS2.* Includes sequence coding His tag	pBz13, 16, 19, 31, 32, 33	BsaI	cacaccaggtctcagtcccaccatcatcatcatcat aatcctgccgacctgagcg
TC71	Primes 3'-end of *csoS2.* Includes sequence coding Strep-II tag	pBz16 and 19	BsaI	ggtctcacgcttacttttcgaactgcgggtggctcca tcctcgtgctcctccggagtaag
TC77	Primes nucleotide 1283-1302 of *csoS2* in the reverse direction. The BsaI recombination site is compatible with TC78.	pBz19	BsaI	cacaccaggtctcagcaggtgctccgttgatcg
TC78	Primes 5'-end of the csoS2-NonStop gBlock in the forward direction. The recombination site is compatible with TC77	pBz19	BsaI	cacaccaggtctcactgcaaaagttgccagaaccca
TC105	Primes 3'-end the sequence speculated to encode the end of CsoS2A. A nucleotide is inserted so that DVR peptide from -1 frame is expressed in 0 frame.	pBz31	BsaI	aggtctcacgcttcacttaattatcgtacatccttaggtgcatgatctagctgttggcca
TC106	Primes upstream of the slippery sequence in *csoS2* in the reverse direction. BsaI recombination site is compatible with TC107	pBz32	BsaI	cacaccaggtctcaaggtgcatgatctagctgttggcca
TC107	Primes the slippery sequence in the forward direction and introduces 2 point mutations in the slippery sequence. BsaI recombination site is compatible with TC106	pBz32	BsaI	cacaccaggtctcaacctaagatgtccggtgacgag
TC108	Primes the stem-loop region in the forward direction and introduces 4 out of 6 point mutations in the stem-loop. BsaI recombination site is compatible with TC108	pBz33	BsaI	cacaccaggtctcatacatccacctcgttcgtcaccggaca
TC109	Primes the beginning of the stem-loop region in the reverse direction and introduces 2 out of 6 point mutations in the stem-loop. BsaI recombination site is compatible with TC109	pBz33	BsaI	cacaccaggtctcatgtatgccagttacaggtaatgagtactacggt
TC161	Used with CN6 to amplify genes upstream of *csoS2* and a part of backbone of pHnCB10LC	pBz51, 52	BsaI	cacaccaggtctcattcctgactgtgaaggcatg

Oligo name	Description	Product	Restriction site	Sequence
TC171	Used with CN7 to amplify genes downstream of *csoS2* and a part of backbone of pHnCB10LC	pBz51, 52	BsaI	cacaccaggtctcaattaagtaaagtgtaacggtatatcatgaacac
CN6	Used with TC161 to amplify genes upstream of *csoS2* and a part of backbone of pHnCB10LC	pBz51, 52	BsaI	cacaccaggtctcatacgacgataccgaagacagct
CN7	Used with TC171 to amplify genes downstream of *csoS2* and a part of backbone of pHnCB10LC	pBz51, 52	BsaI	cacaccaggtctcacgtatcccactaccgagatatccg
TC172	Primes 5'-end of *csoS2* in forward direction. BsaI recombination site is compatible with TC161	pBz51, 52	BsaI	cacaccaggtctcaggaatgaatcctgccgacctgagcg
TC173	Similar to TC105, but with BsaI recombination site compatible with TC171	pBz51	BsaI	cacaccaggtctcataatttatcgtacatccttaggtgcatgatct
TC174	Primes 3'-end of *csoS2* in pBz32. BsaI recombination site compatible with T 171	pBz51	BsaI	cacaccaggtctcataatttatcctcgtgctccgccggagta
RJN305	Primes 5'-end of *egfp* gene in a forward direction	pRJN-FS, pRJN-control	BsmBI	cacaccaggtctctgtcccaccatcatcatcatcacgtgagcaagggcgagg
RJN306	Primes 3'-end of *egfp* gene in a reverse direction	pRJN-FS, pRJN-control	BsmBI	cacaccaggtctcttcagttcaggttgactcgtctcttatggtacagctcgtccatgccg
RJN307	Primes 5'-end of *mcherry* gene in a forward direction	pRJN-FS, pRJN-control	BsmBI	cacaccaggtctctctgagactcgtctcagctaggcgggaaaacctgtacttccagtccg
RJN308	Primes 3'-end of *mcherry* gene in a reverse direction	pRJN-FS, pRJN-control	BsmBI	cacaccaggtctcacgctttaaccgctcttgtacagttcg
RJN309	An oligo that can anneal with RJN310 to create a double-stranded DNA containing the slippery sequence and the stem-loop	pRJN-FS	BsmBI	cacaccacgtctctcatagcccaaagatgtccggtgacgagcgcggcgggtgcatgcccgtcaccggtaatgagtactacggtcgtgctaagagacgtggtgtg
RJN310	A complementary oligo to RJN309	pRJN-FS	BsmBI	cacaccacgtctcttagcacgaccgtagtactcattaccggtgacgggcatgcacccgccgcgctcgtcaccggacatctttgggctatgagagacgtggtgtg
RJN311	An oligo that can anneal with RJN312 to create a double-stranded DNA expressing the short linker	pRJN-control	BsmBI	cacaccacgtctctcatagcgcggcgggcggtgcaggagctaagagacgtggtgtg
RJN312	A complementary oligo to RJN311	pRJN-control	BsmBI	cacaccacgtctcttagctcctgcaccgccccgccgcgctatgagagacgtggtgtg

2.3 Results

2.3.1 CsoS2 Isoforms were not Caused by Differential Glycosylation or Proteolysis

As briefly discussed in the Introduction, the primary DNA sequence of *csoS2* from H. neapolitanus encodes a 92.0-kDa protein. Inspection of this amino acid sequence

reveals a protein with three regions, an N-terminal Domain, a Middle Region, and a C-terminal Domain (NTD, MR, and CTD) (Figure 2-1A). Previous bioinformatic analysis suggested that the NTD possesses four short homologous repeats and that the MR possesses six homologous repeats of roughly ~60 amino acids in length each possessing three canonical [V/I][S/T]G motifs (R1–R6; Figure 2-1A). Inspection of the primary sequence reveals the possibility of three additional repeats in the CTD. These repeats (R7–R9) possess extended interstitial amino acids compared to repeats in the MR. CsoS2 also possesses an extremely conserved terminus, which we term the C-terminal peptide (CTP) (Figure 2-1B). Finally, CsoS2 is also predicted to be highly intrinsically disordered, which has been experimentally validated in solution scattering experiments (Cai et al., 2015).

In order to link this sequence with function, we sought to understand how one such gene can generate two gene products. Heterologous expression of CsoS2 indicated that the mechanism of differential expression used to generate CsoS2A and CsoS2B is conserved between the gamma proteobacteria *E. coli* and *H. neapolitanus*. Carboxysomes purified from either *H. neapolitanus* (HnCB) or *E. coli* (EcCB; see (Bonacci et al., 2012)) produce bands at the characteristic, anomalous molecular masses of 120 kDa and 70 kDa and with roughly 1:1 stoichiometry (Figure 2-1C). Likewise, expression of CsoS2 protein bearing an N-terminal 6× polyhistidine tag (EcS2) also produced both the A form and the B form, with a preference for CsoS2A possibly due to higher degradation of CsoS2B or selective enrichment of CsoS2A during the purification procedure. Previous results indicate that differential expression may arise from post-translational processing (Baker et al., 1999) but attempts to verify glycosylation or proteolytic activation were negative (Figure 2-2A, B).

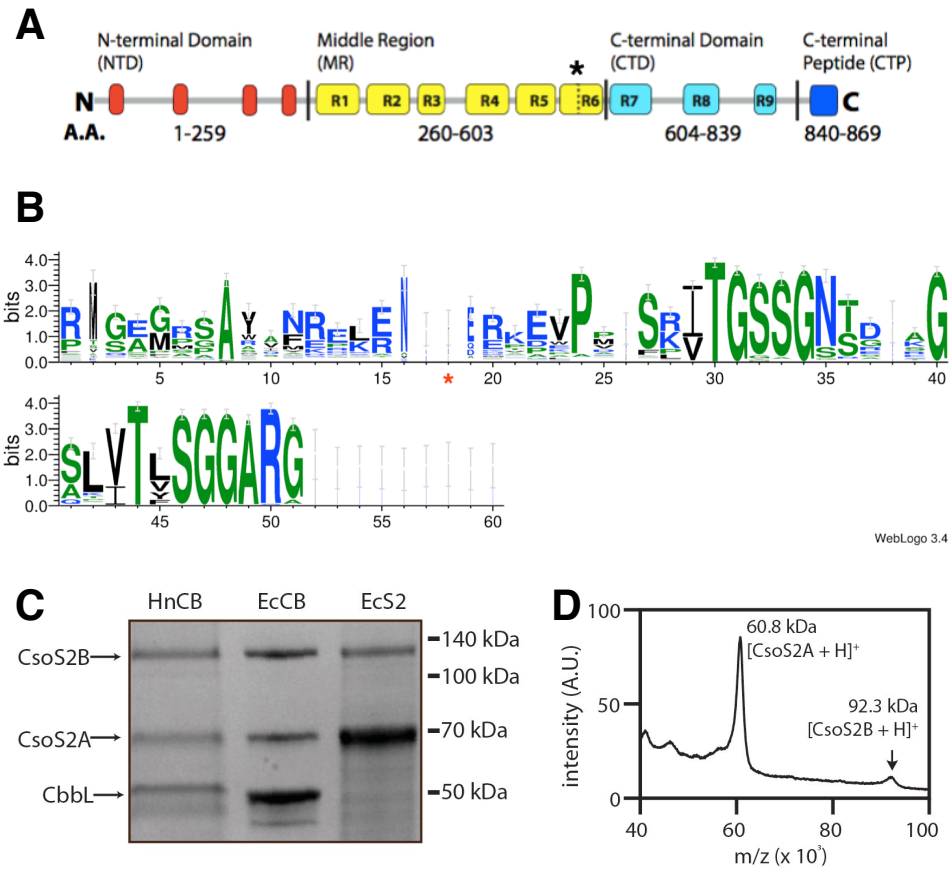

Figure 2-1: CsoS2 is differentially expressed as two proteins without any post-translational modifications.
(A) A schematic representation of different regions in CsoS2, including N-terminal Domain (NTD), Middle Region (MR), C-terminal Domain (CTD), and a conserved C-terminal peptide (CTP). NTD contains four short repeats (yellow). MR has six longer repeats (R1– R6). CTD has three repeats (R7–R9). The repeats are divided by interstitial amino acids (gray), which do not align with any of the repeats. The short isoform, CsoS2A, ends within R6 (asterisk). (B) HMM logo of the C-terminus of CsoS2. The red asterisk denotes the start of the CTP region. (C) SDS-PAGE gel of carboxysomes purified from *H. neapolitanus* (HnCB) and from recombinant expression in *E. coli* (EcCB) and CsoS2 recombinantly expressed *in E. coli* (EcS2). The gel is annotated with the identity of each protein band and MW markers. (D) MALDI-TOF spectrum obtained from purified EcS2.

Intrinsically disordered proteins often run at larger apparent MWs in denaturing PAGE, offering a partial explanation for the observed anomalous masses (Tompa, 2002). We next verified that the EcS2 mass matched that of the theoretical gene product using matrix-assisted laser desorption/ionization–time of flight (MALDI-TOF) mass spectrometry (MS). This revealed two peaks with large m/z, corresponding to masses of 60.8 kDa and 92.3 kDa (Figure 2-1D). The latter is an identical (within 0.01%) match to the predicted mass of full-length EcS2 and is expected to be CsoS2B. We therefore concluded that the 60.8-kDa species corresponded to CsoS2A and sought to identify, using this MW as a guide, the mechanism of CsoS2A expression.

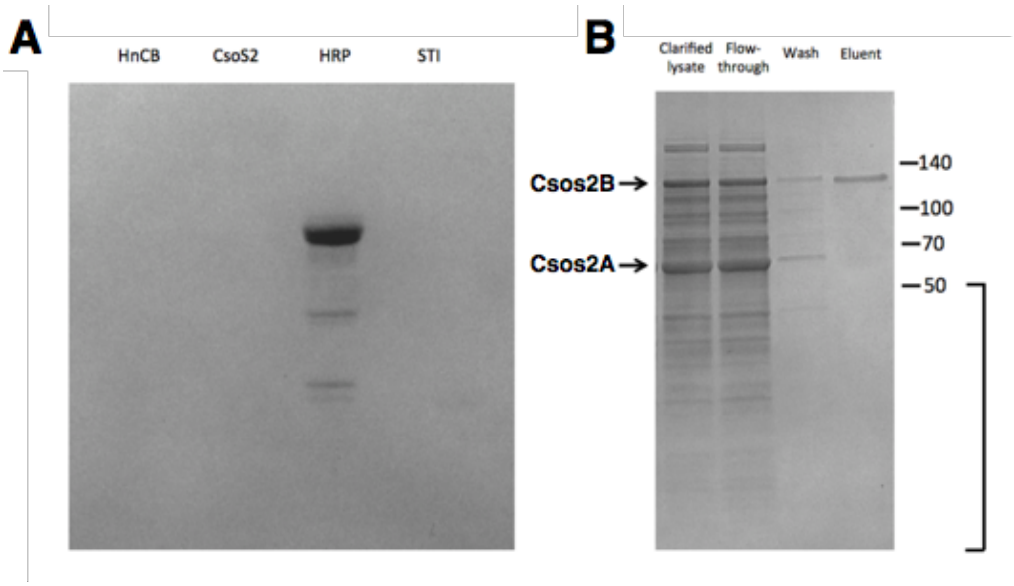

Figure 2-2: CsoS2 is neither glycosylated nor expressed as a protein precursor.
(A) Pierce glycoprotein staining assay was applied to an SDS-PAGE gel containing purified *H. neapolitanus* carboxysomes (HnCB, lane 1), wild-type CsoS2 (CsoS2, lane 2), Horse-radish peroxidase (HRP, lane 3), Soy-bean trypsin inhibitor (STI, lane 4). HRP and STI were positive and negative controls provided with the kit. (B) Strep-tag II affinity purification of C-terminally StrepII tagged CsoS2 yielded only CsoS2B. A Coomassie-stained SDS-PAGE gel of fractions from the purification is shown here. No C-terminal cleavage product was co-eluted and observed in the bracketed region of SDS-PAGE gel, suggesting that the CsoS2A did not arise from specific proteolysis of a CsoS2 precursor. Arrows indicate the locations to which CsoS2B and CsoS2A usually migrate.

2.3.2 -1 Programmed Ribosomal Frameshifting is Responsible for CsoS2A Formation
The lack of post-translational processing led us to ask whether an alternative mechanism, such as co-translational regulation via frameshifting, could explain the presence of CsoS2A. Based on the MALDI-TOF data, we hypothesized that the terminal stop codon for CsoS2A was likely to be in the latter half of the gene due to its mass of 60.8 kDa. To probe for PRF, we designed a construct, EcCsoS2-NonStop, possessing synonymous mutations ablating the 22 stop codons in the − 1 and +1 frames of the latter half of the gene. The construct also contained stop codons in the −1 and +1 frames just downstream of the normal 0 frame stop. Strikingly, expression of EcCsoS2-NonStop generated only a single band with MW similar to CsoS2B (Figure 2-3A). Therefore, expression of a synthetic CsoS2 revealed that frameshifting is likely responsible for the differential expression of the two isoforms.

Further MS analysis confirmed this hypothesis and identified the site of frameshifting. CsoS2A and CsoS2B were isolated from PAGE bands of purified EcS2, trypsinized, and analyzed using tandem mass spectrometry (MS/MS). Masses from this experiment were matched to a theoretical MW library generated from the CsoS2 sequence in order to identify peptides specific to each isoform. CsoS2A and CsoS2B shared a large number of peptides in the first two-thirds of the protein sequence,

but there was a dramatic reduction in CsoS2A peptides beyond amino acid 570 (Figure 2-3B). The recent work of Cai *et al.* noted a similar observation (Cai et al., 2015).

PRF generates proteins possessing an N-terminus from the 0 frame and a C-terminus from either the − 1 or the +1 frame. We therefore repeated the peptide identification analysis but instead used a theoretical frameshifting-specific library, containing every possible frameshifted polypeptide from the *csos2* gene. Surprisingly, this analysis revealed a unique peptide—HAPKDVR—containing a fusion of amino acids from the 0 and −1 frames (Figure 2-2C). The theoretical mass of a CsoS2A protein ending in this sequence exactly matches the experimental value of 60.8 kDa from Figure 2-1D, thus providing additional confirmation.

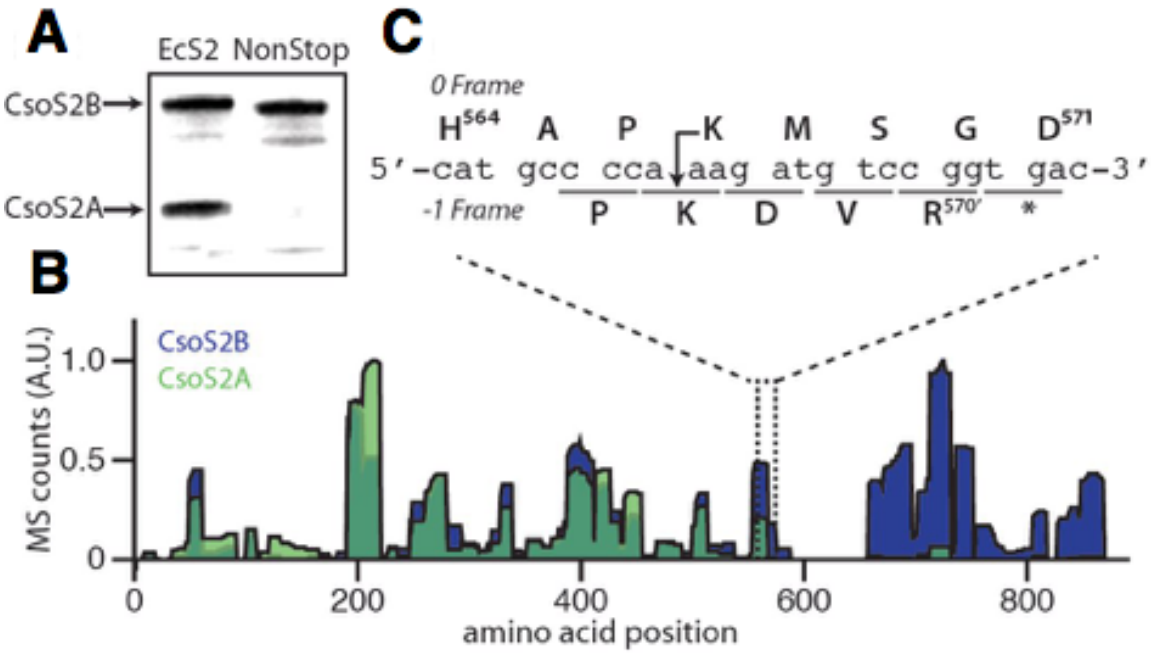

Figure 2-3: The *csoS2* gene encodes a programmed ribosomal frameshift event.
(A) Western blot of CsoS2 (EcS2) and CsoS2-NonStop (NonStop) expressed in *E. coli*. (B) Piling diagram prepared from the list of peptides discovered by MS/MS. Scale is normalized to maximum count for each sample. CsoS2A is colored translucent green; CsoS2B is in blue. (C) Diagram showing *csoS2* codons in the 0 and −1 frames where the −1 PRF occurs. Codons in the −1 frame are underlined. Amino acids in the 0 and − 1 frames are written above and below the codons, respectively. The arrow denotes the last amino acid shared between CsoS2A and CsoS2B. The diagram is mapped back to the corresponding region in the piling diagram (marked by broken lines).

Strikingly, the *csoS2* gene contains similar *cis* elements to *dnaX* described by Kornberg *et al.* at the gene sequence neighboring the sequence encoding the HAPKDVR peptide (Fig. 2-4B). Most importantly, *csoS2* contains a putative slippery sequence, C-CCA-AAG, matching the X-XXY-YYZ motif, directly where there is

ambiguity between the 0 frame and the −1 frame in the MS/MS data (arrow in Figure 2-3C). In addition, RNAstructure analysis (Reuter and Mathews, 2010) of the primary nucleotide sequence indicates a potential stem–loop structure with significant stability (25.7 kcal/mol), downstream of the slippery sequence. This loop contains the C-terminus of the putative HAPKDVR peptide and a stop codon immediately following the arginine codon, suggesting that this is the terminus of CsoS2A.

2.3.3 Frameshifting Elements are Necessary and Sufficient for -1 PRF in CsoS2
Our model featuring PRF was tested by generating constructs deficient in frameshifting elements. First, *csoS2* was recoded to end with the HAPKDVR-stop sequence in the 0 frame (EcS2A). Expression of this construct was equal in size to CsoS2A and confirmed its anomalous PAGE behavior (Figure 2-4C). To test the role of the slippery sequence and secondary structure elements, we next designed constructs with a series of synonymous mutations to ablate either the slippery sequence (NoSlip) or the stem–loop (NoStem) and to perturb PRF (orange and red circles in Fig. 2-4B). As with EcCsoS2-NonStop, these constructs expressed a long peptide matching the MW of CsoS2B (Fig. 2-4C). Thus, both the slippery sequence and the downstream secondary structure are required for − 1 PRF.

To unambiguously assay the function of the PRF elements, we tested for frameshifting in a completely synthetic context. A region of the *csoS2* gene containing the complete PRF elements and additional 15 base pairs downstream of the stem–loop was cloned between a green fluorescent protein (GFP) and mCherry red fluorescent protein (RFP) reporter such that the 0 frame of the PRF elements would yield a full-length GFP–RFP fusion (Figure 2-4E). Expression of this construct generated two fluorescent species—the expected full-length protein displaying both GFP and RFP fluorescence and a short form containing only GFP—thus confirming frameshifting (Figure 2-4D). Interestingly, the ratio of this construct favors the short form while the ratio of CsoS2A and CsoS2B is roughly 1:1 from the carboxysome operon. Regardless, this result demonstrates that the PRF elements are robust to context and are sufficient to recode the expression of a translating message. Taken together, the abovementioned experiments suggest that *csoS2* undergoes a programmed ribosomal frameshift to yield the different forms of CsoS2. More specifically, we conclude that CsoS2A is a 570-amino-acid protein whose terminal amino acids are encoded within the −1 frame of the gene.

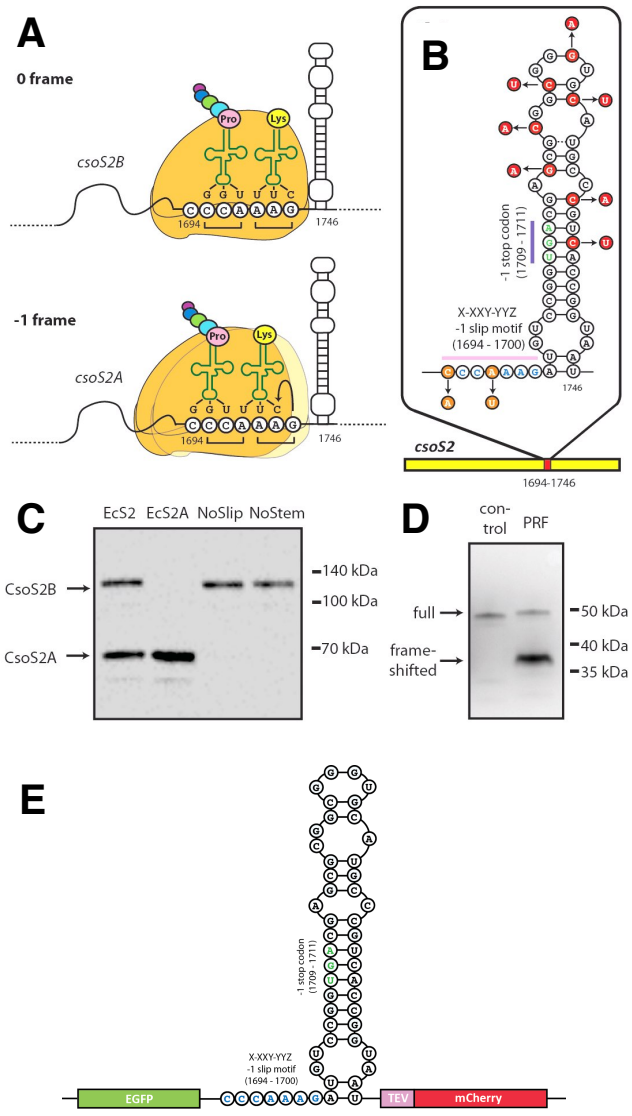

Figure 2-4: Both the slippery sequence and the secondary structure in *csoS2* mRNA are necessary and sufficient for frameshifting.
(A) A predicted stem–loop in the *csoS2* mRNA causes the ribosome to stall during translation. The ribosome may continue translating in the 0 frame, resulting in CsoS2B. Alternatively, the slippery sequence allows the ribosome to slip backward to wobble base pair and continue translating in the – 1 frame to generate CsoS2A. Codons in the original frame are marked by horizontal lines underneath. (B) Predicted secondary structure of the slip sequence and downstream stem–loop (base pairs 1694–1746). Point mutations were made either in the slippery sequence (orange) or in the stem–loop (red). Bases at the end of the arrows outside of the structure are those in the mutated products. Pink and blue bars indicate the slippery sequence and the −1 frame stop codon, respectively. (C) Western blot of recombinant WT CsoS2 (EcS2), CsoS2A (EcS2A), CsoS2 with an ablated slippery sequence (NoSlip), and CsoS2 with an ablated stem–loop (NoStem). (D) GFP fluorescence detected on the SDS-PAGE gel of synthetic EGFP-RFP fluorescent reporters containing a linker (control) or PRF elements (PRF). (E) Schematic diagram of the frameshift reporter construct. The plasmid expressing the reporter construct contains the sequence expressing enhanced green fluorescent protein (EGFP), the slip sequence and stem-loop region from *H. neapolitanus csoS2*, the TEV protease cleavage site, and mCherry. The slippery sequence is highlighted in blue and the -1 frame stop codon is highlighted in green. In the control the sequence in the middle is replaced by an Ala-Ala-Gly-Gly-Ala-Gly-Ala linker.

2.3.4 -1 PRF is Conserved Among CsoS2 Homologs

We next investigated the conservation of −1 PRF in CsoS2 homologs from other organisms. The −1 PRF prediction server KnotInFrame (Theis et al., 2008) with additional criteria (see Materials and Methods), was used to search for −1 PRF candidates among known CsoS2-bearing species (Figure 2-5). Intriguingly, *csoS2* genes from 79 out of 162 species were predicted to undergo frameshifting and the distribution of PRF across the phylogenetic tree appears to be somewhat clustered, suggesting that this mechanism is widely used and conserved among the species with closely related CsoS2 homologs. Although further experimental data are needed to validate each prediction, we found that a number of them matched previously reported observations. For instance, *H. neapolitanus* CsoS2 is predicted to frameshift at the location we experimentally confirmed in this study and the predictions matched previous unexplained observations of carboxysomes possessing either one or two CsoS2 isoforms in three different species (*Thiomonas intermedia*, *Thiomicrospira crunogena*, and *Prochlorococcus marinus* str. MED4) (Dou, 2009; Roberts et al., 2012). Closer inspection of each organism's characteristics, habitat, and CCM function may shed light on why −1 PRF exists in some bacteria but not in others.

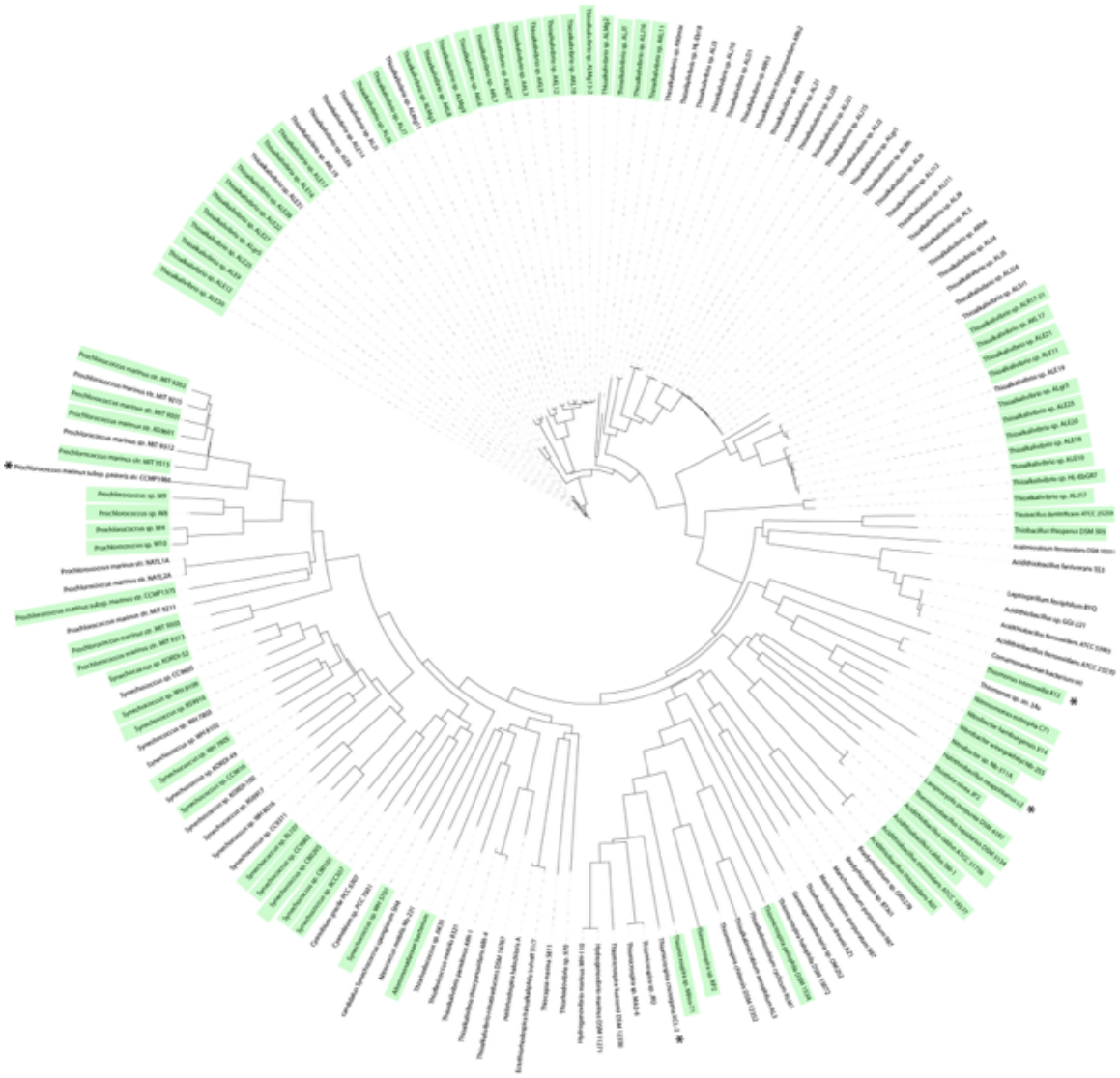

Figure 2-5: Conservation of candidate -1 PRF sequences amongst 162 *csoS2* bearing species.
For the 162 species mapped on this phylogeny, 79 (marked in green) are predicted to have a significant –1 PRF event in the last two thirds of the *csoS2* mRNA as determined by the KnotInFrame -1 PRF prediction tool. Note that *P. marinus* str. MED4 is equivalent to *Prochlorococcus marinus* subsp. *pastoris* strain CCMP1986

2.3.5 CsoS2B is Required for the Carboxysome Formation

We sought to determine the significance of each isoform for carboxysome assembly. Mapping the location of PRF onto the primary sequence of CsoS2 shows that the site is within Repeat 6 at the end of the MR. Thus, CsoS2A contains the NTD and five complete repeats within the MR (asterisk in Figure 2-1A), while CsoS2B contains the entire CTD, including the highly conserved CTP. We therefore hypothesized that this variation in sequence could create a functional difference between CsoS2A and CsoS2B.

The importance of these sequences was assayed using heterologous carboxysome expression in *E. coli*. It has been shown previously that 10 genes from two *H. neapolitanus* operons can be refactored into a single IPTG-inducible operon and expressed in *E. coli* (EcCBs) (Bonacci et al., 2012). EcCBs are generally well-formed and are functional for CO_2 fixation but, in contrast to native HnCBs, are somewhat more heterogeneous in morphology and can possess broken shells. Regardless, this provides a simple platform to assay an *in vivo* reconstitution of carboxysome assembly as a function of CsoS2 sequence.

Expression of EcCBs in this assay revealed a critical functional difference between CsoS2A and CsoS2B. The EcCB expression plasmid was modified to encode CsoS2A, CsoS2B (with ablated slippery sequence), or wild-type (WT) CsoS2, and carboxysomes were expressed and analyzed for formation by sucrose gradient purification and EM. In an SDS-PAGE analysis, WT and CsoS2B- containing carboxysomes produced similar gradient fractionation patterns, but CsoS2A yielded no fractions characteristic of carboxysomes (Figure 2-6A). EM of WT and CsoS2B showed the formation of carboxysome-like structures while the CsoS2A-derived fraction lacked any significant assembly (Figure 2-6B). Thus, CsoS2A, in the absence of CsoS2B, is incapable of forming carboxysomes while CsoS2B by itself is sufficient for shell assembly around the lumen and form carboxysome-like structures.

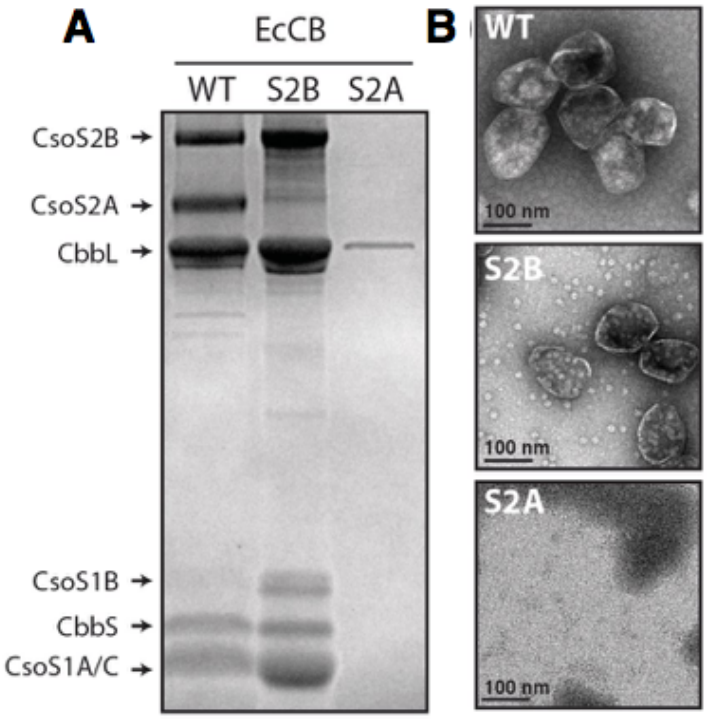

Figure 2-6: Heterologous expression of carboxysomes without either CsoS2A or CsoS2B.
(A) SDS-PAGE of purified carboxysomes containing both CsoS2 isoforms (WT), CsoS2B only (S2B), and CsoS2A only (S2A). Each sample was prepared from fraction 16 of the sucrose gradient. (B) Electron micrographs of the carboxysome samples used for SDS-PAGE gel (scale bar: 100 nm).

2.3.6 Validation of CsoS2 Frameshifting in H. neapolitanus

Thus far, our efforts to study the frameshifting rely on heterologous expression of CsoS2 and its mutants in *E. coli*. Our experimental design was based on the assumption that the findings from *E. coli* would apply to *H. neapolitanus*, considering that both are gammaproteobacteria. To validate this assumption and test the generalizability of our results, we set out to determine whether perturbing the frameshifting elements would result in the same outcome in *H. neapolitanus.*

Transformation of *H. neapolitanus* with foreign DNA was previously shown to be feasible (English et al., 1995). Using electroporation, plasmids can be introduced into the cells, after which the homologous recombination between the gene cassette on the plasmid and the target gene locus in the chromosome can take place. This strategy allows *csoS1A*- and *cbbLS*-knockout mutants of *H. neapolitanus* to be generated (English et al., 1995; Menon et al., 2008). However, a former graduate rotation student in our lab, Ryan Protzko, discovered that *H. neapolitanus* is in fact naturally competent and the transformation only requires incubation with linearized DNA. This simple transformation protocol was used for the rest of the study.

To abolish frameshifting in CsoS2, the point mutations must be introduced to the wild-type *csoS2* gene in the cso operon. The simplest strategy was to replace the entire wild-type *csoS2* gene with the NoSlip variant using one-step homologous recombination. However, this option would necessitate the inclusion of an antibiotic resistance marker in addition to the gene of interest, which could potentially give rise to a polar effect. Therefore, we developed a homologous recombination strategy that would not leave the antibiotic resistance marker in the product strain, based on SacB-dependent positive selection. This strategy relies on the toxicity of polymerized sucrose, called levansucrose, towards certain bacteria such as *E. coli*. When expressed in susceptible bacteria in the presence of sucrose, *Bacillus subtilis* levansucrase (SacB), encoded by *sacB*, can accumulate levansucrose which is lethal to the cells (Gay et al., 1985). Homologous recombination carried out between chromosomally-integrated *sacB* and the gene of interest would result in viable clones only if *sacB* is successfully replaced by the desired gene. For this reason, no antibiotic resistance marker is needed in the incoming gene cassette.

To carry out this strategy, we first replaced *csoS2* gene in *H. neapolitanus* with a *sacB-kan*^{r} cassette (Figure 2-7). We then spread the resulting *csoS2wt*::*sacB-kan*^{r} mutant on an agar plate containing 10% sucrose and allowed the cells to grow in 5 % CO_2. As a control, wild-type cells were grown under the same condition. Unexpectedly, while *csoS2wt*::*sacB-kan*^{r} mutant formed colonies later than the wild-type, the number of colonies were not significantly different between both strains. This result indicates that levansucrose may not be as toxic to *H. neapolitanus* as it is to *E. coli* and an alternative counter selection method may be required.

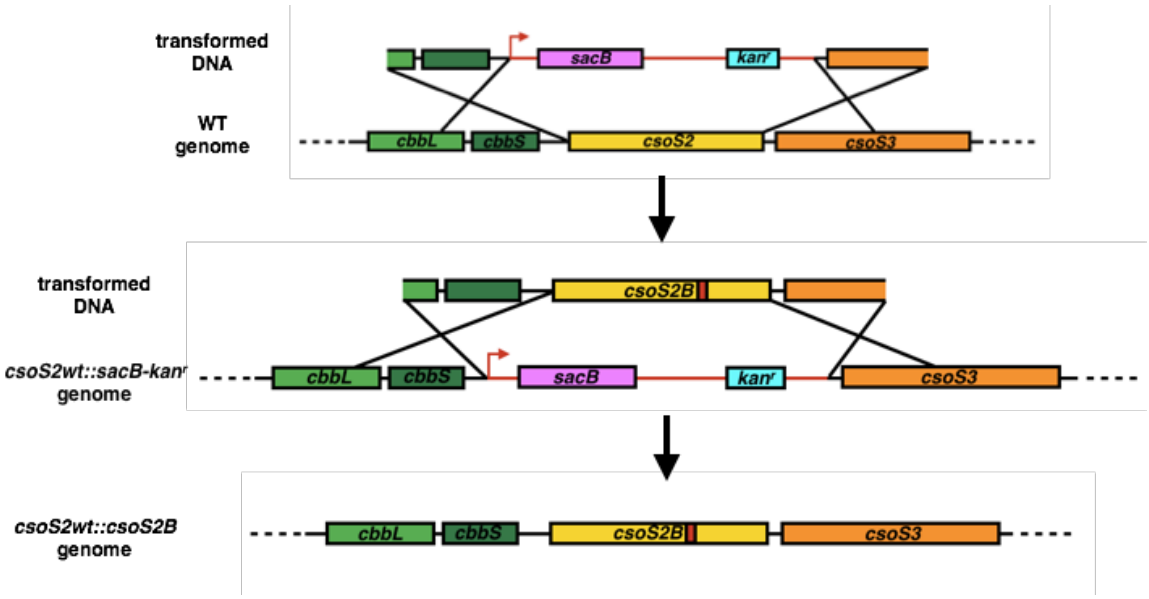

Figure 2-7: Two-step construction of *csoS2wt::csoS2B* mutant from wild-type *H. neapolitanus*. First, *csoS2* ORF is replaced with *sacB-kan*^{r} cassette via homologous recombination, yielding *csoS2wt::sacB-kan*^{r} strain. The homology regions (~1000 bp each) are designed such that the upstream homology region ends immediately before *csoS2* start codon and the downstream homology region begins right after the stop codon. Homologous recombination is carried out once more to replace *sacB-kan*^{r} with NoSlip *csoS2* (displayed as "*csoS2B*"). The desired colony is positively selected by spreading the transformants on a sucrose-containing agar.

As an alternative to the SacB-dependent selection, we developed a new selection assay based on the expected phenotypic difference between *csoS2wt*::*sacB-kan*^{r} and *csoS2wt*::*csoS2B* strains. Specifically, *csoS2wt*::*sacB-kan*^{r} strain was expected to display a high-CO_2 requiring phenotype due to its inability to form carboxysomes. On the other hand, *csoS2wt*::*csoS2B* was expected to form intact carboxysomes and to be able to grow at the atmospheric level of CO_2. Following this rationale, a linear piece of DNA containing NoSlip *csoS2* variant flanked by upstream and downstream homology regions was introduced into the *sacB-kan*^{r} mutant (Figure 2-7). The transformants were plated and incubated in the atmospheric CO_2 level. Although only one colony was able to form after a week (corresponding to the transformation efficiency of 12 cfu/μg of DNA), the *csoS2* gene amplified by colony PCR contained the desired point mutations, as verified by Sanger sequencing. The growth of this colony in liquid media at atmospheric level of CO_2 was comparable to wild-type *H. neapolitanus*. Therefore, the success of this method confirms that carboxysomes can form in the absence of CsoS2A and possess sufficient catalytic function for cell growth at atmospheric level of CO_2.

To investigate the morphology of the carboxysome lacking CsoS2A (“S2B-CB”) , mutant carboxysomes were isolated and visualized by TEM. First, 10 L of *csoS2wt*::*csoS2B* was grown in the fermentor to the stationary phase and harvested. Cell lysis and differential ultracentrifugation were performed as described for *E. coli*-derived carboxysomes. As a positive control, carboxysomes were also prepared from wild-type *H. neapolitanus*. Strikingly, the morphology of the acquired S2B-CB

was indistinguishable from the wild-type, despite the slightly smaller size (Figure 2-8B, C). Denaturing SDS-PAGE analysis of S2B-CB showed a 120-kDa band corresponding to CsoS2B, but not the 70-kDa CsoS2A band (Figure 2-8A). Collectively, these results confirm that -1 PRF is employed to generate CsoS2 isoforms in the native organism, although disrupting this process does not appear to be detrimental to the cell under our experimental conditions.

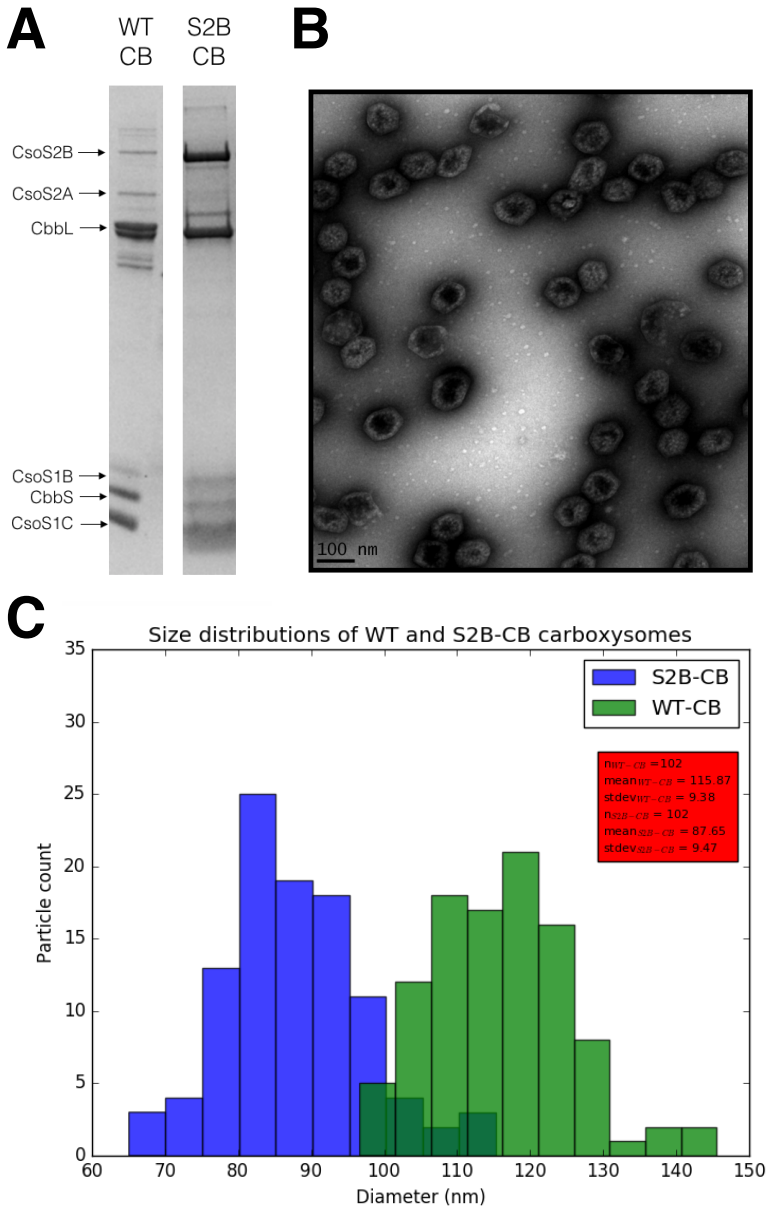

Figure 2-8: *H. neapolitanus* mutant without CsoS2A produces intact but smaller carboxysomes
(A) SDS-PAGE gel of carboxysome-containing fractions from sucrose-fractionation of WT-CB and S2B-CB. (B) Electron micrograph of purified S2B-CB (scale bar: 100 nm). (C) Histogram showing diameter distributions of WT (green) and S2B-CB (purple) carboxysomes. The number of particles measured, the average diameter, and the standard deviation are shown in the red box.

2.4 Discussion

2.4.1 Frameshifting and the Two Forms of CsoS2

Since the initial biochemical characterization of α-carboxysomes, the anomalous behavior of CsoS2 expression has confounded the determination of its molecular function. Here, we show that −1 PRF is responsible for the genesis of the two forms of CsoS2.

−1 PRF occurs when the ribosome is stalled during translation by 3′ secondary structure and, in certain cases, a 5′ internal ribosomal binding site-like sequence. A slippery sequence of the mRNA allows the tRNAs to shift backward to wobble base pair with the codons in the −1 frame. Slipped ribosomes that continue translating in the −1 frame can then encounter a stop codon and can then lead to early protein truncation (Figure 2-4A). Inspection of the *csoS2* nucleotide sequence revealed that it possesses both a potential slippery sequence and a significant predicted 3′ stem–loop. A relatively purine-rich sequence (GGCCAA) is observed 14 nucleotides upstream of the slippery sequence. However, since the sequence is significantly different from the consensus, it is uncertain whether this sequence could act similar to a ribosome-binding site and could contribute to PRF. Removal of either the slippery sequence or stem–loop abolishes CsoS2A, suggesting that these elements are necessary for frameshifting. In addition, PRF function is robust to sequence context, as frameshifting can occur when the *csoS2* gene is expressed by itself in *E. coli* or completely out of context, as seen in the synthetic fluorescent reporter containing only the slippery sequence and the stem–loop. Therefore, these elements are sufficient to explain the generation of two CsoS2 isoforms.

The efficiency of known −1 PRF elements can range from a few percent to as high as 80% (Caliskan et al., 2015). The CsoS2 PRF is roughly 50% and is therefore near the upper range of characterized examples across all kingdoms of life. The dearth of model frameshifting systems in prokaryotes has made the prediction of efficiency difficult; thus, CsoS2 provides a novel example to better inform this understanding. Clearly, extensive Watson–Crick base pairing and high GC content in the stem–loop (Figure 2-4B) are important for efficiency, but the overall efficiency is likely to be multivariate and context-dependent (Mouzakis et al., 2013).

While we portray the − 1 PRF of CsoS2 as entering the − 1 frame when CCA and AAG codons are at the P and A sites of the ribosome (Figure 2-4A), respectively, a recent study has shown that PRF can operate through multiple pathways simultaneously (Yan et al., 2015). Specifically, the translation apparatus can enter the −1 frame from several codons along the slippery sequence and can also slip by − 4 or + 2 nucleotides. Therefore, it is formally possible that CsoS2A in its native context could be the product of not one but of several translational pathways. Further MS studies, by probing for peptides from multiple frames, could elucidate the exact slip events that occur during CsoS2 translation.

Finally, it is not known what advantages are conferred by using −1 PRF as opposed to other strategies in generating isoforms. As −1 PRF is known to result in tight stoichiometric control, it is possible that α-carboxysome assembly benefits from having a fixed ratio between CsoS2AB (Xu et al., 2014). Notably, in β-carboxysomes, CcmM has two isoforms, the short 35-kDa protein (M35) and the long 58-kDa protein (M58), the former generated via an internal ribosome entry site (IRES) (Long et al., 2007; 2010). Despite different mechanisms, both −1 PRF and IRES involve mRNA secondary structure, but direct comparison between the biological advantages of PRF and IRES is unclear. One possibility is that size heterogeneity differences between α- and β- carboxysomes (Rae et al., 2013) are derived from the differences in function between CsoS2 and CcmM/CcmN and the regulation of their expression by PRF and IRES elements, respectively.

2.4.2 Frameshifting and the Mechanism of Carboxysome Assembly

Several lines of evidence hint at the possibility that CsoS2 may play a central role in organizing carboxysome assembly. However, since it was previously not possible to individually express each form of CsoS2, the functional difference between CsoS2A and CsoS2B has not been elucidated. Here, we use our findings on PRF to specifically produce carboxysomes with either CsoS2A or CsoS2B in *E. coli*. Surprisingly, we find that CsoS2B is necessary and, at least to the resolution of our experiments, is sufficient for the assembly of carboxysomes similar to those possessing both isoforms. Since CsoS2A is not required in this process, it is perplexing that this isoform is produced in a significant amount in the native organism and that the frameshifting elements are widely conserved across clades in the phylogeny of CsoS2. Future experiments, probing the structure and function of carboxysome in a more fine-grained fashion, may reveal a role for the CsoS2A isoform.

While we have clarified the molecular identity of each CsoS2 isoform, their exact functions remain unclear. We speculate on their involvement in α-carboxysome assembly by analogy to the CcmMN system. M35 has been shown to nucleate RuBisCO, and M58 can connect the RuBisCO aggregate to the shell (Cot et al., 2008; Long et al., 2007; 2010). Along these lines, CsoS2A may interact with RuBisCO in the core of the lumen, while CsoS2B binds RuBisCO close to the periphery and uses the C-terminus to anchor them to the shell. Recent results place the C-terminus of CsoS2B either bound to the shell or outside it, which supports this hypothesis. Strikingly, EM images and in silico modeling suggested that the ratio of RuBisCO located immediately underneath the shell to those distributed in the remainder of the lumen is close to 1:1, matching the ratio of CsoS2A to CsoS2B in a carboxysome(Iancu et al., 2007). It is therefore possible that CsoS2A and CsoS2B, driven by the latter's localization to the shell, reside in different zones within the carboxysome (Figure 2-9). Further protein–protein interaction studies and higher resolution studies of α-carboxysome assembly are needed to test this hypothesis.

A major difference between CsoS2 and CcmM lies in the necessity of the isoforms. In β-carboxysomes, both CcmM isoforms are necessary for the assembly process (Long

et al., 2010). RuBisCO nucleation, mediated by M35, is required as the first step in the assembly (Cameron et al., 2013). On the other hand, the formation of α-carboxysomes likely proceeds without RuBisCO nucleation or at least in concert with it, as *H. neapolitanus* carboxysomes can assemble in the absence of lumenal RuBisCO (Iancu et al., 2010; Menon et al., 2008). The lack of requirement for CsoS2A, therefore, may indicate a different assembly pathway.

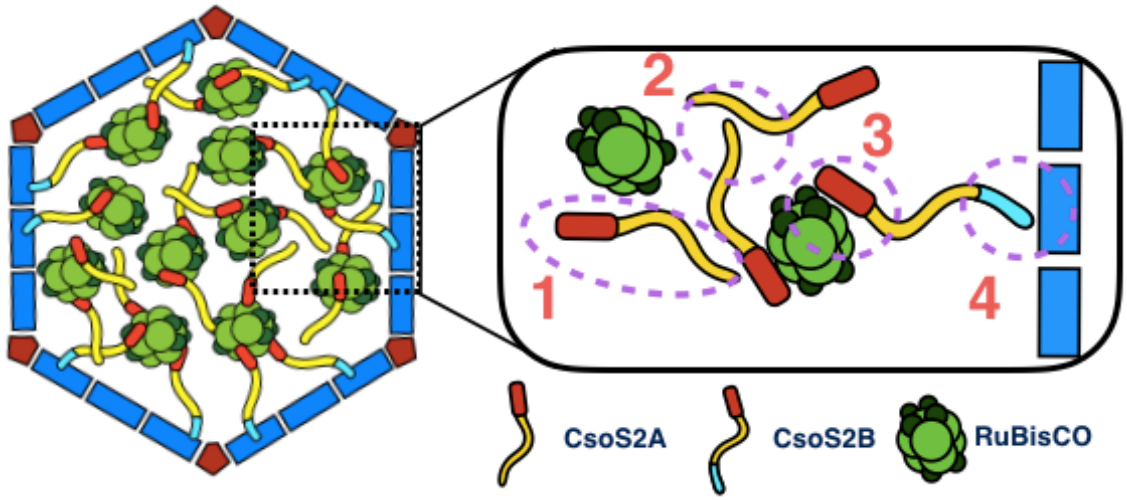

Figure 2-9: Model of possible differential functions of CsoS2 isoforms, in which CsoS2A organizes lumenal RuBisCO while CsoS2B bridges the cargo to the shell.
C-terminus of CsoS2B is colored blue. Areas with unresolved questions are circled with magenta broken lines: (1) the location of CsoS2A, (2) the possibility and the nature of an interaction between CsoS2A and CsoS2B, (3) the exact regions of CsoS2A or CsoS2B that interact with RuBisCO, and (4) the location of the C-terminus of CsoS2B.

Combining our findings above with previous work (Cai et al., 2015; Espie and Kimber, 2011; Rae et al., 2013), we arrive at a potential model describing assembly of the α-carboxysome. Given previous results, CsoS2B is localized near the shell and may facilitate RuBisCO's previously observed tight packing (Iancu et al., 2007) in this region, while CsoS2A primarily organizes RuBisCO in the remainder of the carboxysomal lumen (Fig. 2-9). Several unresolved questions in this model warrant further experimental investigation. First, the molecular recognition elements facilitating interaction between CsoS2AB and the other proteins remain unknown. M35, having three RuBisCO-like repeats, can organize RuBisCO into the previously observed aggregate. Repeats in CsoS2AB share no sequence similarity to RuBisCO; thus, the outcome and quality of any such aggregation may be different. In addition, there is a strong conservation of the last 30 amino acids of CsoS2B (i.e., the CTP), suggesting functional importance, but a specific binding partner has not been identified. Second, while we portray CsoS2A as residing with RuBisCO in the core of a carboxysome, we cannot rule out the possibility that some CsoS2A could exist in the subshell layer. Third, besides RuBisCO and shell proteins, CsoS2 may also interact with carbonic anhydrase. Finally, due to a number of conserved Cys

residues, the two isoforms of CsoS2 may covalently interact with one another (Cai et al., 2015).

2.4.3 Evolutionary Aspects of CsoS2

Although there is no direct evidence for a viral origin to carboxysomes and other bacterial microcompartments, their similarities are striking. Both types of particles are polyhedral, proteinaceous, and self-assembling. CsoS2 also shares additional characteristics of viral proteins: it is highly intrinsically disordered and undergoes PRF. The latter is common among viral proteins as viral constraints necessitate tight stoichiometric regulation and genetic economy. For example, a functionally similar protein to CsoS2, VP4, is present in foot-and-mouth-disease virus. Like CsoS2, VP4 is an intrinsically disordered protein, binds to the interior side of the capsid shell, and is necessary for assembly of the viral capsid (Lea et al., 1994; Xue et al., 2014). It remains to be seen whether these functional properties are, in fact, viral in origin or simply the result of the functionality required to facilitate the structural self-assembly of protein capsids.

While α- and β-carboxysomes appear to be evolutionarily independent, both of them employ two protein isoforms and repeating units to mediate the assembly. Such a strategy might provide a number of advantages. It has been reported that a multivalent protein can mediate the aggregation of its partner in a switch-like manner (Jiang et al., 2015; Li et al., 2012; Molliex et al., 2015). Repeats in CsoS2 and CcmM may provide multivalency for a similar mechanism. Moreover, in both classes of carboxysomes, the shell interaction appears to be mediated by the long form of these proteins yet a shorter, multivalent species is often present. These observations may hint at the nature and rules governing self-assembly of these protein complexes and therefore serve as the basis for future experiments interrogating the design principles of carboxysome assembly and function in the cytosolic milieu.

2.5 References

Baker, S.H., Lorbach, S.C., Rodriguez-Buey, M., Williams, D.S., Aldrich, H.C., and Shively, J.M. (1999). The correlation of the gene *csoS2* of the carboxysome operon with two polypeptides of the carboxysome in *Thiobacillus neapolitanus*. Arch. Microbiol. *172*, 233–239.

Bonacci, W., Teng, P.K., Afonso, B., Niederholtmeyer, H., Grob, P., Silver, P.A., and Savage, D.F. (2012). Modularity of a carbon-fixing protein organelle. Proc Natl Acad Sci USA *109*, 478–483.

Cai, F., Dou, Z., Bernstein, S., Leverenz, R., Williams, E., Heinhorst, S., Shively, J., Cannon, G., and Kerfeld, C. (2015). Advances in understanding carboxysome assembly in *Prochlorococcus* and *Synechococcus* implicate CsoS2 as a critical

component. Life *5*, 1141–1171.

Caliskan, N., Peske, F., and Rodnina, M.V. (2015). Changed in translation: mRNA recoding by −1 programmed ribosomal frameshifting. Trends Biochem. Sci. *40*, 265–274.

Cameron, J.C., Wilson, S.C., Bernstein, S.L., and Kerfeld, C.A. (2013). Biogenesis of a bacterial organelle: the carboxysome assembly pathway. Cell *155*, 1131–1140.

Cot, S.S.-W., So, A.K.-C., and Espie, G.S. (2008). A multiprotein bicarbonate dehydration complex essential to carboxysome function in cyanobacteria. J. Bacteriol. *190*, 936–945.

Dou, Z. (2009). Functional characterization and assembly studies of carboxysomes in *Halothiobacillus neapolitanus*. 1–176.

English, R.S., Jin, S., and Shively, J.M. (1995). Use of electroporation to generate a *Thiobacillus neapolitanus* carboxysome mutant. Appl. Environ. Microbiol. *61*, 3256–3260.

Espie, G.S., and Kimber, M.S. (2011). Carboxysomes: cyanobacterial RubisCO comes in small packages. Photosyn. Res. *109*, 7–20.

Flower, A.M., and McHenry, C.S. (1990). The gamma subunit of DNA polymerase III holoenzyme of *Escherichia coli* is produced by ribosomal frameshifting. Proc Natl Acad Sci USA *87*, 3713–3717.

Gay, P., Le Coq, D., Steinmetz, M., Berkelman, T., and Kado, C.I. (1985). Positive selection procedure for entrapment of insertion sequence elements in gram-negative bacteria. J. Bacteriol. *164*, 918–921.

Gonzales, A.D., Light, Y.K., and Zhang, Z. (2005). Proteomic analysis of the CO_2-concentrating mechanism in the open-ocean cyanobacterium *Synechococcus* WH8102. Canadian Journal of Botany *83*, 735–745.

Iancu, C.V., Ding, H.J., Morris, D.M., Dias, D.P., Gonzales, A.D., Martino, A., and Jensen, G.J. (2007). The structure of isolated *Synechococcus* strain WH8102 carboxysomes as revealed by electron cryotomography. J. Mol. Biol. *372*, 764–773.

Iancu, C.V., Morris, D.M., Dou, Z., Heinhorst, S., Cannon, G.C., and Jensen, G.J. (2010). Organization, structure, and assembly of alpha-carboxysomes determined by electron cryotomography of intact cells. J. Mol. Biol. *396*, 105–117.

Jiang, H., Wang, S., Huang, Y., He, X., Cui, H., Zhu, X., and Zheng, Y. (2015). Phase transition of spindle-associated protein regulate spindle apparatus assembly. Cell *163*, 108–122.

Lea, S., Hernéndez, J., Blakemore, W., Brocchi, E., and Curry, S. (1994). The structure and antigenicity of a type C foot-and-mouth disease virus. Structure 123–139.

Li, P., Banjade, S., Cheng, H., Kim, S., Chen, B., Guo, L., Llaguno, M., Hollingsworth, J.V., King, D.S., Banani, S.F., et al. (2012). Phase transitions in the assembly of multivalent signalling proteins. Nature *483*, 336–340.

Long, B.M., Badger, M.R., Whitney, S.M., and Price, G.D. (2007). Analysis of carboxysomes from *Synechococcus* PCC7942 reveals multiple RuBisCO complexes with carboxysomal proteins CcmM and CcaA. J. Biol. Chem. *282*, 29323–29335.

Long, B.M., Tucker, L., Badger, M.R., and Price, G.D. (2010). Functional cyanobacterial beta-carboxysomes have an absolute requirement for both long and short forms of the CcmM protein. Plant Physiol. *153*, 285–293.

Menon, B.B., Dou, Z., Heinhorst, S., Shively, J.M., and Cannon, G.C. (2008). *Halothiobacillus neapolitanus* carboxysomes sequester heterologous and chimeric RuBisCO Species. PLoS ONE *3*, e3570.

Molliex, A., Temirov, J., Lee, J., Coughlin, M., Kanagaraj, A.P., Kim, H.J., Mittag, T., and Taylor, J.P. (2015). Phase separation by low complexity domains promotes stress granule assembly and drives pathological fibrillization. Cell *163*, 123–133.

Mouzakis, K.D., Lang, A.L., Vander Meulen, K.A., Easterday, P.D., and Butcher, S.E. (2013). HIV-1 frameshift efficiency is primarily determined by the stability of base pairs positioned at the mRNA entrance channel of the ribosome. Nucleic Acids Res. *41*, 1901–1913.

Rae, B.D., Long, B.M., Badger, M.R., and Price, G.D. (2013). Functions, compositions, and evolution of the two types of carboxysomes: polyhedral microcompartments that facilitate CO_2 fixation in cyanobacteria and some proteobacteria. Microbiol. Mol. Biol. Rev. *77*, 357–379.

Reuter, J.S., and Mathews, D.H. (2010). RNAstructure: software for RNA secondary structure prediction and analysis. BMC Bioinformatics *11*.

Roberts, E.W., Cai, F., Kerfeld, C.A., Cannon, G.C., and Heinhorst, S. (2012). Isolation and characterization of the *Prochlorococcus* carboxysome reveal the presence of the novel shell protein CsoS1D. J. Bacteriol. *194*, 787–795.

Santer, M., and Vishniac, W. (1957). The thiobacilli. Bacteriol Rev *21*, 195–213.

Theis, C., Reeder, J., and Giegerich, R. (2008). KnotInFrame: prediction of -1 ribosomal frameshift events. Nucleic Acids Res. *36*, 6013–6020.

Tompa, P. (2002). Intrinsically unstructured proteins. Trends Biochem. Sci. *27*, 527–533.

Tsuchihashi, Z., and Kornberg, A. (1990). Translational frameshifting generates the gamma subunit of DNA polymerase III holoenzyme. Proc Natl Acad Sci USA *87*, 2516–2520.

Xu, J., Hendrix, R.W., and Duda, R.L. (2014). Chaperone-protein interactions that mediate assembly of the bacteriophage lambda tail to the correct length. J. Mol. Biol. *426*, 1004–1018.

Xue, B., Blocquel, D., Habchi, J., Uversky, A.V., Kurgan, L., Uversky, V.N., and Longhi, S. (2014). Structural disorder in viral proteins. Chem. Rev. *114*, 6880–6911.

Yan, S., Wen, J.-D., Bustamante, C., and Tinoco, I. (2015). Ribosome excursions during mRNA translocation mediate broad branching of frameshift pathways. Cell *160*, 870–881.

Chapter 3 CsoS2 Forms Protein-Protein Interaction with RuBisCO

3.1 Introduction

In Chapter 2, we have shown that *csoS2* gene expresses two protein isoforms via a cotranslational mechanism known as -1 programmed ribosomal frameshifting. Taking advantage of this fact, we were able to show CsoS2B was critical for carboxysome formation whereas CsoS2A was deemed dispensable. Despite the evidence of the biological significance, the molecular mechanism whereby CsoS2 facilitates the carboxysome assembly has yet to be elucidated. In this chapter, we turn our attention to a major protein-protein interaction mediated by CsoS2 as well as how CsoS2's unique characteristics, including its disorderedness and repetitiveness, contribute to its function.

Recent advances in bioinformatics and genome sequencing have led to the prediction of 23 types of bacterial microcompartments in at least 23 bacterial phyla (Axen et al., 2014). In contrast to the rate of BMC discovery, we are only beginning to understand the principles that govern their assembly (Cameron et al., 2013). Thus far, short metabolic pathways have been reconstructed inside few of catabolic BMCs such as Pdu (Fan and Bobik, 2011; Fan et al., 2012; Lawrence et al., 2014), Eut (Choudhary et al., 2012; Held et al., 2016), and the unnamed BMC from *Haliangium ochraceum* (Lassila et al., 2014). However, the design of these custom nanoreactors is still far from being plug-and-play and has not taken the full advantage of their exceptional modularity and self-assembling properties (Choudhary et al., 2012; Held et al., 2016; Lawrence et al., 2014). To gain an ability to fine-tune these systems, the deep understanding of their properties and assembly strategy are essential. In addition, the "perfect" nanoreactor for a certain application may ultimately be a hybrid that borrow different elements from variety of BMCs (Cai et al., 2015b; Kerfeld and Erbilgin, 2015). Therefore, the knowledge of different principles employed by nature in putting together BMCs will grant us greater latitude in designing and optimizing nanoreactors.

3.1.1 Assembly Mechanisms of α- and β- Carboxysomes

Major advances in the understanding of BMC assembly have come from studies on carboxysomes, especially the β-lineage. Past biochemical studies have revealed that the assembly of β-carboxysomes is primarily driven by protein-protein interactions mediated by two proteins: CcmM and CcmN (Kinney et al., 2012; Long et al., 2007; 2010). CcmM contains two domain: the N-terminal γ-carbonic anhydrase (γ-CA) domain and three RuBisCO small subunit (RbcS)-like repeats (Peña et al., 2010). To initiate the assembly, CcmM organizes RuBisCO into a large aggregate called the procarboxysome. Specifically, the three RbcS-like repeats can replace the true small subunits in multiple RuBisCO hexadecamers, effectively crosslinking RuBisCO into an expansive protein particle. Two isoforms of CcmM, the 35-kDa (CcmM35) and 58-kDa (CcmM58) perform slightly different functions (Long et al., 2011). CcmM35

possesses only the RbcS-like repeats domain, and thus its sole function is to hold RuBisCO molecules together in an aggregate. In contrast, CcmM58 contains both the γ-CA and RbcS-like domain. As the γ-CA domain can interact with CcmN, which in turn interacts with the carboxysome shell, CcmM58's main function is presumably to link the procarboxysome to the shell (Cameron et al., 2013).

Unlike β-carboxysomes, the assembly pathway of α-carboxysomes remains largely unknown. The importance of CcmM and CcmN in the β-carboxysome formation raises the possibility that the α-lineage may require some form of scaffold as well. Among the proteins encoded in the cso operon, those that have already been characterized are known to perform distinct functions from CcmM and CcmN, leaving CsoS2 as the only remaining candidate. While CsoS2 bears no sequence similarity to neither CcmM nor CcmN, its characteristics suggest a similar role the carboxysome assembly. First, deletion of CsoS2 in *H. neapolitanus* completely abolished carboxysome formation, indicating its potentially critical role in the process (Cai et al., 2015a). Moreover, based on pKa, CsoS2 is predicted to possess an opposite charge to other carboxysomal proteins at the physiological pH (Heinhorst et al., 2006), and thus CsoS2 may recruit all other proteins via electrostatic interaction. Most importantly, CsoS2 is predicted to be an intrinsically-disordered protein with multiple repeats. The significance of this particular characteristic in the context of macromolecular assembly will be discussed in details in the next section.

Even though CsoS2 may be functionally analogous to CcmM/CcmN, there is accumulating evidence that the assembly pathways of α- and β- carboxysomes are distinct. β-carboxysomes have been shown to follow a "RuBisCO-first" pathway, in which RuBisCO aggregation precedes all other steps (Chen et al., 2013) (Figure 3-1). Indeed, CcmM-mediated RuBisCO nucleation was shown to be critical to the assembly, and mutants that lost ability to form a procarboxysome were also not capable of carboxysome biogenesis (Cameron et al., 2013). An exception is that transient expression of CcmM and shell proteins in *Nicotiana benthamiana* was able to produce carboxysome-like inclusions without RuBisCO, although the size and shape were starkly different from regular carboxysomes (Lin et al., 2014). In contrast, α-carboxysome assembly does not appear to exhibit the same level of RuBisCO-dependency. Menon *et al*. discovered that the CbbLS-knockout *H. neapolitanus* mutant still formed hollow carboxysomes with relatively unchanged morphology (Menon et al., 2008). A similar result was observed in our group when the genes encoding carboxysome shell proteins from *H. neapolitanus* were expressed together in *E. coli*, although the resulting particles were irregularly shaped. In addition, cryo-EM study of *H. neapolitanus* carboxysomes unveiled partially-formed shell fragments lined with a thin layer of RuBisCO particles (Iancu et al., 2010). These findings suggest a possibility that the assembly mechanism of α-carboxysomes follows a distinct “shell-first” paradigm, where the shell formation initiates RuBisCO nucleation, followed by simultaneous shell expansion and RuBisCO in-filling until the assembly is complete (Figure 3-2)(Rae et al., 2013).

Considering that this assembly mechanism appears to allow carboxysomes to form in the absence of RuBisCO, it is unclear how α-carboxysomes retain the geometric integrity without their RuBisCO "filling." One possibility is that, unlike CcmM/N, CsoS2 may act as a molecular ruler (Marshall, 2004) or contain built-in instruction on how the shell should assemble. Regardless of the exact mechanism, it is becoming clear that α- and β-carboxysomes follow different assembly pathways, and the insights gained from studying one lineage may not directly apply to the other.

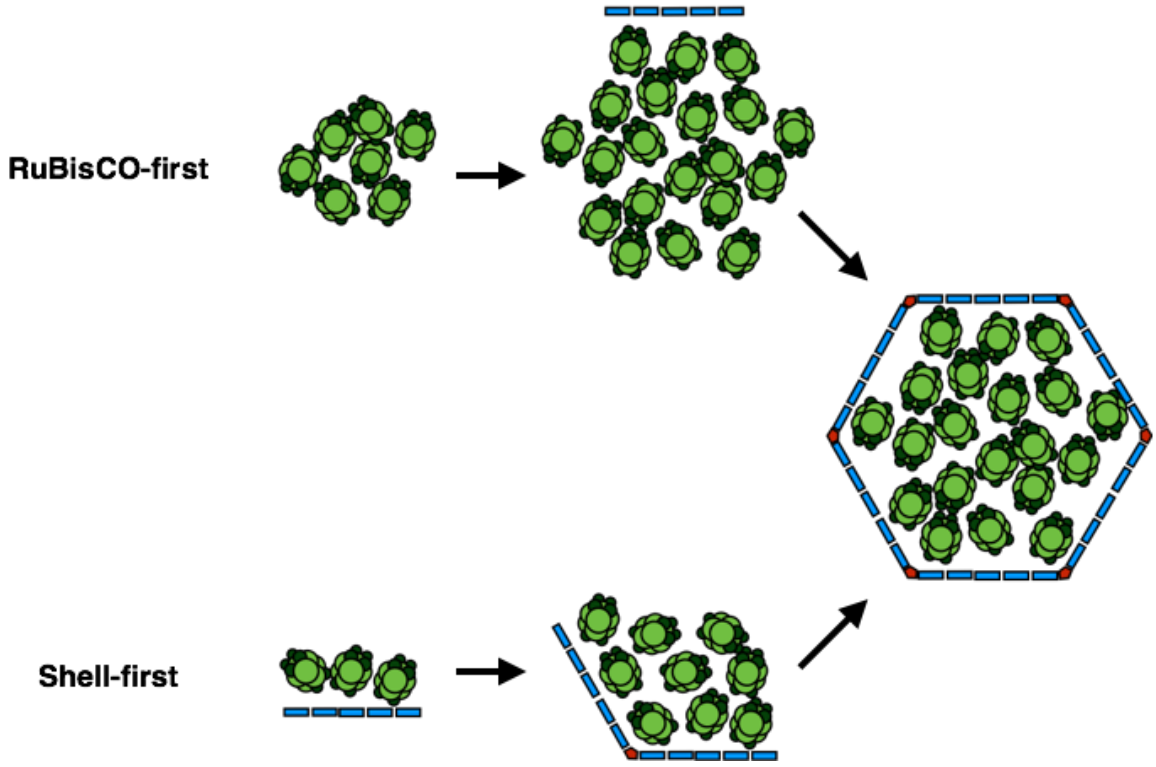

Figure 3-1: Schematic depicting two possible assembly models for carboxysomes.
Only shell proteins and RuBisCO are shown. RuBisCO-first assembly, found in β-carboxysomes, starts with nucleation of RuBisCO into the procarboxysome, followed by the encapsulation by the shell. Shell-first assembly begins with the formation of a shell facet from shell protomers. The nascent shell facet then seeds RuBisCO nucleation and additional shell may be recruited to the growing particle.

3.1.2 CsoS2 as an Intrinsically-Disordered Protein

One of the most important clues regarding the potential function of CsoS2 comes from the fact that this protein belongs to a family called intrinsically-disordered proteins (IDPs). IDPs are a widespread class of proteins, characterized by their high conformational flexibility and the lack of stable tertiary structures under native physiological conditions. In lieu of overall well-defined structures, IDPs exist as dynamic ensemble of conformational states (Tompa, 2012; Uversky and Dunker, 2010). In many cases, IDPs possess large solvent-exposed surface area and high ratio of charged to hydrophobic residues, which allow them to remain unfolded

(Uversky et al., 2000). However, with exception to a small number of proteins, the majority of IDPs are not completely unstructured. They can contain a folded domain fused to a long unstructured region, and multiple secondary structure elements may even be present in the disordered region (Tompa, 2012; Weatheritt and Gibson, 2012). IDPs been found to perform variety of functions that demand high flexibility and expanded conformation (Tompa, 2012). In particular, IDPs are exceptional at binding other proteins and unite them into a large assembly in a short amount of time (Burke et al., 2015; Ekman et al., 2006; Jiang et al., 2015; Li et al., 2012; Nott et al., 2015). These unique characteristics allow IDPs to play complementary roles to globular proteins in many biological processes, such as signal transduction, DNA condensation, transcriptional regulation, and macromolecular assembly (Tompa, 2012).

While IDPs are highly diverse, they share a number of common elements in their protein-protein interactions. They often possess short amino-acid segments that act as binding hotspots, known as MoRFs (Molecular Recognition Features) (Oldfield et al., 2005). These binding units can adopt secondary structures such as α-helix ("α-MoRF") and β-sheet ("β-MoRF"), or they may have an irregular structure ("i-MoRF") (Vacic et al., 2007). In many cases, binding sites on an IDP are unstructured and become folded as they bind to the target and fit more snugly to the partner's binding site (Sugase et al., 2007). Because folding does not happen until binding partners are encountered, unbounded IDPs exhibit large radius of gyration and can reach farther to capture their targets, analogous to "fly-casting." In addition to the large spatial coverage, a simulation showed that the fly-casting behavior helps minimize the number of encounters between two binding partners before the interaction is finished, thus allowing for fast association (Huang and Liu, 2009).

In contrast to its high specificity, IDP-mediated binding is generally of low affinity. Weak binding, and in turn high off-rate, is beneficial for processes that favor transient interaction such as signal transduction. To achieve higher affinity, the weak individual binding units must cooperate via multivalent binding. Many IDPs have amino-acid repeats that serve as duplicate binding sites (Ekman et al., 2006), linked by unstructured regions. The presence of many binding sites on the same polymer increases their effective concentration and decrease the entropy loss associated with the multivalent interaction (Karush, 1976). In certain cases, multiple molecules of IDP may self-associate via hydrophobic motifs in order to increase the density of binding sites (Barbar and Nyarko, 2015). Because of multivalency and inherent self-association propensity, IDPs and their binding partners can rapidly crosslink into a micron-sized macromolecular assembly, sometimes via liquid-liquid phase transition (Lin et al., 2015; Nott et al., 2015). For example, the RNA-binding protein called Fused in Sarcoma (FUS) uses [S/G]Y[S/G] motifs to self-associate and the RNA-binding domain to bind RNA, turning into large hydrogels known as RNA granules in a switch-like manner (Lin et al., 2015). For the reasons described, IDPs are often found to play a central role in cellular processes

where hundreds or thousands of building blocks must be promptly assembled in response to a signal (Smock and Gierasch, 2009).

CsoS2 has several structural features that normally facilitate IDP's binding function. According to the prediction by PONDR-FIT algorithm (Xue et al., 2010), CsoS2 is an IDP with alternating short segments of structured and disordered regions (Figure 3-2A). On the primary sequence level, CsoS2 has several amino-acid repeats divided by long interstitial regions (Figure 3-2B,C). These repeats can be categorized into 3 domains based on the sequence similarity: N-terminal domain (NTD), Middle Region (MR), and C-terminal domain (CTD). NTD consists of four repeats, each with roughly 15-aa per repeat and is significantly shorter than MR and CTD repeats. The four repeats in NTD are highly similar to each other and enriched in charged amino acids such as lysine, arginine, aspartate, and glutamate. These charged residues may mediate electrostatic interaction upon CsoS2's encounter with its binding partner. The MR region contains 6 repeats, each with ~60 aa, except for the 3rd repeat which appears to be missing 12 amino acids found at the end of other repeats. These repeats bear almost no resemblance to NTD in terms of composition and length. While MR repeats are not as similar to each other as in NTD, certain residues are conserved across all the repeats. Most notable is the [V/I][S/T]G motifs that appear in relatively the same location in all repeats. In fact, these motifs are even present in CTD, whose repeats are poorly conserved in comparison to those in NTD and MR. Similar small motifs in some IDPs were reported to be essential for hydrophobic interaction and self-association, including FG-repeats in FG-nucleoporins (NTRs) (Bayliss et al., 2000; lde and Kehlenbach, 2010) and [S/G]Y[S/G] motifs in FUS previously described (Burke et al., 2015). Therefore, if all the repeats in CsoS2 indeed function as binding units, CsoS2 may be able to rapidly drive the carboxysome formation via extensive multivalent interaction with other building blocks.

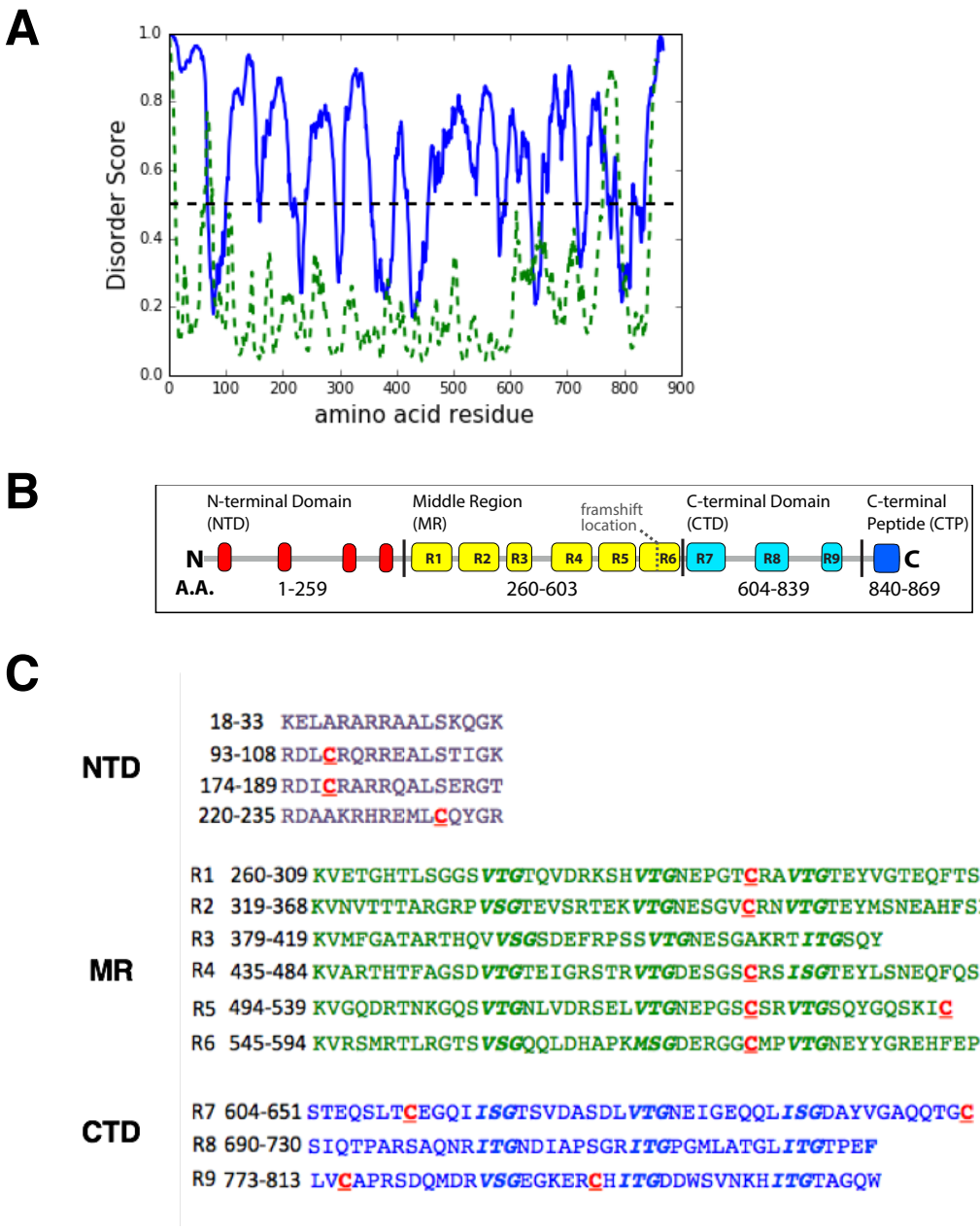

Figure 3-2: CsoS2 as a repetitive intrinsically-disordered protein
(A) PONDR-FIT analysis of CsoS2 showing disorder score of each residue (blue solid line). The same analysis performed on first 869 aa of Cas9 protein from *Streptococcus pyogenes* is included for comparison. (B) Schematic of the domains and repeats in CsoS2.(C) Amino acid sequences of the repeats. Interstitial regions between repeats are omitted. Conserved [V/I][S/T]G motifs are italicized. All cysteine residues are displayed in red and underlined.

Even though BMCs characterized so far do not involve IDP-mediated interaction network, a close example exists in algae. Pyrenoids are shell-less aggregates of RuBisCO responsible for CO_2 fixation, thus functionally analogous to carboxysomes (Badger et al., 1998). It was recently discovered that in the model green algae

Chlamydomonas reinhardtii, an IDP called Essential Pyrenoid Component 1 (EPYC1) may play an important role in pyrenoid morphogenesis (Mackinder et al., 2016). EPYC1 was shown to interact with RuBisCO, and *epyc1* mutant formed defective pyrenoids, resulting in compromised photosynthesis at low CO_2. Interestingly, EPYC1 was found to be a repetitive protein, containing four repeats of roughly 60 amino acids each. These findings led to the putative models that EPYC1 and RuBisCO may be enmeshed together as part of a codependent network, or EPYC1 may form a scaffold where RuBisCO can bind. Although EPYC1 shares no sequence similarity to CsoS2, its intrinsic disordered-ness and amino-acid repeats may hint at a general strategy employed by biological systems to achieve intracellular multi-protein aggregates.

3.1.3 Protein-protein Interaction Involving CsoS2

Formation of bacterial microcompartments from individual building blocks relies on extensive network of protein-protein interactions (Cameron et al., 2013; Fan et al., 2010; Kerfeld and Erbilgin, 2015). Because the actual identity of the cargo proteins matters much less to the assembly outcome than the binding elements that participate in protein-protein interactions, BMC assembly is a highly modular process that tolerates cargo replacement as long as the interaction network is maintained. Therefore, understanding how CsoS2 participates in the complex web of interactions inside a BMC is a requisite to solving the puzzle of carboxysome assembly and utilizing the compartment for engineering applications.

Little is known about the types of protein-protein interactions that CsoS2 can participate in, and most of the evidence was derived from experiments that, while informative, did not directly elucidate the molecular mechanism that imparts CsoS2's function. Past two-hybrid studies investigated pairwise protein-protein interactions between carboxysomal proteins (Gonzales et al., 2005; Williams, 2006); however, due to the false positive rate of two hybrid asssays (Serebriiskii and Golemis, 2001), further investigations using complementary techniques are needed. In addition, two-hybrid assays do not provide insights into the possible mode of interactions of the CsoS2. Quantitative methods that offer more than binary validation are needed in order to reach a comprehensive understanding of how CsoS2 functions.

Being an IDP gives CsoS2 intractable qualities in addition to the previously described functional advantages. Flexibility and lack of tertiary structure make IDP a prime target for proteolysis, while high surface exposure and the presence of self-association domains increase self-aggregation propensity. In addition, if CsoS2 does multivalently interact with its binding partners, the product may unavoidably be in the form of large aggregates. Such tendency to aggregate is problematic for measurements and data analysis by commonly-used quantitative techniques such as Isothermal Calorimetry (ITC) (Velazquez-Campoy et al., 2004) and Microscale Thermophoresis (MST) (Jerabek-Willemsen et al., 2014), which are optimized for studying 1:1 binding stoichiometry and can be complicated by light scattering

caused by the presence of aggregates. These difficulties are possibly the main reason why protein-protein interaction studies on CsoS2, especially quantitative ones, have been scarce relative to other carboxysomal proteins.

3.1.4 Objective

The goal of this work is to elucidate the interaction between CsoS2 and RuBisCO, the main enzymatic component of the carboxysome. We also used this opportunity to experiment with different strategies that may alleviate obstacles in studying an IDP like CsoS2. First, to decrease the degree of multivalency and possibly self-aggregation, CsoS2 was dissected into three domains – NTD, MR, and CTD. Each domain was assayed individually and compared to the result from using the full CsoS2. This strategy may also provide insights into the function of each domain as an added benefit. To validate the results from previous two-hybrid studies, we used *in vivo* pull-down experiments, where tagged bait and untagged prey proteins were coexpressed, and copurification of the prey with the bait was taken to indicate the interaction. This method was expected to allow CsoS2 or its domains to interact with the putative binding partner in a more biologically-relevant environment before they are extensively degraded during the purification. In addition, this strategy would not require fusion of bait and prey proteins with bulky two-hybrid reporters, which should minimize the false positive results as well as disruption of folding and oligomerization due to inclusion of the reporters. In order to complement the qualitative *in vivo* pull-down assay, a quantitatative surface-based technique called Bio-Layer Interferometry (BLI) (Shah and Duncan, 2014) was employed. In BLI assays, CsoS2 variants would be tethered to a probe surface and allowed to interact with RuBisCO in solution. We expected that by using appropriate density of CsoS2 variants on the surface, each molecule should localize sufficiently far from the others and would not be able to partake in self-aggregation. Finally, to visualize the appearance of assembly intermediates from CsoS2 and RuBisCO, we employed transmission electron microscopy (TEM) and fluorescent microscopy to investigate different combinations of CsoS2 variants and RuBisCO. In this process, we discovered a possibility that CsoS2 by itself may undergo liquid-liquid phase transition often involved in the assembly of membraneless organelles in eukaryotes, and we validated this possibility using various biochemical techniques.

3.2 Materials and Methods

3.2.1 In Vivo *Pull-down*

Chemically competent BL21(AI) was cotransformed with pGro7 and the expression plasmid expressing two orthogonally-tagged proteins (StrepII-tag and His-tag). A colony from successful transformation was grown overnight to the stationary phase before the culture was backdiluted 50 times in 100 mL LB media. The culture was grown at 37 ^{o}C until the OD_{600} reached 0.3-0.5, followed by induction with arabinose at the final concentration of 0.1% w/v. After induction, the culture was allowed to grow at 22 ^{o}C, shaken at 175 RPM, for 16-18 h. Cell pellet was harvested by

centrifugation at 3000g for 10 min. If not used immediately, the pellet was stored at -80 °C until needed.

To lyse the cells, the pellet was first resuspended in 4 mL of Strep-Tactin wash buffer [100 mM Tris 150 mM NaCl (pH 7.4)] containing 0.1 mg/mL lysozyme (Sigma-Aldrich), 0.1 mg/mL DNaseI (Worthington), and Roche cOmplete protease inhibitor cocktail. The cell suspension was sonicated at 60% amplitude in 5-second pulses, separated by 10-second cooling periods, for the total of 1 min. The lysate was clarified by centrifugation at 8,000g for 20 min. The clarified lysate was applied to a 1 mL Bio-Rad spin column packed with 100-μL Strep-Tactin Superflow Agarose resin. To avoid overflowing the column, the clarified lysate was applied 500 μL at a time until the entire volume was flown through. The bound resin was washed with 5 column volumes of Strep-Tactin wash buffer and eluted in 100-μL fractions with Strep-Tactin elution buffer (100 mM Tris 150 mM NaCl 2.5 mM desthiobiotin pH 7.4). Western blotting was then performed as previously discussed (see Materials and Methods in Chapter 2)

3.2.2 Bio-Layer Interferometry

Prior to the measurement, sensors were incubated in the assay buffer [50 mM Tris and 150 mM NaCl (pH 7.4)] for 10 min. The steps were programmed as described in Table 4-1. Briefly, the sensor was first incubated in regeneration buffer [30 mM Tris , 90 mM NaCl, 300 mM imidazole, and 0.5% SDS] and washed once in the assay buffer, then loaded with the bait protein (5-20 μg/mL). As a negative control, the sensor not loaded with the bait protein was also included. In certain cases, the loaded sensor was subsequently passivated with 2% bovine serum albumin (BSA, Sigma-Aldrich) to minimize non-specific binding by the prey protein. After the sensor was washed once more to remove residual reagents, the baseline was collected and set. The sensor was exposed to the prey protein to allow association ("association step") and then to the assay buffer to allow the bound prey protein to dissociate ("dissociation step"). The sensor was regenerated with the regeneration buffer to remove the bound bait and prey, and then reused in the next round of the experiment. To account for different surface density between samples, BLI response graphs were normalized to the amplitude of the signal gained during the loading step.

In an experiment where the streptavidin sensor was required, the bait protein was first labeled with NHS-biotin (Pierce) in 1:3 protein: NHS-biotin molar ratio. The rest of the experiment followed a similar protocol as described for the Ni-NTA sensor, except that a sensor was not regenerated after the association-dissociation of the prey and a new sensor was used after each trial.

3.2.3 Protein Labeling with RED-NHS

Labeling buffer, fluorescent dye, a spin desalting column, and a gravity-flow desalting column were obtained from a Protein Labeling Kit RED-NHS (NanoTemper). The protein to be labeled was first exchanged into labeling buffer

using a spin desalting column. Fluorescent dye was dissolved in 100% DMSO and applied to the protein in 3:1 dye:protein molar ratio. The labeling reaction was allowed to proceed for 30 min in the dark at room temperature. To remove the excess dye, the mixture was desalted into protein storage buffer (50 mM Tris and 150 mM NaCl, pH 7.5) with a gravity-flow desalting column.

3.2.4 Recombinant Expression of Proteins Containing His- or StrepII-tag

To construct expression plasmids, Cloning and Golden-Gate assembly were performed using the method described in "Preparation of plasmids" (Materials and Methods, Chapter 2). Expression plasmids were then used to transform chemically competent *E. coli* BL21(AI). For RuBisCO expression, the cells were cotransformed with both the RuBisCO plasmid and the inducible GroEL/ES plasmid (pGro7) to improve folding and solubility. Successfully transformed cells were cultured in 1-l LB media at 37 °C overnight and back-diluted the next day. The culture was allowed to grow until the OD_{600} reached 0.3-0.5 before it was induced with 0.1 % arabinose. Then, the temperature and the shaking speed were reduced to 22 °C and 175 RPM, respectively. At 16 h postinduction, cells were pelleted by centrifugation at 4000g for 20 minutes and stored at -80 °C until needed.

3.2.5 Large-scale Preparation of Clarified lysates

For Ni-NTA purification, cell pellets were first resuspended in 50 mL of buffer A [20 mM Tris and 300 mM NaCl (pH 7.5)] with cOmplete protease inhibitor (Roche), 0.1 mg/mL lysozyme (Sigma-Aldrich), and 0.1 mg/mL DNaseI (Worthington). For Strep-Tactin affinity purification, buffer A was substituted with Strep-Tactin buffer [100 mM Tris and 150 mM NaCl (pH 7.5)]. Cells were passed through an Avestin EmulsiFlex-C3 homogenizer three times. Crude lysates were centrifuged at 15000g for 20 min to remove cell debris. For known aggregation-prone proteins such as CsoS2, the centrifugation speed was reduced to 8000g to minimize protein loss. Clarified lysates were subjected to suitable purification protocols.

3.2.6 Ni-NTA Affinity Purification

Clarified lysates were incubated with appropriate amount of Ni-NTA resin (Life Sciences) and gently stirred for 1 h. Generally, for purification of CsoS2 and variants, 1 mL of resin was used per 1 L of cells. For RuBisCO purification, 3 mL of resin was used per 1 L of cells. After incubation, the mixture was loaded on a gravity flow column (Bio-Rad) and the liquid was let through. The resin was washed with 30 column volumes of Ni-NTA wash buffer [20 mM Tris, 300 mM NaCl, and 30 mM imidazole (pH 7.5)]. The protein was eluted with Ni-NTA elution buffer [20 mM Tris, 300 mM NaCl, and 300 mM imidazole (pH 7.5)] in 3 fractions, each with 1 column volume. Collected fractions were analysed with denaturing SDS-PAGE. Clean and sufficiently concentrated fractions were pooled and applied to an Econopac 10DG desalting column to exchange into the storage buffer [50 mM Tris and 150 mM NaCl (pH 7.5)]. Absorbance at 280 nm was determined using a NanoDrop 2000 UV-Vis spectrophotometry. Protein concentration was calculated using Beer-Lambert law.

Unless immediately used, purified proteins were mixed with 10% glycerol, frozen in liquid nitrogen, and stored at -80 °C

3.2.7 Strep-Tactin Affinity Purification

A gravity-flow column was packed with 1 mL of Strep-Tactin resin (EMD Millipore) per 1 L of cells. Clarified lysates were passed through the column and the resin was washed with 10 column volumes of Strep-Tactin buffer [100 mM Tris and 150 mM NaCl (pH 7.5)]. Proteins were eluted with Strep-Tactin elution buffer [100 mM Tris and 150 mM NaCl (pH 7.5)] in 3 fractions of 1 column volumn each. Eluted proteins were exchanged into the storage buffer [50 mM Tris and 150 mM NaCl (pH 7.5)] and frozen glycerol stocks were prepared in the same manner as described for Ni-NTA purification.

3.2.8 Fluorescent Microscopy

Solution of fluorescent particles (3 μL) was applied between two pieces of No. 1.5 rectangular micro cover glass (24x50 mm, VWR). Imaging was carried out on a Zeiss AXIO Observer.Z1 inverted microscope with 10x phase-contrast air objective (NA of 1.4) or 100x phase-contrast oil objective (NA 1.4). Fluorescent excitation was provided by an Excelitas Technologies X-Cite 120Q fluorescent light source. Images were acquired with a Hamamatsu Photonics ORCA-Flash 4.0 scientific complementary metal oxide semiconductor (sCMOS) camera and Zeiss ZEN 2012 software. Alexa Fluo 647-labeled particles were imaged using a Cy5 filter set (excitation 640 nm; emission 690 nm). For GFP-labeled particles, an enhanced green fluorescent protein (eGFP) filter set was used (excitation 470 nm; emission 525 nm). In most cases, the optimal exposure time was determined using the algorithm provided by ZEN 2012 software and kept consistent throughout the experiment.

3.2.9 Purification of His-tagged CsoS2 from H. neapolitanus *carboxysomes by cation-exchange chromatography*

Purified carboxysomes from wild-type *H. neapolitanus* were pelleted by ultracentrifugation at 105,000g and resuspended in 200 μL of urea denaturing buffer [20 mM sodium phosphate, 20 mM sodium chloride, and 8 M urea (pH 7.5)] with 5 mM dithiothreitol (DTT). Denaturation was allowed to proceed with gentle agitation at room temperature for 1 h. The solution was centrifuged at 20,000g to remove insoluble particles. The entire volume of the solution was injected into a 500-μL sample loop (Bio-Rad). Cation-exchange chromatography was performed with a mono™ S 5/50 column on NGC™ Chromatography System (Bio-Rad) at the flow rate of 1 mL/min. Two denaturing buffers with different NaCl concentration were used for the run: low-salt buffer (same as the urea denaturing buffer described above) and high-salt buffer [20 mM sodium phosphate, 1 M sodium chloride, and 8 M urea (pH 7.5)]. Both buffers contained 1 mM DTT. The elution was done in gradient, with the percentage of high-salt buffer linearly increased from 0 to until 40% in 7 column volumes and then 40% to 100% in 1 column volume. Fractions of

0.5 mL were collected throughout the run and subsequently analyzed by denaturing SDS-PAGE. Fractions containing CsoS2wt were pooled and kept at -80 °C.

3.2.10 Circular Dichroism Spectroscopy

Purified protein was first exchanged into CD buffer [20 mM sodium phosphate and 20 mM sodium sulfate (pH 7.4)] to minimize the background absorbance. From this solution, 300 μL was transferred to a 1-mm quartz cell. The sample containing only CD buffer was included as a negative control. Data were collected on a J-815 circular dichroism spectrometer (JASCO). Spectra were collected in from 190 to 260 nm in 0.5 nm steps with the scanning speed of 20 nm/min and signal averaging for 1 s for each step. Each sample was measured 3 times and the spectra were averaged. BeStSel server (Micsonai et al., 2015) was used to deconvolute the signal inorder to determine the secondary structure distribution .

3.2.11 Fluorescence Recovery After Bleaching (FRAP)

The specimen was prepared as described for fluorescent microscopy. Fluorescent particles were images with Zeiss LSM 710 AxioObserver with 63x oil objective. Two regions with 1 pixel diameter, one within a spherical CsoS2B-sfGFP particle and the other in the background, were chosen for bleaching. The sample was bleached with 488 nm laser at 50% power for 10 iterations and allowed to recover for 1 min. Change in fluorescent intensity over time was plotted using MATLAB.

Table 3-1: An overview of the experimental steps in a BLI experiment

	Step	Solution in 96-well plate	Explanation
1	Regeneration	Regeneration buffer	Strip the His-tagged bait from the Ni-NTA sensor. Not applicable if a streptavidin sensor is used
2	Washing	Assay buffer*	Wash residual protein and regeneration buffer from step 1
3	Loading	Bait His-tagged protein*	Load the His-tagged protein on to the probe surface
4	Washing	Assay buffer	Wash unbound His-tagged protein. If the bait is denatured, it is renatured in this step
5	Blocking	1% BSA	Passivate the probe to reduce non-specific interaction of the prey protein
6	Washing	Assay buffer	Wash unbound His-tagged protein
7	Baseline	Assay buffer	Set the baseline
8	Association	Prey protein	Allow the prey protein to associate to the bait
9	Dissociation	Assay buffer	Allow the prey protein to dissociate from the bait
10	Washing	Assay buffer	Remove unbound proteins

* For experiments involving denatured His-tagged CsoS2, the assay buffer in these steps included 8 M urea

Table 3-2. Expression plasmids used in Chapter 3.
(Destination vectors used in constructing these plasmids are listed in Table 3-3)

Plasmid	Description	Resistance	Reference
pBz12	pET14-based plasmid expressing CbbL (RuBisCO large subunit) and CbbS (RuBisCO small subunit) from *H. neapolitanus*. A hexahistidine tag is included on the C-terminus of CbbL. The intergenic region between *cbbL* and *cbbS* is kept the same as in *H. neapolitanus* genome	Amp	This work
pBz13	pET14-based plasmid expressing CsoS2wt from *H. neapolitanus* with N-terminal hexahistidine tag	Amp	This work
pBz15	pET14-based plasmid expressing CbbL (RuBisCO large subunit) and CbbS (RuBisCO small subunit) from *H. neapolitanus*. A StrepII-affinity tag is included on the C-terminus of CbbL. The intergenic region between *cbbL* and *cbbS* is kept the same as in *H. neapolitanus* genome	Amp	This work
pGro7	Arabinose-inducible expression plasmid for chaperonin GroEL/ES. Purchased from Takara Bio	Cm	N/A
pBz14	pET14-based plasmid expressing C-terminally StrepII-tagged CbbL, untagged CbbS, and N-terminally His-tagged CsoS2wt. Intergenic regions between *cbbL* and *cbbS*, and between *cbbS* and *csoS2*, are preserved. This construct is used in in vivo pull-down experiments desbribed in Results	Amp	This work
pBz64	pET14-based plasmid expressing C-terminally StrepII-tagged CbbL, untagged CbbS, and N-terminally His-tagged CsoS2A. CsoS2B is abolished by removing the coding sequence beyond the segment that encodes CsoS2A. Intergenic regions between *cbbL* and *cbbS*, and between *cbbS* and *csoS2*, are preserved. This construct is used in in vivo pull-down experiments desbribed in Results	Amp	This work
pBz65	pET14-based plasmid expressing C-terminally StrepII-tagged CbbL, untagged CbbS, and N-terminally His-tagged CsoS2B. CsoS2A is abolished by including the NoSlip mutations (see Chapter 2). Intergenic regions between *cbbL* and *cbbS*, and between *cbbS* and *csoS2*, are preserved. This construct is used in in vivo pull-down experiments desbribed in Results	Amp	This work
pBz107	pET14-based plasmid expressing C-terminally StrepII-tagged CbbL, untagged CbbS, and N-terminally His-tagged NTD. Intergenic regions between *cbbL* and *cbbS*, and between *cbbS* and *csoS2*, are preserved. This construct is used in in vivo pull-down experiments desbribed in Results	Amp	This work
pBz104	pET14-based plasmid expressing C-terminally StrepII-tagged CbbL, untagged CbbS, and N-terminally His-tagged MR. Intergenic regions between *cbbL* and *cbbS*, and between *cbbS* and *csoS2*, are preserved. This construct is used in in vivo pull-down experiments desbribed in Results	Amp	This work
pBz97	pET14-based plasmid expressing C-terminally StrepII-tagged CbbL, untagged CbbS, and N-terminally His-tagged CTD. Intergenic regions between *cbbL* and *cbbS*, and between *cbbS* and *csoS2*, are preserved. This construct is used in in vivo pull-down experiments desbribed in Results	Amp	This work
pBz35-1	pET14-based plasmid expressing CsoS2A with C-terminus fused to sfGFP. A hexahistidine tag and a Strep-II tag are present on the N- and C-termini of the fusion protein, respectively	Amp	This work
pBz35-2	pET14-based plasmid expressing CsoS2A with C-terminus fused to sfGFP. A hexahistidine tag and a Strep-II tag are present on the N- and C-termini of the fusion protein, respectively	Amp	This work
pBz109	pET14-based plasmid expressing NTD with N-terminal hexahistidine tag	Amp	This work
pBz106	pET14-based plasmid expressing MR with N-terminal hexahistidine tag	Amp	This work
pBz110	pET14-based plasmid expressing CTD with N-terminal hexahistidine tag	Amp	This work

Plasmid	Description	Resistance	Reference
pBz124	pHnCB10 with *csoS2* truncated such that only MR and CTD are expressed . This construct expresses carboxysomes without NTD in CsoS2	Cm	This work
pBz58	pHnCB10LC with *csoS2* gene truncated on the 5' end to remove the gene segment that encodes repeat 1 and 2 of the NTD. This construct expresses carboxysomes without repeat 1 and 2 of NTD in CsoS2	Amp	This work
pBz102	pHnCB10LC with a hexahistidine-encoding sequence on the 5'-end of *csoS2* gene. This construct is used in the preparation of denatured His-tagged CsoS2 directly from purified carboxysomes	Amp	This work
pBz123	pET14-based plasmid expressing head-to-tail fusion of two NTDs, separated by a 12-aa linker ("double-NTD"). A His-tag and TEV cleavage sequence are included on the N-terminus of the first NTD, and a Strep-II tag is present on the C-terminus of the second NTD	Amp	This work
pBz140	pET14-based plasmid expressing a dimeric leucine zipper fused to NTD, separated by a 10-aa linker ("Doduo"). A His-tag and TEV cleavage sequence are included on the N-terminus of the leucine zipper	Amp	This work

Table 3-3. Cloning destination vectors used in Chapter 3

Plasmid	Description	Resistance	Reference
pSAV038	pET14-based plasmid modified in Savage Group to include BsaI restriction sites on the 5'- and 3'-ends of the LacZ cassette	Amp	N/A
pSAV039	pET14-based plasmid modified in Savage Group to include BsaI restriction sites on the 5'- and 3'-ends of the LacZ cassette. In addition, the coding sequence for a His-tag and a TEV cleavage sequence is included upstream of the first Golden-Gate recombination site, such that the expressed protein contains a N-terminal His tag and a TEV cleavage sequence by default	Amp	N/A
pBz10	A plasmid modified from pSAV038. The canonical *E. coli* RBS downstream of T7 promoter is deleted to allow for the usage of native *H. neapolitanus* RBS	Amp	This work
pBz22C	A plasmid modified from pSAV038. The coding sequences for a short flexible amino acid linker and C-terminally His-tagged sfGFP are located downstream of the BsaI insertion site, such that the expressed protein is fused with C-terminal sfGFP	Amp	This work
pBz57	A plasmid modified from pHnCB10LC, with *csoS2* ORF replaced with a LacZ cassette.	Amp	This work
pBz115	A plasmid modified from pHnCB10, with *csoS2* ORF replaced with a LacZ cassette.	Cm	This work

Table 3-4. Oligos used in Chapter 3

Oligo name	Description	Product	Restriction site	Destination vector	Sequence
TC61	primes the 5'-end of *cbbL* in the forward direction	pBz12, pBz02s	BsaI	pSAV038	CACACCAGGTCTCAGTCCGCAGTTAAAAAGTATAGTGCTGG
TC25	primes the 3'-end of *cbbL,* in the reverse direction. Also include a coding sequece for C-terminal His-tag	pBz12	BsaI	pSAV038	CACACCAGGTCTCAATCAATGATGATGATGATGGTGACGATTTTGAGTGTCGAGTTTGT

Oligo name	Description	Product	Restriction site	Destination vector	Sequence
TC23	primes the beginning of the intergenic region between *cbbL* and *cbbS* in the forward direction	pBz12, pBz02s	BsaI	pSAV038	CACACCAGGTCTCATGATCCCTCGTACCACACAA
TC27	primes the 3' end of *cbbS* in the reverse direction	pBz12	BsaI	pSAV038	CACACCAGGTCTCACGCTTTAGTTGCCGCGGTAGACC
TC62	Primes the 5'-end of *csoS2* in the forward direction. Also include a coding sequece for N-terminal His-tag	pBz13	BsaI	pSAV038	CACACCAGGTCTCAGTCCCACCATCATCATCATCATAATCCTGCCGACCTGAGCG
TC26	Primes the 3'-end of *csoS2* in the reverse direction.	pBz13	BsaI	pSAV038	CACACCAGGTCTCACGCTTTACTTAATCAACCGCGCG
TC24	Primes the 3'-end of *cbbL*, in the reverse direction. Also include a coding sequece for C-terminal StrepII-tag	pBz02s	BsaI	N/A	CACACCAGGTCTCAATCACTTTTCGAACTGCGGGTGGCTCCAACGATTTTGAGTGTCGAG
TC67	primes the 5'-end of *cbbL* in the forward direction. The plasmid pBz02S is used as the template, as its *cbbLS* already includes the StrepII-tag on the 3'-end of *cbbL*	pBz15	BsaI	pSAV038	CACACCAGGTCTCATCCAACGATTTTGAGTGTCGAGTTT
TC68	primes the 3'-end of *cbbS* in the reverse direction. The plasmid pBz02S is used as the template, as its *cbbLS* already includes the StrepII-tag on the 3'-end of *cbbL*	pBz15	BsaI	pSAV038	CACACCAGGTCTCATGGAGCCACCCGCAGTTCGAAAAGTGATCCCTCGTACCACAC
TC59	Primes the end of *cbbS*-*csoS2* intergenic region in the pBz02S plasmid. Also adds an N-terminal His-tag to *csoS2*. Used with TC60, this will amplify around the pBz02S plasmid	pBz14	BsaI	N/A	CACACCAGGTCTCAGCATGTTTGACCCCATCTT
TC60	Primes the end of *cbbS*-*csoS2* intergenic region in the pBz02S plasmid. Also adds an N-terminal His-tag to *csoS2*.	pBz14	BsaI	N/A	CACACCAGGTCTCAATGCACCATCACCATCACCATCCTTCACAGTCAGGAATGAATC
TC207	With TC206 or TC106, amplifies the first insert for constructing pBz64 or 65, respectively.	pBz64, 65	BsmBI	N/A	CACACCAGGTCTCATAGTAGGTTGAGGCCGTTGAGC
TC206	With TC207, amplifies the first insert for constructing pBz64	pBz64	BsmBI	N/A	CACACCAGGTCTCACATCCTTTGGGGCATGATCTAGCT
TC205	With TC208, amplifies the second insert for constructing pBz64	pBz64	BsmBI	N/A	CACACCAGGTCTCAGATGTACGATGATTAAGTAAAGCGGTTAAGATCC

Oligo name	Description	Product	Restriction site	Destination vector	Sequence
TC208	With TC207 or 107 , amplifies the second insert for constructing pBz64 or 65, respecitvely	pBz64, pBz65	BsmBI	N/A	CACACCAGGTCTCAACTACTGGGCTGCTTCCTAATG
TC106	With TC207, amplifies the first insert for constructing pBz65	pBz65	BsmBI	N/A	CACACCAGGTCTCAAGGTGCATGATCTAGCTGTTGGCCA
TC107	With TC208, amplifies the second insert for constructing pBz65	pBz65	BsmBI	N/A	CACACCAGGTCTCAACCTAAGATGTCCGGTGACGAG
TC53	With TC54, amplifies around pSAV038 vector, leaving out the E. coli ribosomal binding site	pBz10	BamHI	N/A	CACACCAGGATCCGACACCCGCCAACACGGTCTCTAGCGGTTAAGATCC
TC54	See TC53	pBz10	BamHI	N/A	CACACCAGGATCCCAGCTGCATTAATGGTCTCTGGACGGTGACCCTATAGTGAGTCGTA
TC264	With TC261, amplifies the region from the 5'-end of *cbbL* to the end of *cbbS-csoS2* intergenic region using pBz14 as template	pBz107, pBz104, pBz97	BsaI	pBz10	CACCAGGTCTCAGTCCCTCAGCTAGGTGCG
TC261	See TC264	pBz107, pBz104, pBz97	BsaI	pBz10	CACACCAGGTCTCAATGTTTGACCCCATCTTGAAT
TC297	With TC298, amplifies the coding sequence of the NTD of CsoS2 from pHnCB10 plasmid	pBz107	BsaI	pBz10	ACACCAGGTCTCAACATGCACCATCATCATCATCATCCTTCACAGTCAGGAATGAATCCT
TC298	See TC297 and TC299	pBz107, pBz109	BsaI	pBz10	CACACCAGGTCTCACGCTTTAAACCTTTTTTGGCGCTGC
TC276	With TC277, amplifies the coding sequence of the CTD of CsoS2 from pHnCB10 plasmid. This particular insert is for constructing the *in vivo* pull-down construct	pBz97	BsaI	pBz10	CACACCAGGTCTCAACATGCACCATCATCATCATCATACGAGCACCCCAGAGC
TC277	See TC276 and TC300	pBz97, pBz110	BsaI	pBz10 (for pBz97), pSAV038 (for pBz110)	CACACCAGGTCTCACGCTTTACTTAATCAACCGCG
TC295	With TC263, amplifies the coding sequence of the MR of CsoS2 from pHnCB10 plasmid. This particular insert is for constructing the *in vivo* pull-down construct	pBz104	BsaI	pBz10	CACACCAGGTCTCAACATGCACCATCATCATCATCATGCGCCAAAAAAGGTTGAAAC
TC263	See TC295 and TC296	pBz104, pBz106	BsaI	pBz10 (for pBz104), pSAV038 (for pBz106)	CACACCAGGTCTCACGCTTTATTGAGCTTCGGGCTCTG

Oligo name	Description	Product	Restriction site	Destination vector	Sequence
TC118	With TC119, amplifies around the pSAV038. The sticky ends from BsmBI digestion will allow ligation with the insert containing *sfGFP*, amplified with TC120 and 121.	pBz22C	BsmBI	N/A	CACACCACGTCTCAACCG CCGCTTCCGCTACCGCTA GAGACCAGCTTGTCTG
TC119	See TC118	pBz22C	BsmBI	N/A	CACACCACGTCTCAAGCG AAGATCCGGCTGCTAACA AAGC
TC120	With TC121, amplifies the coding sequence for sfGFP from a sfGFP-expression plasmid available in Savage Group (DFS724)	pBz22C	BsmBI	N/A	CACACCACGTCTCACGGT TCTAAAGGTGAAGAACT GTTCACC
TC121	See TC120	pBz22C	BsmBI	N/A	CACCACGTCTCACGCTTT AATGATGATGATGATGG TGTTTGTAGAGCTCATCC ATGCCGT
TC124	With TC125 or 126, amplifies *csoS2A* or *B*, respectively	pBz35-1, pBz35-2	BsaI	pBz22C	CACACGGTCTCAGTCCTG GAGTCATCCACAGTTTGA AAAGAATCCTGCCGACCT
TC125	See TC124	pBz35-1	BsaI	pBz22C	CACACCAGGTCTCACGCT TCGTACATCCTTAGGTGC ATGATCTAGCTGTTGGCC A
TC126	See TC124	pBz35-2	BsaI	pBz22C	CACACGGTCTCACGCTTC CACGAGCACCTCCGGAGT AAG
TC299	With TC298, amplifies the coding sequence of the NTD from pHnCB10 plasmid	pBz109	BsaI	pSAV038	CACACCAGGTCTCAGTCC CACCATCATCATCATCAT CCTTCACAGTCAGGAATG AATCCT
TC296	With TC263, amplifies the coding sequence of the MR of CsoS2 from pHnCB10 plasmid. This particular insert is for constructing the the plasmid for expressing N-terminally His-tagged CTD	pBz106	BsaI	pSAV038	CACACCAGGTCTCAGTCC CACCATCATCATCATCAT GCGCCAAAAAAGGTTGA AAC
TC300	With TC277, amplifies the coding sequence of the CTD of CsoS2 from pHnCB10 plasmid. This particular insert is for constructing the plasmid for expressing N-terminally His-tagged CTD	pBz110	BsaI	pSAV038	CACACCAGGTCTCAGTCC CACCATCATCATCATCAT ACGAGCACCCCAGAGC
TC321	With TC322, amplifies the first insert for constructing pBz115, using pHnCB10 as a template	pBz115	BsmBI	N/A	ACGCCGACTGTCTCGAGT ATCTAGTC
TC322	See TC321	pBz115	BsmBI	N/A	CAGCTGCATTAATGGTCT CTATGTTTGACCCCATCT TGAATAAGTGC

Oligo name	Description	Product	Restriction site	Destination vector	Sequence
TC323	With TC324, amplifies the second insert (LacZ fragment) from pSAV038	pBz115	BsmBI	N/A	TTCAAGATGGGGTCAAA CATAGAGACCATTAATG CAGCTGGCAC
TC324	See TC323	pBz115	BsmBI	N/A	GATCGTTACACTTTACTT AACGCTAGAGACCAGCTT GTCTG
TC325	With TC326, amplifies the third insert for constructing pBz115, using pHnCB10 as a template	pBz115	BsmBI	N/A	AGACAAGCTGGTCTCTAG CGTTAAGTAAAGTGTAA CGATCATGAACACCCGTA A
TC326	See TC325	pBz115	BsmBI	N/A	CGATAAACGAATAGACG CGCAAATCAAT
TC327	With TC328, amplifies the fourth insert for constructing pBz115, using pHnCB10 as a template	pBz115	BsmBI	N/A	CGCATGAACAACATTGA TTTGC
TC328	See TC327	pBz115	BsmBI	N/A	AACGCCAGCAACGCGACT A
TC344	With TC345, amplifies *csoS2wt* and adds an N-terminal His-tag	pBz124	BsaI	pBz115	CACACCAGGTCTCAACAT GGGCACAGCACCTTCCTG CAAG
TC345	See TC344	pBz124	BsaI	pBz115	CACACCAGGTCTCACGCT TCATCCTCGTGCTCCGCC GGAGTA
TC188N	With CN6 (See Table 2-2), amplifies the first fragment for construcing pBz57 from pHnCB10LC	pBz57	BsmBI	N/A	CACACCAGGTCTCTGGAC ATTCCTGACTGTGAAGGC A
TC189N	With CN7 (See Table 2-2), amplifies the second fragment for construcing pBz57 from pHnCB10LC	pBz57	BsmBI	N/A	CACACCAGGTCTCAAGCG TTAAGTAAAGTGTAACG GTATATCATGAACAC
TC190N	With TC82, amplifies the LacZ cassette from pSAV038	pBz57	BsmBI	N/A	CACACCACGTCTCAGTCC AGAGACCATTAATGCAG C
TC82	See TC190N	pBz57	BsmBI	N/A	CACACCACGTCTCACGCT AGAGACCAGCTTGTC
TC192	With TC193N, amplifies the region on *csoS2* from the middle of the inter-repeat region between NTD's repeat 2 and 3 to the end of NTD	pBz58	BsaI	pBz57	CACACCAGGTCTCAGTCC GTTCGACCACAGCAATCA GT
TC193N	See TC192	pBz58	BsaI	pBz57	CACACCAGGTCTCTCGCT TTATCCTCGTGCTCCGCC GGAGT
TC285	With TC291, amplifies the first piece for constructing pBz102, using pHnCB10LC as a template	pBz102	BsaI	N/A	CACACCAGGTCTCAGCAT GTTTGACCCCATCTTG
TC291	See TC285	pBz102	BsaI	N/A	CACACCAGGTCTCACAGC AAAAGGCCAGGAACC
TC286	With TC293, amplifies the second piece for constructing pBz102, using pHnCB10LC as a template	pBz102	BsaI	N/A	CACACCAGGTCTCACGAT CATGAACACCCGTAACAC ACG

Oligo name	Description	Product	Restriction site	Destination vector	Sequence
TC293	See TC286	pBz102	BsaI	N/A	CACACCAGGTCTCAACTCATACTCTTCCTTTTTCAATATTATTGAAGC
TC287	With TC288, amplifies the third piece for constructing pBz102, using pHnCB10 (not pHnCB10LC) as a template	pBz102	BsaI	N/A	CACACCAGGTCTCAATGCACCATCATCATCATCATCCTTCACAGTCAGGAATGAATCC
TC288	See TC288	pBz102	BsaI	N/A	CACACCAGGTCTCAATCGTTACACTTTACTTAATCAACCGCG
TC294	With TC292, amplifies the fourth piece for constructing pBz102, using pHnCB10LC as a template	pBz102	BsaI	N/A	CACACCAGGTCTCAGAGTATTCAACATTTCCGTGTCG
TC292	See TC294	pBz102	BsaI	N/A	CACACCAGGTCTCAGCTGGCCTTTTGCTCACATG
TC336	With TC337, amplifies the first NTD moiety for constructing pBz123, using pBz109 as a template (which already contains the N-terminal His-tag coding sequence)	pBz123	BsaI	pSAV039	CACACCAGGTCTCAGTCCCCTTCACAGTCAGGAATGAATCCTG
TC337	See TC336	pBz123	BsaI	pSAV039	CACACCAGGTCTCACCAGAACCTCCTGATCCACCTTTTGGCGCTGCGTTGCC
TC338	With TC343, amplifies the second NTD moiety for constructing pBz123	pBz123	BsaI	pSAV039	CACACCAGGTCTCACTGGTGGATCTGGTGGTTCACCTTCACAGTCAGGAATGAATCCTG
TC343	See TC338	pBz123	BsaI	pSAV039	ACCAGGTCTCACGCTTTACTTTTCGAACTGCGGGTGGCTCCAAACCTTTTTTGGCGCTGC
TC380	With TC381, amplifies the dimeric variant of leucine zipper coding sequence from the synthesized DNA gBlock (IDT DNA)	pBz140	BsaI	pSAV039	CACACCAGGTCTCAGTCCAGAATGAAACAACTTGAAGACAAGGTTGAAG
TC381	See TC380	pBz140	BsaI	pSAV039	CACACCAGGTCTCTCACCGCGTTCACCAAC
TC384	With TC384, amplifies the NTD coding sequence	pBz140	BsaI	pSAV039	CACACCAGGTCTCACCTTCACAGTCAGGAATGAATCCTGCC
TC385	See TC385	pBz140	BsaI	pSAV039	CACACCAGGTCTCACGCTTTAAACCTTTTTTGGCGCTGC

3.3 Results

3.3.1 Preparing of RuBisCO and CsoS2 for In Vitro *Biochemical Assays*

In vitro experiments require protein samples to be purified in order to remove cellular components that may interfere with the downstream assays. To this end, we prepared recombinant *H. neapolitanus* RuBisCO and CsoS2 in *E. coli* and subsequently purified the proteins via affinity chromatography. ORFs of each protein was amplified from the cso operon in pHnCB10 plasmid and inserted in the expression vector downstream of the T7 promoter. A His-tag coding sequence was also included in the 3'-end of the *cbbL* gene for RuBisCO, and 5'-end of the *csoS2* gene for CsoS2. As *csoS2* still contained unaltered frameshifting elements, the protein product was expected to contain both CsoS2A and B in roughly 1:1 stoichiometry. To distinguish *csoS2*-encoded protein from pure isoforms, it will be referred to as "CsoS2wt." The resulting plasmids were transformed to BL21(AI), which is an *E. coli* expression strain with exceptionally tight repression in absence of the inducer, arabinose. Clarified lysates prepared from the induced BL21(AI) carrying the expression plasmids were purified with Ni-NTA affinity chromatography.

Initially, majority of RuBisCO was lost after the centrifugation to produce clarified lysate, suggesting that most of the protein formed inclusion bodies (Figure 3-3A). A number of RuBisCO homologs were previously found to require chaperonin GroEL/ES for the proper folding and oligomerization (Georgiou and Valax, 1996; Liu et al., 2010; Wheatley et al., 2014). While *E. coli* expresses a basal level of GroEL/ES, the chaperonin may not be present at a sufficient concentration to accommodate the folding of abundantly expressed RuBisCO. Therefore, RuBisCO was subsequently coexpressed with the GroEL/ES-overexpressing plasmid (pGro7, Takara), and this strain produced highly soluble RuBisCO between 4-10 mg/L cells (Figure 3-3B, C). While this work was in progress, it was reported that a protein named acRAF from *H. neapolitanus* may improve RuBisCO solubility even further (Wheatley et al., 2014), but we did not attempt this strategy considering that a sufficient amount of RuBisCO had already been acquired by including GroEL/ES alone.

According to previous studies that attempted to purify recombinant *H. neapolitanus* CsoS2wt, the protein was largely insoluble (Baker et al., 1999; Dou, 2009). As a precaution, therefore, CsoS2 was coexpressed with GroEL/ES, in a case that the poor solubility was due to the absence of chaperones. To our surprise, CsoS2 was expressed solubly with yield of 1 mg/1 L cells (Figure 3-3D, E). The same level of yield and solubility were achieved without GroEL/ES overexpression. This disagreement with the literature possibly arose from the use of a different expression strain and plasmid. SDS-PAGE analysis of the purified CsoS2 revealed two prominent bands corresponding to CsoS2A and B, as well as "ladders" below each band. These laddering patterns were likely the result of protease degradation,

to which IDPs are susceptible. Addition of a strong protease inhibitor cocktail (Roche's cOmplete) did not significantly remedy this problem, suggesting that degradation of CsoS2 might have occurred *in vivo*. Nonetheless, we did not subject CsoS2wt to further purification steps as the protein tended to aggregate significantly and behaved differently after prolonged purification.

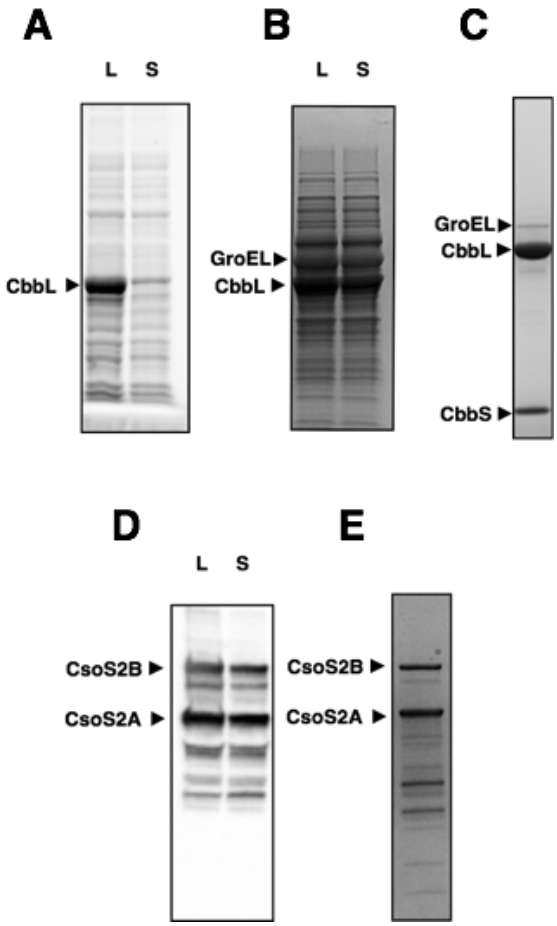

Figure 3-3: Purification of His-tagged RuBisCO and His-tagged CsoS2wt
(A) Coomassie-stained SDS-PAGE gel showing lysate (L) and supernatant (S) from BL21(AI) expressing His-tagged RuBisCO without GroEL/ES overexpression plasmid (pGro7). Lysate and supernatant from BL21(AI) coexpressing RuBisCO and GroEL/ES are shown in (B). (C) Coomassie-stained gel of RuBisCO purified via Ni-NTA affinity purification. (D) Anti-His Western blot of lysate and supernatant from BL21(AI) coexpressing His-CsoS2wt and GroEL/ES. (E) Coomassie-stained gel showing CsoS2wt purified by Ni-NTA affinity purification. Note that all the gel segments shown in (A)-(E) are from different SDS-PAGE gels run at different parameters.

3.3.2 Determining Protein-Protein Interaction between CsoS2wt and RuBisCO

Non-covalent protein-protein interactions are employed to target the cargo protein to lumen by various types of biological compartments, including Pdu (Fan and Bobik, 2011; Fan et al., 2012; Lawrence et al., 2014), Eut (Choudhary et al., 2012;

Held et al., 2016), encapsulins (Sutter et al., 2008), and viral capsids (Minten et al., 2009). In carboxysomes, CsoS2wt may interact with RuBisCO and bring it into an assembling carboxysome. Protein-protein interaction between CsoS2wt and RuBisCO was previously reported based on two-hybrid assays (Gonzales et al., 2005; Williams, 2006), but the inherently high false-positive rate of this method demands additional confirmation.

To this end, we first employed Native Polyacrylamide Gel Electrophoresis (Native PAGE) to assess the interaction between CsoS2wt and RuBisCO. Both proteins were recombinantly expressed in *E. coli* as described. The purified CsoS2wt and RuBisCO were mixed to the final concentration of 3.5 and 0.2 μM, respectively, and incubated at room temperature for 1 h. The samples were applied to a polyacrylamide gel and run at 100 V for 1.5 h. Coomassie staining of this gel revealed that that RuBisCO or CsoS2wt by itself did not migrate significantly down the gel (Figure 3-4A), understandably due to the large size of the RuBisCO hexadecamer (~560 kDa) and the positively-charged nature of CsoS2 at the physiological pH. When present together, however, RuBisCO and CsoS2wt formed extensive smear that coincided with disappearance of the RuBisCO band, suggesting the formation of new high MW species

To visualize the protein complex that gave rise to the smear, we resorted to Negative Staining TEM. The specimen was prepared by incubating RuBisCO and CsoS2 (0.2 and 3.5 μM, respectively) at 4 °C overnight, followed by application to a carbon-coated EM grid. Grids containing only RuBisCO or CsoS2wt were also prepared as controls. By itself, RuBisCO appeared as doughnut-shaped particles with a diameter of roughly 10 nm (Figure 3-4B, top left) while CsoS2wt was not visible under EM (Figure 3-4B, bottom left). In contrast, when RuBisCO and CsoS2wt were present together, expansive paracrystalline aggregates filled with the doughnut-shaped RuBisCO particles, some larger than 1 micron in diameter, were observed (Figure 3-4B, right). These paracrystalline aggregates were reminiscent of the dense interior of α-carboxysomes seen under cryo-electron microscopy (cryo-EM) (Iancu et al., 2010; Schmid et al., 2006), suggesting that they might be a carboxysome assembly intermediate.

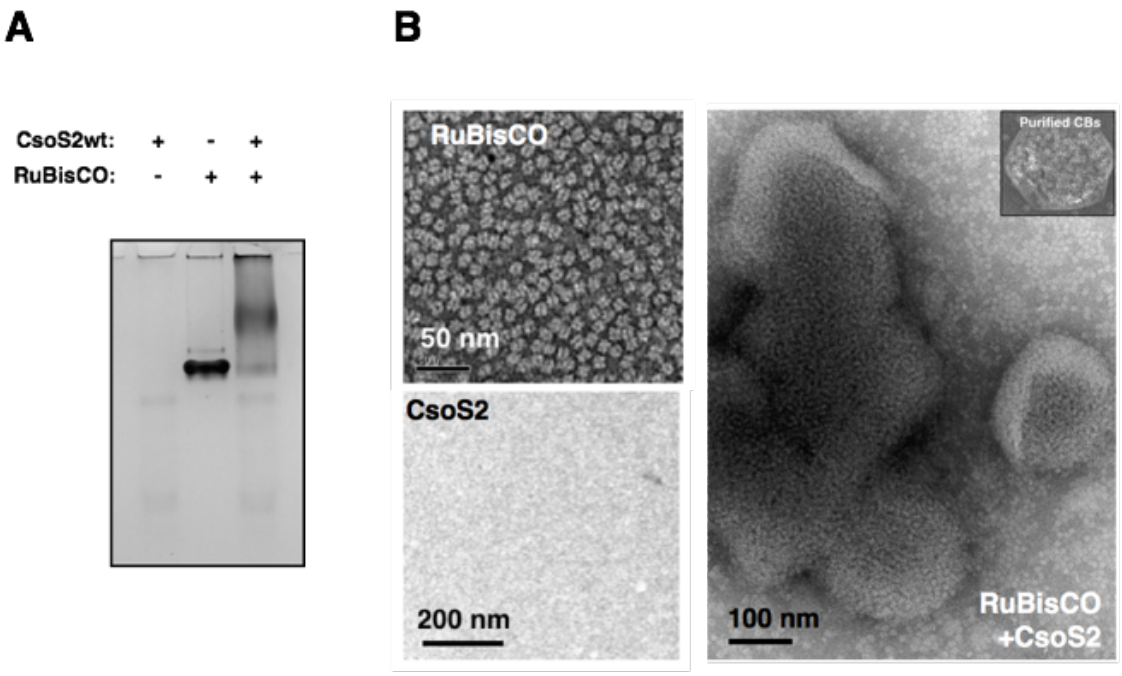

Figure 3-4: Native PAGE and electron microscopic investigation of RuBisCO and CsoS2wt mixture (A) Coomassie-stained Native-PAGE of CsoS2wt, RuBisCO, and CsoS2wt+RuBisCO. (B) Electron micrographs of negative-stained RuBisCO, CsoS2wt, and RuBisCO+CsoS2wt: (Top left) Purified RuBisCO; (Bottom left) Purified CsoS2wt; (Right) RuBisCO and CsoS2wt incubated overnight. The inset shows the image of a purified *H. neapolitanus* carboxysome with highly structured RuBisCO in the interior.

3.3.3 Evaluating Protein-Protein Interaction Between Single CsoS2 Isoforms and RuBisCO

Recombinant CsoS2wt used in the previous section contained both CsoS2A and B in roughly equimolar amount. It was therefore unclear if it was CsoS2A, CsoS2B, or both isoforms that gave rise to the higher-order band in Native PAGE and the aggregates in TEM. Taking advantage of our ability to produce each isoform individually, we recombinantly expressed and purified CsoS2A and CsoS2B using the genetic manipulation demonstrated in Chapter 2 and subjected each isoform to the TEM experiment as described for CsoSwt-RuBisCO. Surprisingly, we were not able to visualize any paracrystalline aggregates (data not shown).

Weak protein complexes could disassemble due to the prolonged and harsh treatment during the preparation of EM grids and negative staining. As an alternative, we employed fluorescent microscopy to visualize the presence of the aggregates. CsoS2A and B labeled with superfolder Green Fluorescent Protein (sfGFP) on the C-terminus were expressed and purified. The mixture between 6 μM CsoS2A-sfGFP or 1.6 μM CsoS2B-sfGFP and 1.7 μM RuBisCO was then visualized for the GFP signal. Without RuBisCO, CsoS2A- or CsoS2B-sfGFP formed a small number of bright fluorescent aggregates. However, when RuBisCO was included with either of the CsoS2 isoforms, a large number of dimly fluorescent particles appeared in addition to a small number of brightly fluorescent aggregates (Figure 3-5B). The

bright aggregates might have resulted from spontaneously aggregated CsoS2A or B, which was either not active in binding to RuBisCO or did not undergo dramatic morphological changes upon binding. On the other hand, the abundant dim aggregates likely resulted from *de novo* aggregation between RuBisCO the soluble form of the CsoS2 proteins. These observations indicate that either isoform of CsoS2 was capable of nucleating RuBisCO, although the resulting aggregates may not be as robust as when two isoforms are present together, causing them to elude the electron microscopic detection.

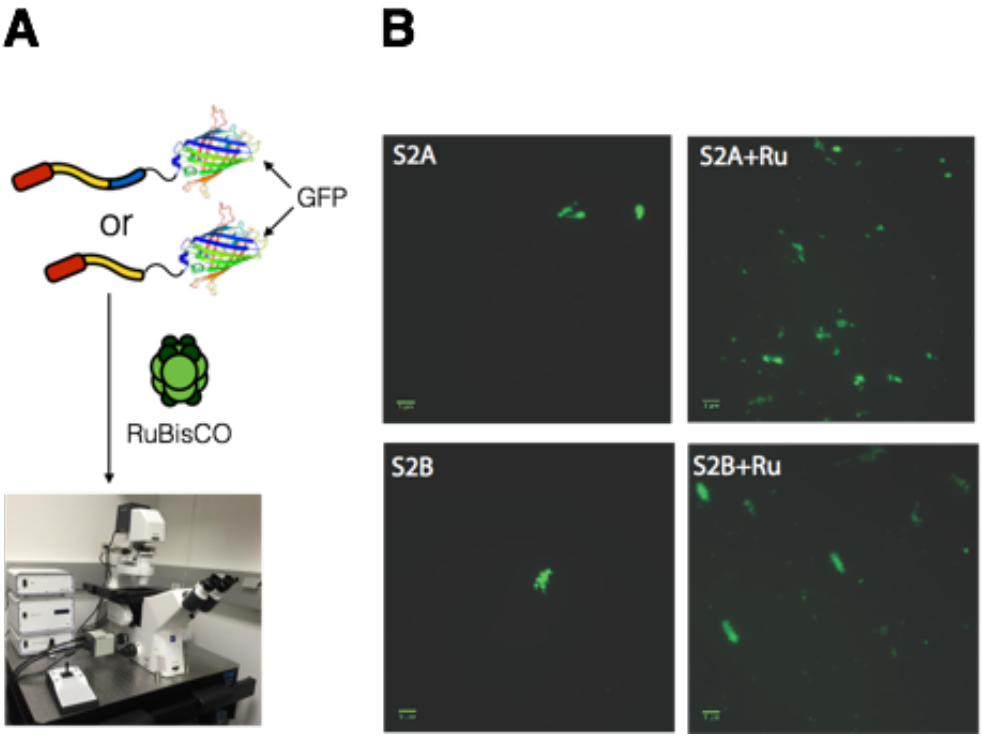

Figure 3-5: Visualization of CsoS2A+RuBisCO and CsoS2B+RuBisCO with fluorescent microscopy
(A) Schematic showing the preparation of the samples. CsoS2A-sfGFP and CsoS2B-sfGFP were purified and incubated with RuBisCO before subjected to fluorescent microscopy. (B) Fluorescent images of CsoS2A-sfGFP, CsoS2B-sfGFP, CsoS2A-sfGFP+RuBisCO, and CsoS2B-sfGFP+RuBisCO (Scale bar = 5 μm)

To verify that the interaction we observed could take place *in vivo*, we co-expressed the His-tagged CsoS2A or B with Strep-tagged RuBisCO on a single plasmid and captured the RuBisCO with Strep-Tactin resin. Copurification of a CsoS2 isoform with Strep-tagged RuBisCO would serve as a strong indication that both proteins associated *in vivo*. Cells co-expressing CsoS2wt and RuBisCO were subjected to the same procedure as a positive control. The eluted proteins were then evaluated with SDS-PAGE and anti-His Western blotting. To our surprise, the eluates of all samples displayed an exaggerated ratio of CsoS2A to CsoS2B (Figure 3-6A); in fact, 1:1 CsoS2A:CsoS2B was observed in the RuBisCO-CsoS2B sample, which should have

contained no CsoS2A. As RuBisCO's preferential binding to CsoS2A might account to the observed phenomenon, Western blotting was performed on the clarified lysates prior to the pull-down and compared to the result from the eluates. No difference in the isoform distribution was observed, suggesting that the presence of RuBisCO *in vivo* somehow altered the frameshifting efficiency. Therefore, while this pull-down experiment confirms the interaction between RuBisCO and CsoS2A, it could not provide a conclusion for CsoS2B.

To complement the result from the *in vivo* pull-down assay, we employed native PAGE to detect the formation of large aggregates between each CsoS2 isoform and RuBisCO. We performed this experiment on a 0.7% agarose gel instead of a PAGE gel, as we discovered around this time that the former provided a much better resolution for large protein complexes (Cai et al., 2015a). In addition, as our set-up for performing agarose gel electrophoresis required the gel to be run horizontally, it was possible to load the proteins in the center of the gel to allow them to run towards both electrodes. This experimental design would be beneficial for studying the positively-charged CsoS2 species, which normally would not run down a vertical gel at the physiological pH . CsoS2A-sfGFP, CsoS2B-sfGFP, and CsoS2wt were mixed with RuBisCO and incubated for 10 min at room temperature. The samples were applied to the agarose gel and run with 100 V for 1 h 15 min. Coomassie staining of the gel revealed that CsoS2wt and RuBisCO formed intense smear and depleted the band corresponding to RuBisCO. Both CsoS2A-sfGFP and CsoS2B-sfGFP displayed a band corresponding to higher-order species, but the mobility shift was musch less dramatic than the CsoS2wt-RuBisCO control and the band corresponding to RuBisCO was not depleted (Figure 3-6B). Taken together with the observation from fluorescent microscopy, this result confirms that CsoS2A and CsoS2B are capable of associating with RuBisCO into higher-order species, albeit to a lesser extent than CsoS2wt.

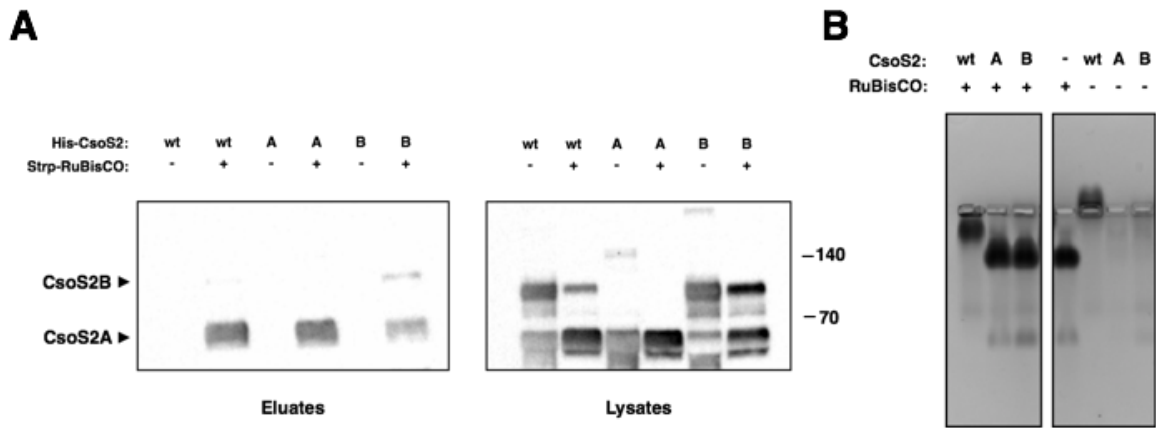

Figure 3-6: Protein-protein interaction between CsoS2 isoforms and RuBisCO validated by *in vivo* pull-down and native agarose electrophoresis
(A) Anti-His Western blot of the the eluates (left) and lysates (right) from the Strep-tactin purification. The presence and absence of each protein was shown above the images. For His-tagged CsoS2, the type of CsoS2 was also indicated: wt = CsoS2wt; A = CsoS2A; B = CsoS2B. (B) Coomassie-stained gel from native agarose electrophoresis. The gel was run horizontally, with the top of the gel facing the anode and bottom the cathode. The type of CsoS2 was indicated on top of the gel image. Note that A and B are CsoS2A-sfGFP and CsoS2B-sfGFP, respectively, while CsoS2wt is not tagged with sfGFP.

3.3.4 Elucidating the Domain in CsoS2 that Interacts with RuBisCO

CsoS2 can be divided into three domains based on the characteristics of their amino acid repeats. NTD contains four ~15-aa repeats that are enriched in positively charged amino acids such as lysine and arginine. MR has six ~60-aa repeats that contain highly conserved [V/I][S/T]G motifs. While repeats in CTD display similar lengths and [V/I][S/T]G motifs to those in MR, they are significantly less conserved. These differences imply that the three domains may perform different functions or interact with different sets of binding partners. We proposed at the end of Chapter 2 (Figure 2-9) that the NTD is available for binding RuBisCO while CTD acts in shell recruitment. According to our finding that both CsoS2A and B bound RuBisCO, the responsible domain was likely NTD or MR.

To this end, we examined which domains could copurify with RuBisCO in the *in vivo* pull-down experiment. The coexpression plasmids were designed and prepared in a similar manner to the previous pull-down experiments, except that the *csoS2* gene was truncated to encode only a single domain. The affinity purification with Strep-Tactin resin was also performed in a similar manner, using samples that contained only the CsoS2 domains as negative controls. Detection with anti-His tag Western blotting revealed that StrepII-tagged RuBisCO was able to enrich NTD and MR, but not CTD (Figure 3-7). This finding verified our assumption that the RuBisCO-binding regions were within the domains shared between CsoS2A and CsoS2B.

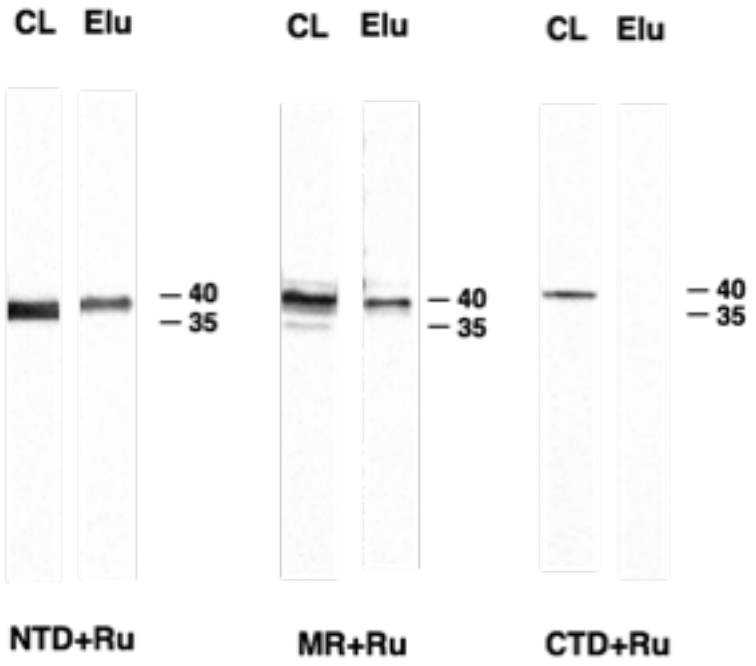

Figure 3-7: Determination of interaction between RuBisCO and single domains from CsoS2 using *in vivo* pull-down.
The results of anti-His Western blot performed on the eluates (EL) from Strep-Tactin affinity purification were shown. Clarified lysates (CL) prior to the purification were also included for comparison.

To additionally validate our pull-down result and gain further insights into the plausible binding mechanism of NTD and MR, we employed a surface-based protein-protein interaction measurement by Bio-Layer Interferometry (BLI). In this technique, the bait protein is first loaded to a chemically-functionalized sensor via the non-covalent interaction between the affinity tag and the functional group on the surface. The BLI sensor is able to quantify the change in surface thickness and the surface loading density, therefore, is correlated with the amplitude of the signal gained during this step (Figure 3-8A). With the bait on the surface, the sensor is then exposed to the putative prey in solution ("association step"). Interaction between bait and prey would result in the deposit of prey on the probe surface, leading to gradual increase in thickness and in turn the BLI signal. Then, the probe is immersed in a buffer solution to allow the prey to dissociate, causing the signal to decrease ("dissociation step"). The rate of signal increase and decrease during the association and dissociation steps can be used to infer the kinetics of the protein-protein interaction. A representative plot of BLI signal during association and dissociation steps is shown in Figure 3-8B

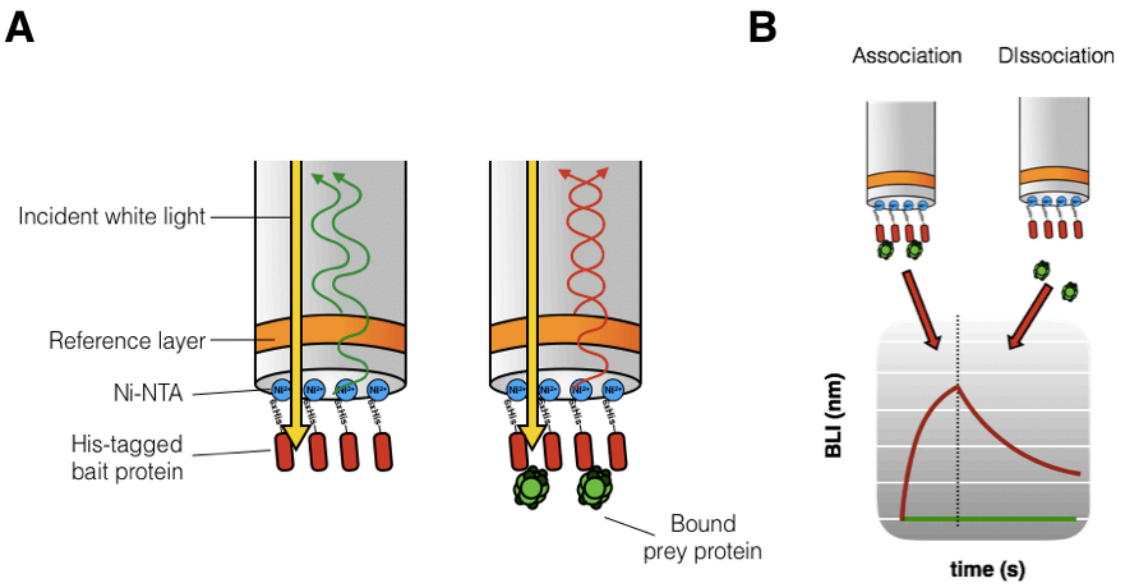

Figure 3-8: Principles of Bio-Layer Interferometry (BLI)
(A) A schematic explaining how BLI detects protein-protein interaction. The His-tagged bait protein (red rounded rectangle) was first tethered to the Ni-NTA functionalized probe surface. Upon the application of the incident white light, the reference layer (orange strip) and the bait-covered surface can reflect the light. The product from the interference of the light waves that are reflected from two different layers is recorded by the detector as the baseline. As more and more prey protein binds to the probe, the thickness of the tip changes, leading to a different interference product of the reflected light waves. The difference from the baseline is then converted to the BLI signal to indicate the extent of bait-prey interaction. (B) The appearance of the BLI signal during association and dissociation steps. The red line represents the signal if the interaction does take place. The flat green line corresponds to the absence of interaction between bait and prey.

To prepare for the BLI experiment, we constructed plasmids expressing N-terminally His-tagged NTD, MR, and CTD individually, and recombinantly expressed the proteins in *E. coli*. Ni-NTA affinity purification yielded satisfactory amount of each domain and products of protease degradation were found to be minimal. Although it was important to include CsoS2wt as positive control, we were concerned about its extensive aggregation and degradation that could significantly affect the quantitative assays. Therefore, we developed a new strategy to prevent the proteolysis and self-aggregation of CsoS2wt, taking advantage of the following facts: (i) CsoS2 is inaccessible to proteases when inside a carboxysome; (ii) a denaturing condition should not affect the function of CsoS2, considering that it is an IDP that lacks significant structure to begin with; and (iii) renaturation of denatured CsoS2 after it is loaded to the BLI sensor is unlikely to lead to self-aggregation if the surface coverage is low enough. Our workflow is outlined in Figure 3-10A. Briefly, we prepared a pHnCB10 variant that expresses a His-tag on the N-terminus of CsoS2wt and confirmed by Negative-Staining TEM that the expressed carboxysomes did not differ from the wild-type (Figure 3-9B). The purified carboxysomes were denatured with 8 M urea and subjected to cation exchange chromatography, which was expected to capture a positively-charged species like CsoS2wt. We used the

gradient elution strategy with two different buffers: low-salt buffer [20 mM sodium phosphate, 20 mM sodium chloride, 8 M urea, pH 7.5] and high-salt buffer [20 mM sodium phosphate, 1 M sodium chloride, 8 M urea, pH 7.5]. Percentage of the high-salt buffer was linearly increased from 0-40% over 8 column volumes and then from 40-100% in 1 column volume. His-CsoS2wt was eluted when the percentage of high-salt buffer reached 30%, corresponding to 312 mM NaCl (Figure 3-9C). SDS-PAGE analysis shows that our purification scheme successfully enriched CsoS2wt (fraction 12 and 13, Figure 3-9D) with negligible amounts of contaminating RuBisCO and shell protein. Notably, products from CsoS2 proteolysis was not observed, confirming that CsoS2 was likely located inside the carboxysome and thus was inaccessible to proteases.

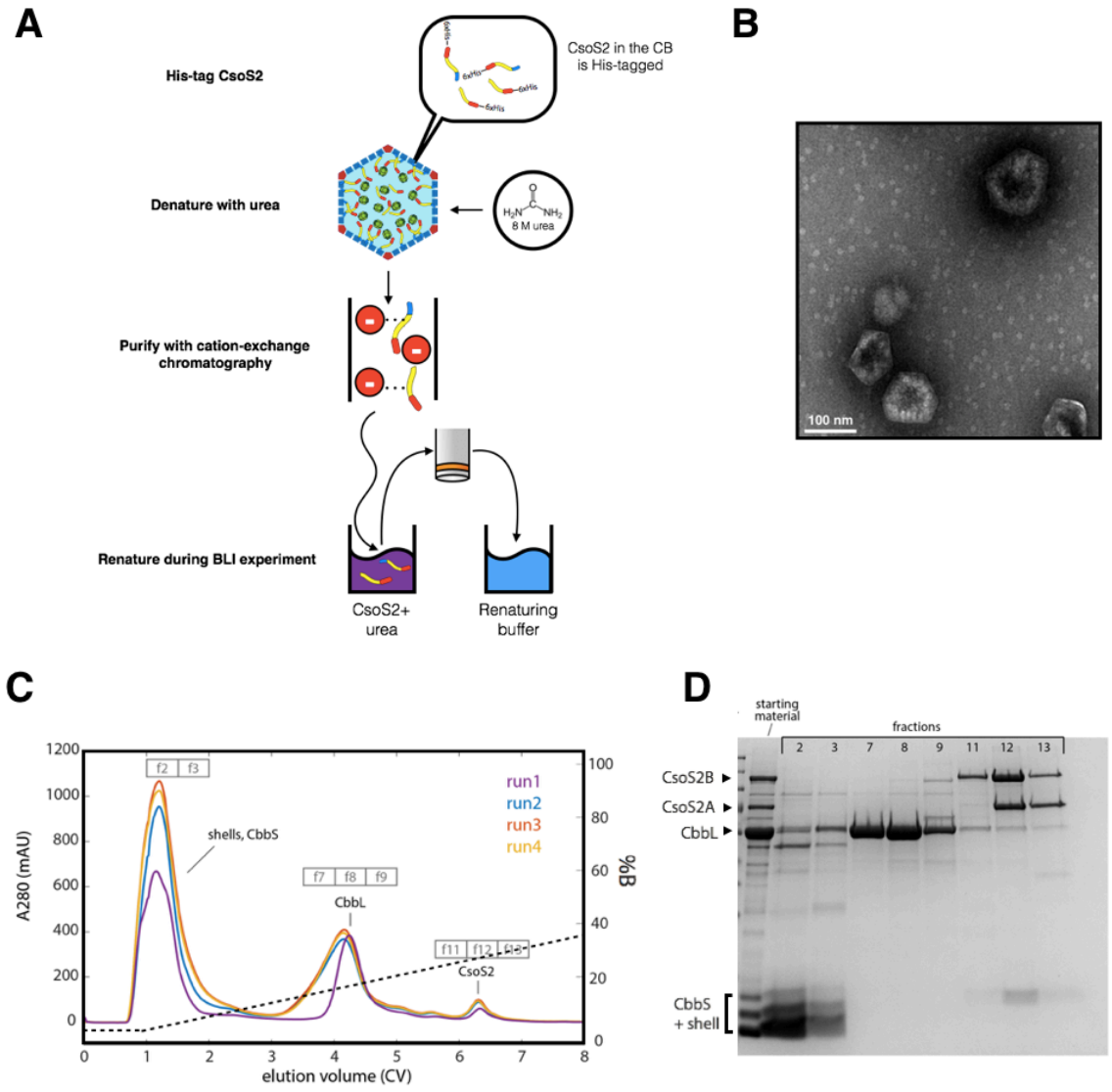

Figure 3-9: Purification of His-tagged CsoS2wt from carboxysomes using denaturing cation-exchange chromatography
(A) Cartoon depicting the purification workflow. Carboxysomes with His-tagged CsoS2wt are purified and denatured with 8 M urea. CsoS2wt is separated from all other carboxysomal proteins using cation-exchange chromatography. The acquired protein is kept denatured until it is tethered to the probe and exposed to renaturing buffer. (B) Electron micrograph of purified carboxysomes containing His-tagged CsoS2. (C) Chromatogram from the cation-exchange step. The graphs from four separate runs are overlaid. Three major peaks are annotated with the corresponding fraction numbers and expected proteins. The percentage of the high-salt buffer (% B) over the course of the purification is shown by the dashed line. Note that %B was allowed to increase to 100% after the time period shown in the graph but no other peaks were observed. (D) SDS-PAGE of the fractions that correspond to the major peaks in (C).

BLI measurements were carried out with His-tagged NTD, MR, CTD, or full CsoS2wt as bait and StrepII-tagged RuBisCO as prey. Steps involved in this experiment were outlined in Table 3-1 (See Materials and Methods). Briefly, each bait protein was loaded on to the Ni-NTA sensor, which was then washed with buffer to remove excess bait. For CsoS2wt, the loading was performed under a denaturing condition and the protein was allowed to renature in the subsequent wash step. The loaded sensor was passivated with 2% bovine serum albumin (BSA) to minimize non-specific binding of prey protein to Ni-NTA. To measure the association and dissociation BLI signal, the loaded probe was exposed to RuBisCO in solution and transferred to buffer, respectively. At the beginning of the next experiment cycle, the His-tagged bait was released from the sensor surface by treating with a regeneration buffer containing 300 mM imidazole and 0.05% SDS, followed by a brief wash to remove residual reagents. The now-empty sensor was reloaded with fresh bait protein and ready for the subsequent experimental cycle.

The BLI assay revealed that using CsoS2wt and NTD as bait proteins exhibited significant association signal while MR and CTD yielded no observable increase in signal (Figure 3-10A). Interestingly, NTD appears to yield higher binding signal than CsoS2wt at the same loading amount and RuBisCO concentration. Moreover, the trial using 20 μg/μL NTD and 2 μM RuBisCO gave almost the same level of association signal as when 10 μg/μL NTD was used, suggesting that RuBisCO might have become limiting in this range of NTD concentration. This was not the case for CsoS2wt, whose signal did not become RuBisCO-limited even at 20 μg/μL. Collectively, these results indicate that NTD was as active and potentially more competent at binding RuBisCO than CsoS2wt under our experimental conditions.

Characteristics of the BLI signal from CsoS2wt/NTD indicate the involvement of multivalent interaction in RuBisCO binding. Upon normalizing the graph to the loading density, BLI signal still exhibited dependency to the loading concentration, suggesting a synergistic effect where multiple copies of closely-located CsoS2 or NTD cooperatively bound to RuBisCO. In addition, BLI signal did not return to the baseline at the end of the dissociation step, which indicates the presence of both irreversible and reversible binding elements. Irreversible binding might arise from clusters of CsoS2wt/NTD molecules that could collectively clamp on approaching RuBisCO (Figure 3-10B), while CsoS2wt/NTD in the less dense region of the surface would only be able to achieve loose and reversible binding.

The binding capability of MR remains inconclusive owing to its inconsistent behavior across our interaction assays. While the *in vivo* pull-down experiment (Figure 3-7) appears to suggest RuBisCO-MR interaction, the absence of signal in the BLI assay indicates the opposite (Figure 3-10). This discrepancy could be explained either by the false positive result from the pull-down experiment or the false negative result from the BLI assay. Due to its lack of a well-folded native state, MR may be a somewhat sticky protein capable of non-specific interactions with other abundant proteins in the cell. In this case, overexpression of both RuBisCO and MR

in the same cell could make the *in vivo* pulldown more susceptible to this artifact. For BLI, the false negative could be attributed to improper orientation of MR on the surface; ie. the binding-competent region might have been located close to the probe surface and thus was inaccessible to RuBisCO. To rule out the artifacts inherent to either scenario, the expression of both MR and RuBisCO may need to be down-regulated for an *in vivo* pull-down experiment, and a His-tag on the C-terminus of MR should be additionally used in the next attempt of the BLI assay.

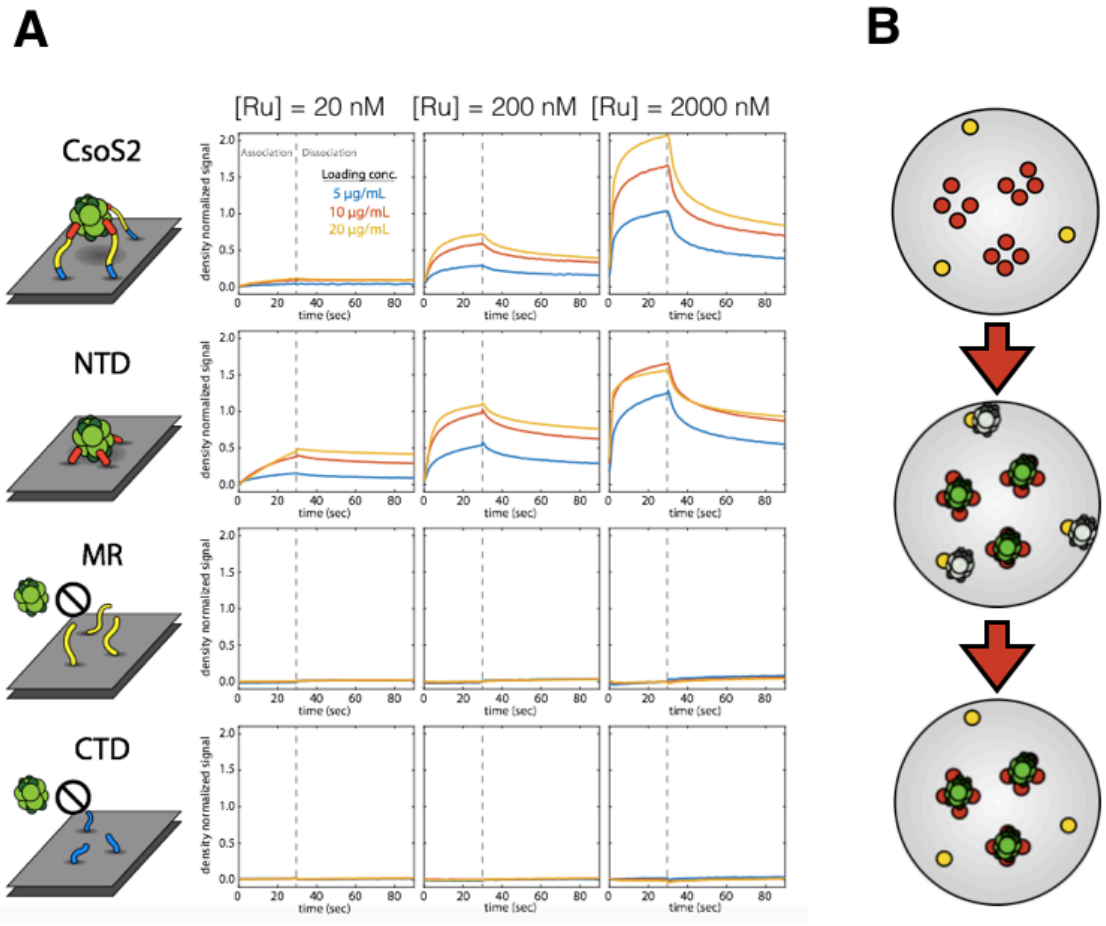

Figure 3-10: BLI measurement of interaction between different CsoS2 domains and RuBisCO
(A) Density-normalized BLI plots when full CsoS2wt, NTD, MR, and CTD were used as bait and RuBisCO as prey. Different rows of the panel correspond to different domains of CsoS2. Different columns correspond to different RuBisCO concentrations (20, 200, and 2000 nM). For each CsoS2 variant, three concentrations were used: 5, 10, and 20 µg/mL. (B) Schematic depicting the irreversible and reversible RuBisCO binding as discussed in text. CsoS2 variants are distributed unevenly on the surface, with some located close to each other in a cluster (red) and the others not close to any neighbors. Upon the association, incoming RuBisCO molecules that land on a cluster (bright green molecules) are held tightly, while those that land near to lone CsoS2 (faintly-colored RuBisCO) are bound weakly. When allowed to dissociate, weakly-bound RuBisCO can come off but those bound by the clusters will still be attached.

3.3.5 Determining the effect of the number of binding sites in NTD-RuBisCO interaction
For protein that uses multiple MoRFs to cooperatively bind its partner, increasing the number of recognition sites may lead to even higher binding affinity (Nyarko et al., 2013). Thus, if NTD indeed has a cooperative binding mode, assembling two NTDs into a single bipartite binding domain should increase the BLI signal per NTD. To this end, two NTD domains were genetically fused head-to-tail, separated by a 12-aa linker (Figure 3-11A, B), and used as the bait protein in a BLI experiment. In

the presence of RuBisCO, this “double-NTD” exhibited BLI signal amplitude roughly two-fold higher than the single NTD when normalized to the density of individual NTD moieties on the surface (Figure 3-11C), suggesting the involvement of cooperative binding.

Despite the positive result from the double-NTD construct, the full effect of the multivalency might have not yet been reached due to the possibility that the optimal binding configuration may prefer NTD to align in parallel. Head-to-tail fusion via a short linker likely prohibits such configuration, leading to compromised binding (Figure 3-11B). To complement the result from using the double-NTD construct, we decided to fuse a leucine-zipper moiety to the N-terminus of NTD. Leucine zipper is an α-helical domain from the GCN4 protein that dimerizes in a parallel orientation. Point mutagenesis of certain hydrophobic residues was reported to alter the oligomeric state from dimer to trimer and tetramer (Harbury et al., 1993), but only the dimeric variant was used in this study. The NTD-leucine zipper fusion, called "Doduo," consisted of His-tagged NTD on the N-terminal side, a 10 aa flexible linker, and a leucine zipper on the opposite end (Figure 3-11D, E). BLI measurement revealed that Doduo bound RuBisCO with approximately 1.3-fold higher BLI amplitude than single NTD during the association (Figure 3-11F), slightly lower than the double-NTD.

Notably, the signal difference between the beginning of the association and the end point of the dissociation (“ΔBLI_{irr}”) was relatively similar between Doduo and single NTD, while the difference between the end points of the association and dissociation (“ΔBLI_{rev}”) was two-fold higher for Doduo. On the other hand, relative to single NTD, double-NTD appears to exhibit almost two-fold ΔBLI_{irr} but similar ΔBLI_{rev}. This discrepancy indicates that the superior binding of Doduo was mostly attributed to reversible interaction, while double-NTD’s binding was dominated by irreversible interaction. In addition, contrary to our initial concern, this would suggest that double-NTD was more capable of strong multivalent binding than Doduo, although the mechanistic reason was not known. Regardless, these findings indicate that the binding of NTD to RuBisCO involves the cooperativity from at least two NTDs from different CsoS2 molecules.

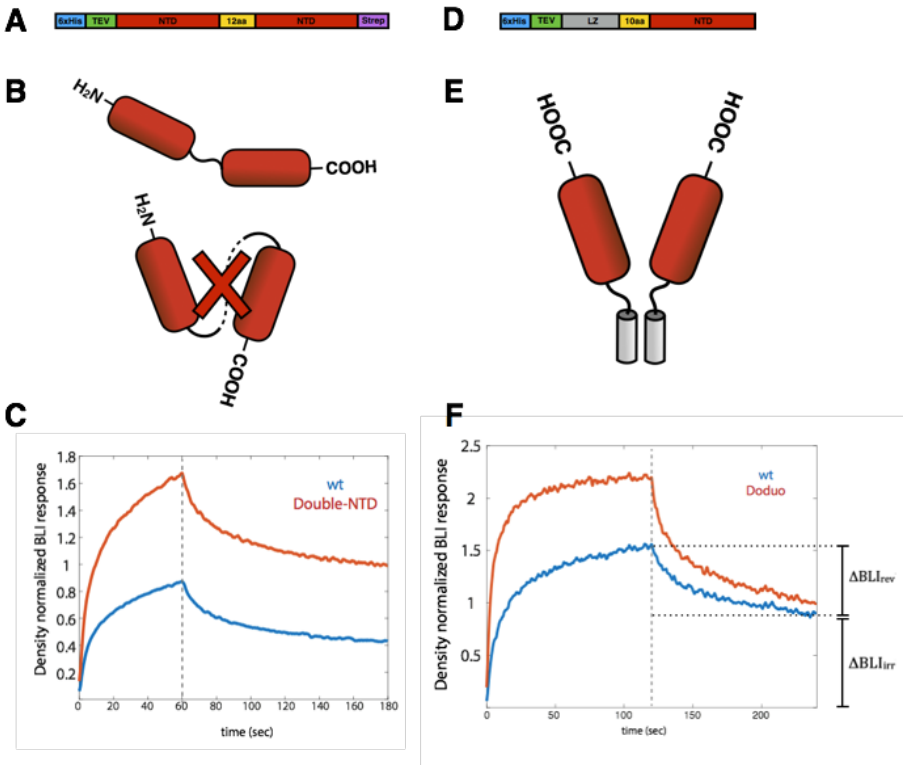

Figure 3-11: BLI measurement of interaction between RuBisCO and dimeric NTD constructs.
(A) Schematic showing the components in double-NTD construct and their arrangement. (B) Cartoon representations of double-NTD in head-to-tail (top) and parallel (bottom) orientations. Parallel orientation is expected to be geometrically forbidden, as discussed in the text. For easy identification of the termini, affinity tags are not shown. (C) Density-normalized BLI response from interaction between double-NTD and RuBisCO. Wild-type NTD was used as a control (D) Schematic showing the components in Doduo construct and their arrangement. Leucine zipper is abbreviated as LZ (E) Cartoon representation of dimerized NTD-LZ. (F) Density-normalized BLI response from interaction between Doduo and RuBisCO. Wild-type NTD was used as a control. Signal differences corresponding to ΔBLI_{irr} and ΔBLI_{rev} discussed in the text are shown on the plot.

Previously, we have shown that CsoS2wt was able to form extensive paracrystalline aggregates with RuBisCO. Since multivalent interaction has been reported to facilitate the formation of large protein assemblies (Burke et al., 2015; Jiang et al., 2015; Li et al., 2012), we set out to determine whether NTD's propensity for multivalent binding might have conferred the observed RuBisCO nucleation. Purified NTD was chemically labeled with RED-NHS (equivalent to Alexa Fluor 647 NHS ester) and incubated with RuBisCO for 10 min at room temperature (Figure 3-12A). Upon visualization by fluorescent microscopy, a large number of red fluorescent aggregates were observed (Figure 3-12B). In contrast, incubation of labeled NTD with BSA or buffer alone produced only negligible amount of small fluorescent aggregates, possibly from the low-level self-aggregation of NTD. This result demonstrates that NTD is capable of multivalent binding and mediating RuBisCO nucleation similarly to the full-length CsoS2.

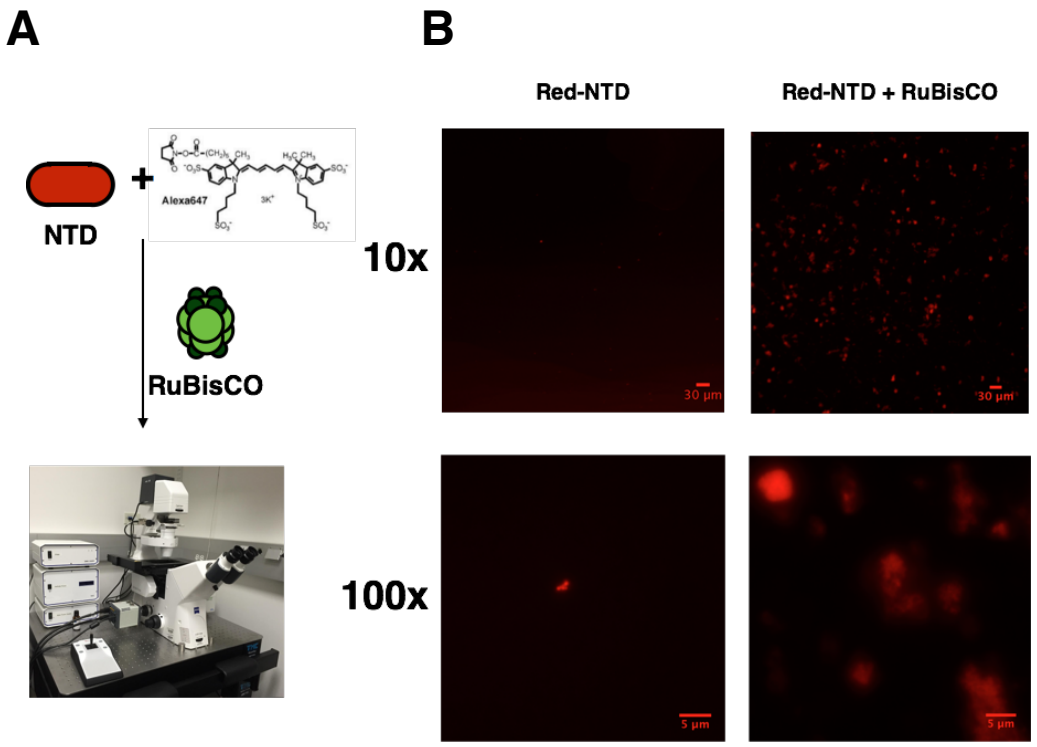

Figure 3-12: Visualization of NTD-RuBisCO aggregates with fluorescent microscopy
(A) Schematic showing the preparation of the sample. Purified His-tagged NTD was labeled with RED-NHS (equivalent to NHS-Alexa 647). The labeled protein (Red-NTD, 4.5 μM) was incubated with RuBisCO (2 μM), before subjected to fluorescent microscopy. (B) Fluorescent images of Red-NTD in the presence and absence of RuBisCO. Images were taken with 10x (top row, scale bar = 30 μm) and 100x magnification (bottom row, scale bar = 5 μm).

3.3.6 Investigating Binding Units within NTD

Despite lacking tertiary structure, an intrinsically-disordered protein often contains multiple MoRFs interspersed along its unstructured backbone (Vacic et al., 2007), analogous to called "beads on a string." The difference in binding competency between NTD and the other two domains may be attributable to the presence of these elements. While we would like to identify the MoRFs in NTD using atomic-resolution structural techniques such as X-ray crystallography, they are not compatible with a highly flexible protein like CsoS2 (Dou, 2009). As a close alternative, we employed bioinformatics prediction combined with lower-resolution structural characterization to gain insights into the structural elements of CsoS2 that may account for its binding ability.

Circular dichroism (CD) spectroscopy is an optical technique that provides low-resolution structural information of proteins (Greenfield, 2006). Taking advantage of the fact that different types of secondary structure exhibit different dichroism, which is the differential absorption of left- and right-handed circularly polarized light, CD spectroscopy can measure the abundance of different secondary structures

in a protein. To prepare NTD, MR, and CD for CD measurement, the previously purified proteins were dialyzed into CD buffer [20 mM sodium phosphate and 20 mM sodium sulfate (pH 7.4)]. This step was performed in order to remove high background absorbance in the UV range due to certain components in the protein storage buffer. The CD measurement revealed that NTD had two local minima, one at 205 nm and the other at 220 nm (Figure 3-13A). MR and CTD, on the other hand, had only one peak centered around 200 nm. As peaks at 200 nm and 220 nm are characteristic of random-coil and α-helical structure, respectively, this result suggests that NTD contained a certain extent of alpha helices while MR and CTD were largely unstructured. Spectra deconvolution using BeStSel server (Micsonai et al., 2015) confirmed that NTD possessed 24.6% of alpha helices while MR and CTD contained only 6.1 and 5.4%, respectively.

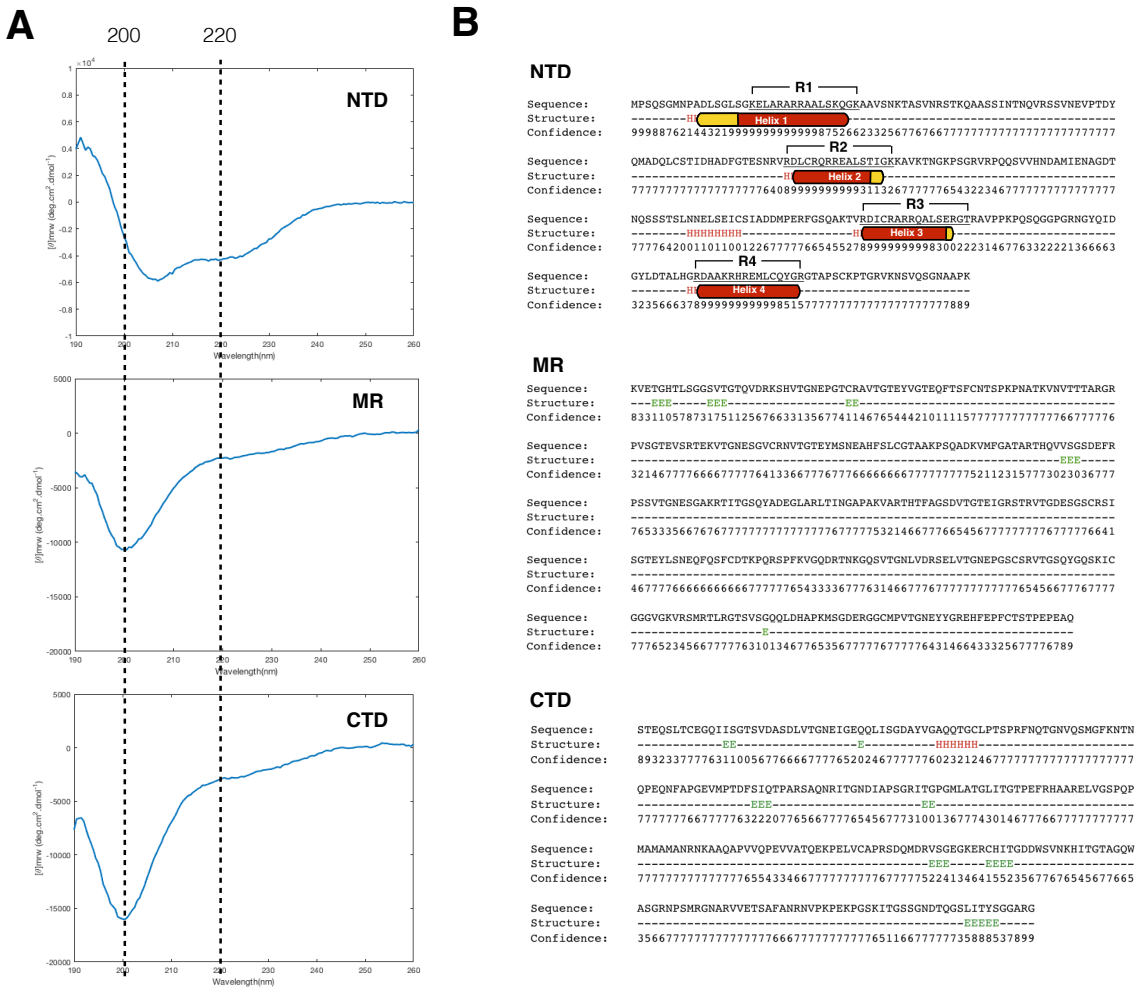

Figure 3-13: Secondary structure of NTD, MR, and CTD:
(A) Circular dichroism spectra of NTD, MR, and CTD, showing mean residue molar ellipticity at the wavelengths between 190 to 260 nm. Two broken lines marked 200 nm and 220 nm. (B) Secondary structure prediction of NTD, MR, and CTD by JPred4 server. Each residue is annotated with the predicted secondary structure (H = α-helix, E = expanded) and the confidence score ranging from 0-9, with higher value indicating higher confidence in prediction. For Helix 1-4 described in the text, each region of consecutive H's is replaced with a cartoon cylinder, with the red color indicating confidence score of at least 5 and yellow indicating the score below 5. Residues corresponding to repeat 1-4 were marked with brackets.

While CD allowed us to estimate the percentage of each secondary structure, it offered no information about the location of these elements. In order to locate α-helical regions in NTD, a protein secondary structure prediction server JPred4 was employed (Drozdetskiy et al., 2015). The algorithm predicted four 14-15 aa alpha helices within NTD with high confidence scores (Helix 1-4, Figure 3-13B). A short helix was also predicted between the Helix 2 and 3, but was likely a false positive considering the low confidence scores. Together, Helix 1-4 accounted for approximately 25% of the total amino acid in NTD, closely matching the percentage

of α-helix derived from CD. Strikingly, the predicted α-helices in NTD significantly overlapped with four NTD repeats, which agrees with the general observation that repeats in IDPs often function as binding domains. On the other hand, predictions on MR and CTD revealed only very short regions of expanded secondary structure (E) and one short alpha helix, all of which were assigned with low confidence scores. Therefore, the JPred4 predictions were in good agreement with the CD result, and collectively, these results suggest that the α-helices in NTD may explain its unique RuBisCO-binding capability.

Many instances of amphipathic helices have been reported to participate in protein-protein interaction via the hydrophobic face (Segrest et al., 1990). To get a clearer picture of the binding topology of NTD α-helices, their primary sequences were plotted in the helical wheel format using an algorithm on the Zidovetzki group's website (University of California – Riverside). Inspection of the helical wheel diagrams showed that all four helices contained one side that was highly populated with charged amino acids such as lysine, arginine, aspartate, and glutamate (Figure 3-14A). In contrast, the opposite face was enriched in hydrophobic residues such as leucine, and methionine, with exception to lysine or histidine sandwiched between residue 1 and 4. Among the repeats, Leu12 and Arg9 were perfectly conserved (Figure 3-14B). Each of positions 2, 3, 6, and 8 was also occupied by amino acids with similar charge property. The polarized distribution of amino acids suggest that NTD helices are amphipathic and may use one of the face in protein binding.

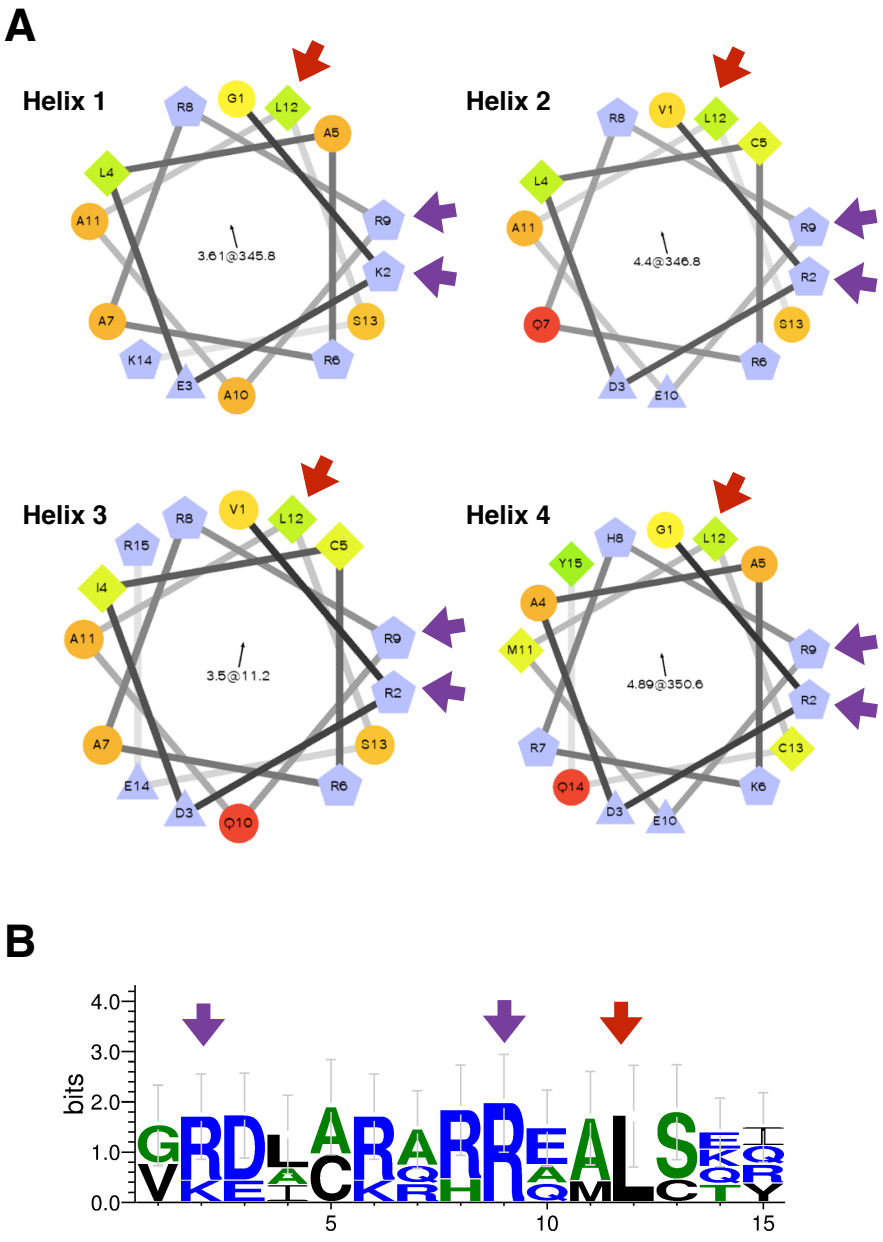

Figure 3-14: Helix 1-4 from NTD are amphipathic
(A) Helical wheel diagramss of helix 1-4 from NTD. The targets for alanine substitution are indicated by arrows: Leu12 (red arrow), Arg/Lys2 and Arg9 (purple arrows). The thin black arrow in the center indicates the direction of the hydrophobic moment of the helix. The magnitude and the angle of the hydrophobic moment are given in the form of X@Y, where X is the magnitude and Y is the angle. (B) WebLogo constructed with the sequences of Helix 1-4.

To investigate the function of the conserved residues on each face of the NTD amphipathic helices, we constructed two mutants of NTD, one with Leu12 on every helix replaced with alanine ("L12A") and the other with both Arg/Lys2 and Arg9

mutated to alanine ("R2A-R9A") (Note that the first residue in each helix is numbered as 1). We chose to mutate these residues because they were were among the most highly conserved across Helix 1-4 and likely involved in different types of intermolecular interaction, with leucine capable of forming hydrophobic interaction and lysine/arginine electrostatic interaction. The His-tagged mutants were expressed in *E. coli* and purified with Ni-NTA chromatography before subjected to BLI with RuBisCO.

For this particular BLI assay, we decided to design the experiment such that we could deconvolve the contributions of reversible and irreversible binding. To this end, in the first trial, we allowed RuBisCO to associate to the loaded NTD variants for 500 s to ensure that the NTD population capable of irreversible binding would have been saturated by the end of the association step. After the dissociation of RuBisCO and a brief wash step, the probes, still with irreversibly bound RuBisCO, were allowed to associate with 20 nM RuBisCO. As all of the irreversible-binding NTDs were expected to have already been occupied by RuBisCO, this round of association should only permit reversible binding. After a dissociation step to allow reversibly binding RuBisCO to come off, we repeated the reversible association-dissociation cycle twice more with 60 nM and 200 nM RuBisCO

In the first long association-dissociation cycle, we observed significant binding signal from the L12A mutant, but the saturation was reached much faster and the maximum signal was half of the wild-type NTD (Figure 3-15). R2-R9 mutant, on the other hand, did not yield an association signal. In the subsequent short association-dissociation cycles, L12A gave relatively the same level of association signal to wild-type NTD. The association signal in these cycles also showed dependence on RuBisCO concentration in both WT and L12A NTD. These findings suggest that the irreversible binding mode of L12A mutant was significantly compromised while the reversible binding was unaffected. In contrast, perturbing R/K2 and R9 led to complete loss of binding capability. This result suggests that either R/K2 or R9 is indispensable in RuBisCO association, possibly by participating directly in the binding or mediating intramolecular interaction between helices. L12, on the other hand, may be important for cooperativity between NTD molecules, potentially by helping align NTDs in the same binding cluster into the correct binding topology.

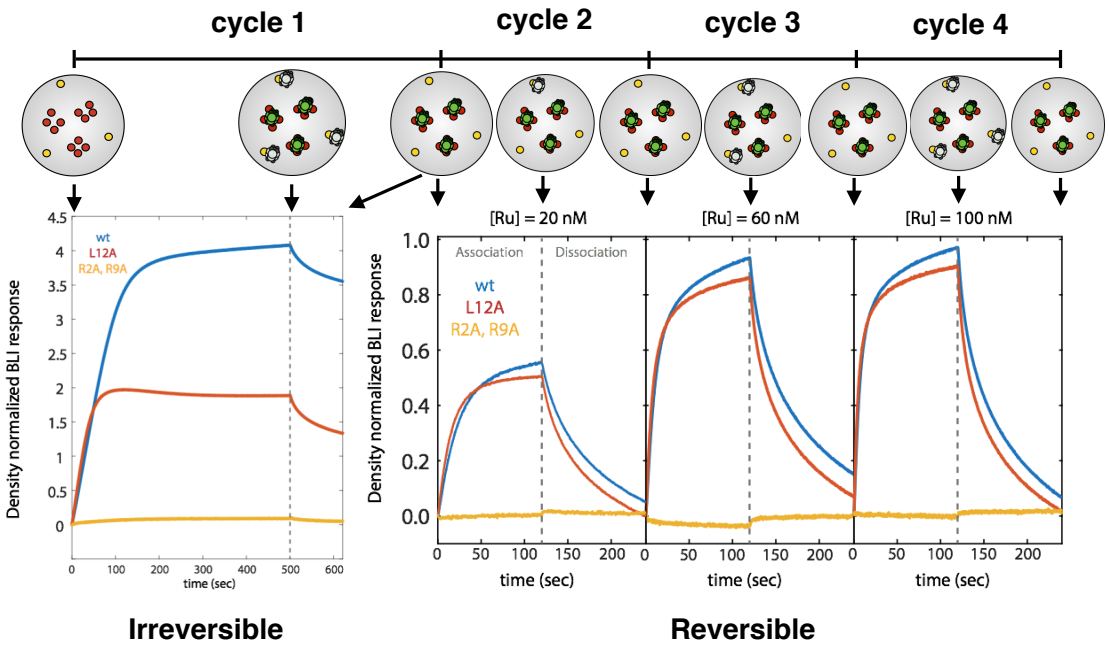

Figure 3-15: BLI measurement of the interaction between NTD mutants and RuBisCO.
The experiment was performed in 4 successive association-dissociation cycles. Association in cycle 1 was allowed to proceed for a long time to saturate the irreversibly-binding NTD clusters with RuBisCO. Cycle 2-4 consist mainly of reversible binding, with RuBisCO concentration increasing after each cycle. Cartoon representation of the expected binding events in each cycle are shown above the BLI plots, with arrows pointing to the corresponding data point in the plots. Symbols were previously defined in Figure 3-10B.

There are a number of ways by which NTD helices can bind RuBisCO, including: (i) each helix associates with one RuBisCO hexadecamer (Figure 3-16A); (ii) four helices wrap around one RuBisCO molecule (Figure 3-16B); and (iii) four helices form a helical bundle that can then interact with RuBisCO (Figure 3-16C). To determine which binding mode NTD may use, we tested the ability of a single helix in binding to RuBisCO. We purchased a synthetic peptide corresponding to the third helix, with Cys5 mutated to alanine to prevent intermolecular disulfide crosslinking. This peptide, called WT-C5A, had the following sequence: VRDIARARRQALSERGT. When a BLI assay was performed with His-tagged RuBisCO as a bait and WT-C5A in solution, no binding signal was observed (data not shown). To ensure that this was indeed a true negative result, we decided switch the bait and prey. WT-C5A was conjugated with NHS-biotin and attached to a streptavidin-functionalized sensor, while His-tagged RuBisCO was in solution. This direction also did not yield a binding signal. Therefore, it is possible that the interaction between a single NTD helix and RuBisCO may be too weak. This scenario, if true, would rule out the model portrayed in Figure 3-16A. As only Helix 3 was tested in our work, future study should examine other individual NTD helices to confirm the finding.

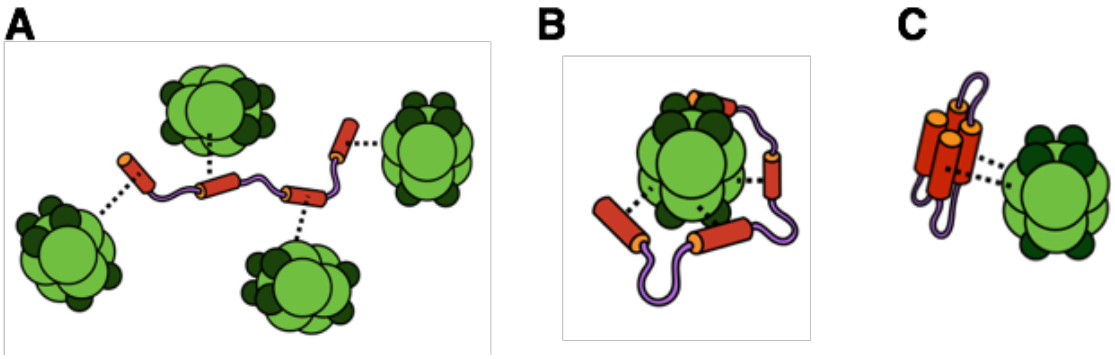

Figure 3-16: Three of the possible binding configurations of NTD helices
(A) One helix binds to one RuBisCO hexadecamer. (B) All four helices chelate one RuBisCO hexadecamer on different locations. (C) Four helices assemble into a four-helical bundle that binds to RuBisCO. Each helix in NTD is symbolized as a red cylinder.

3.3.7 Evaluating the Importance of NTD in Carboxysome Morphogenesis

A previous study demonstrated that carboxysomes in *H. neapolitanus* could be produced in the absence of RuBisCO expression (Menon et al., 2008). Considering that NTD interacts with RuBisCO, the absence of this domain was expected to have no affect on carboxysome formation. To verify this hypothesis, we used an *E. coli*-based carboxysome expression assay described in Chapter 2. We slightly modified the protocol by using the original pHnCB10 plasmid (Table 2-1), as we discovered that pHnCB10 expressed carboxysomes with higher abundance and integrity. To this end, a pHnCB10 variant with NTD removed from *csoS2* gene was constructed (Figure 3-17A) and expressed in *E. coli*, but no carboxysome-like particles were observed under EM. Interestingly, in one of our original attempts to construct the NTD-less pHnCB10 plasmid, we mistakenly deleted only the first and second repeats in NTD as opposed to the whole domain. Surprisingly, expression of this plasmid in *E. coli* yielded particles of comparable sizes to regular carboxysomes, but were highly broken and barely filled with RuBisCO (Figure 3-17B). These findings demonstrated that NTD was crucial to the assembly of carboxysomes and the presence of more than two repeats may be necessary to ensure properly assembled carboxysomes. It should be noted that, while it is tempting to speculate that all four repeats are required for proper carboxysome assembly, a bioinformatic survey of NTD repeats in different CsoS2 homologs showed that the number of repeats can range from three to six per NTD domain (Cai et al., 2015a). Therefore, the number of N-terminal α-helices involved in binding RuBisCO may not necessarily be four in certain CsoS2 homologs.

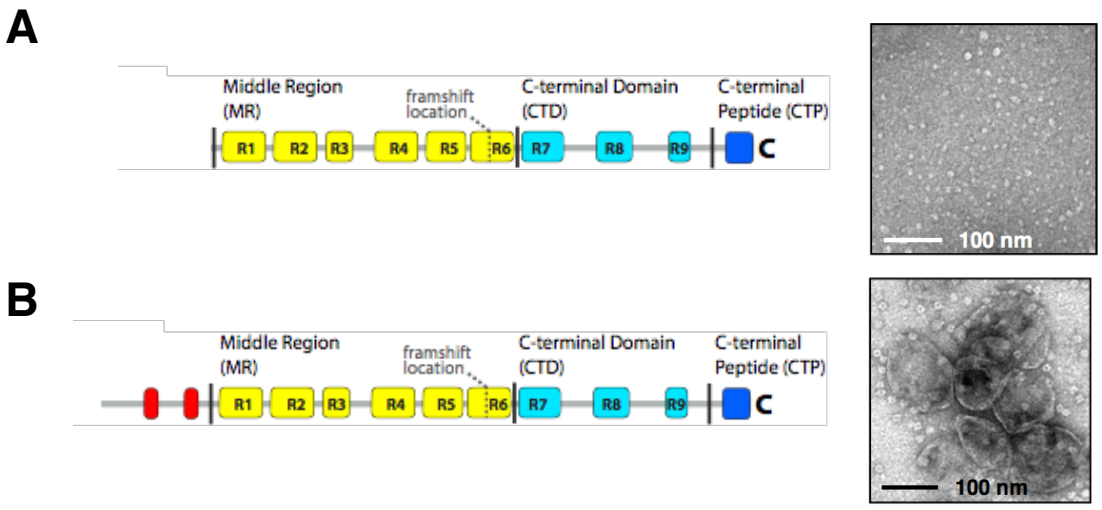

Figure 3-17: Carboxysomes when NTD is completely or partially removed
(A) The entire NTD is removed from CsoS2 in pHnCB10 plasmid, as shown on the left. A negative-staining TEM image of the fraction that normally contains carboxysomes is shown on the right. (B) Repeat 1 and 2 in the NTD are removed (left). Electron micrograph of the purified particle is shown on the right.

3.3.8 Verifying the Liquid-like Property of CsoS2

Smooth spherical appearance of protein in solution can be indicative of liquid-liquid phase separation (Li et al., 2012). Phase-separated macromolecular assemblies, which have been reported to involve repetitive multivalent IDPs in several cases, used their concentrated binding sites to rapidly recruit their binding partners (Burke et al., 2015; Ekman et al., 2006; Jiang et al., 2015; Li et al., 2012; Nott et al., 2015). During our optimization of the sfGFP-labeled CsoS2B purification protocol, we discovered that the presence of the reducing agent TCEP (0.5 mM) during the cell lysis and purification led to spherically-shaped CsoS2B-sfGFP (0.8±0.2 μm), as visualized by fluorescent microscopy (Figure 3-18A). The parallel purification without TCEP, on the other hand, gave rise to branched CsoS2B aggregates. Addition of TCEP (0.5 mM and 5 mM) and dithiothreitol (DTT, 5 and 50 mM) to these aggregates after the purification did not convert them into spherical droplet-like particles (Figure 3-18B). However, if TCEP was removed via buffer exchange, the spherically-shaped CsoS2B-sfGFP turned into the branched morphology similar to the protein purified without TCEP. Collectively, these finding suggests that CsoS2B-sfGFP originally existed as spherical particles, but oxidation during lysis and purification irreversibly turned them into branched aggregates, likely via disulfide crosslinking.

To determine if the droplet-like CsoS2B-sfGFP particles can recruit RuBisCO, purified His-tagged *H. neapolitanus* RuBisCO (3 μM) and CsoS2B-sfGFP (1.6 μM) were incubated together for 10 minutes and inspected with fluorescent microscopy.

The images revealed a large number of net-like aggregates with the diameter of up to 10 μm, and brightly fluorescent spherical particles were occasionally found embedded in these nascent aggregates (Figure 3-18C). Because a significant number of the droplet-like structures still remained, we suspected that the nascent aggregates were assembled from RuBisCO and freely-diffuse CsoS2B-sfGFP, while the spherical bodies did not participate. To confirm this possibility, CsoS2B-sfGFP was subjected to ultracentrifugation at 105,000g to pellet the droplet-like particles. Supernatant was set aside and the pellet fraction was resuspended in the imaging buffer [50 mM Tris and 150 mM NaCl (pH 7.5)]. The supernatant and pellet fractions were then incubated with RuBisCO as described previously. Contrary to our expectation, both fractions were capable of forming expanded aggregates in a similar manner, although some bright spherical particles could still be found in among the dimmer aggregates in the pellet fraction. Therefore, freely-diffuse CsoS2B-sfGFP as well as those located in the periphery of the spherical particles may be in the form that was active in binding RuBisCO.

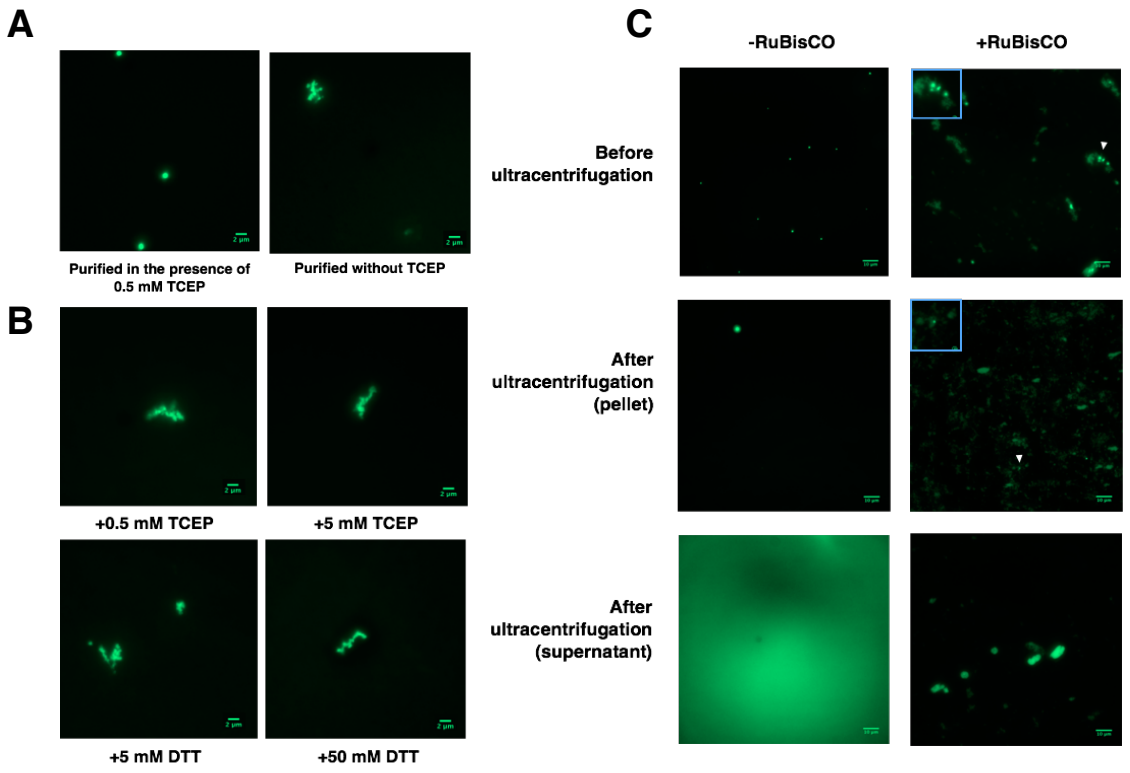

Figure 3-18: Reductant-dependent morphology of CsoS2B-sfGFP particles
(A) Fluorescent microscopic images of CsoS2B-sfGFP lysed and purified in the presence (left) and absence (right) of 0.5 mM TCEP (scale bar = 2 μm). (B) Morphology of CsoS2B-sfGFP purified without TCEP when treated with TCEP (0.5 and 5 mM) and DTT (5 and 50 mM) after the purification. (C) Appearance of CsoS2B-sfGFP purified with 0.5 mM TCEP in the solution before ultracentrifugation at 100,000g (top left), in the pellet after ultracentrifugation (middle left), and in the supernatant after ultracentrifugation (bottom left). The images of these samples after incubation with RuBisCO are shown to the right. White arrows indicate spherical particles embedded in the nascent aggregates and their magnified images are shown in the insets. Scale bar = 10 μm.

Since the behavior exhibited by CsoS2B-sfGFP thus far did not resemble a liquid, we decided to verify whether the spherical droplets we observed were indeed phase-separated liquid. To this end, we employed Fluorescent-Recovery-After-Photobleaching (FRAP) technique, in which liquid-like molecules were expected to be able to quickly replenish the bleached region with other unbleached fluorescent molecules, giving rise to fast fluorescent recovery. After 10 iterations of bleaching, corresponding to 5 s total, the bleached region (~ 1 pixel diameter) in a CsoS2B-sfGFP particle did not recover even after a minute (Figure 3-19A). While this would suggest that CsoS2B-sfGFP was not liquid-like, the wavelength of the laser (488 nm) used in bleaching was relatively close to the diameter of CsoS2-sfGFP particles (0.8±0.2 μm), which might have caused off-target bleaching that covered a wide area around the region of interest. Therefore, we additionally investigated the liquid-like behavior of CsoS2B-sfGFP using aliphatic alcohols, which were shown to disrupt hydrophobic self-association of IDP inside phase-separated liquid bodies (Molliex et al., 2015; Shulga and Goldfarb, 2003). Different concentrations of n-butanol or 1,6-hexanediol, previously reported to be the most effective at disrupting nucleoporins (Shulga and Goldfarb, 2003), were applied to CsoS2B-sfGFP and allowed to incubate for 15 min at room temperature. At 5 and 10% v/v, 1,6-hexanediol did not appear to alter the morphology of the spherical particles (Figure 3-19B) while n-butanol caused free CsoS2B-sfGFP to crash out. When n-butanol was reduced to 1% or 3%, CsoS2B-sfGFP did not crash out, but no change to the spherical structures was observed (data not shown). Finally, we treated the spherical particles with various types of reducing agents to see if the robustness was due to disulfide crosslinks in the core, but no change was observed (Figure 3-19B). Therefore, contrary to our initial expectation, CsoS2B-sfGFP spherical particles did not behave like the phase-separated liquid under the conditions of our experiments.

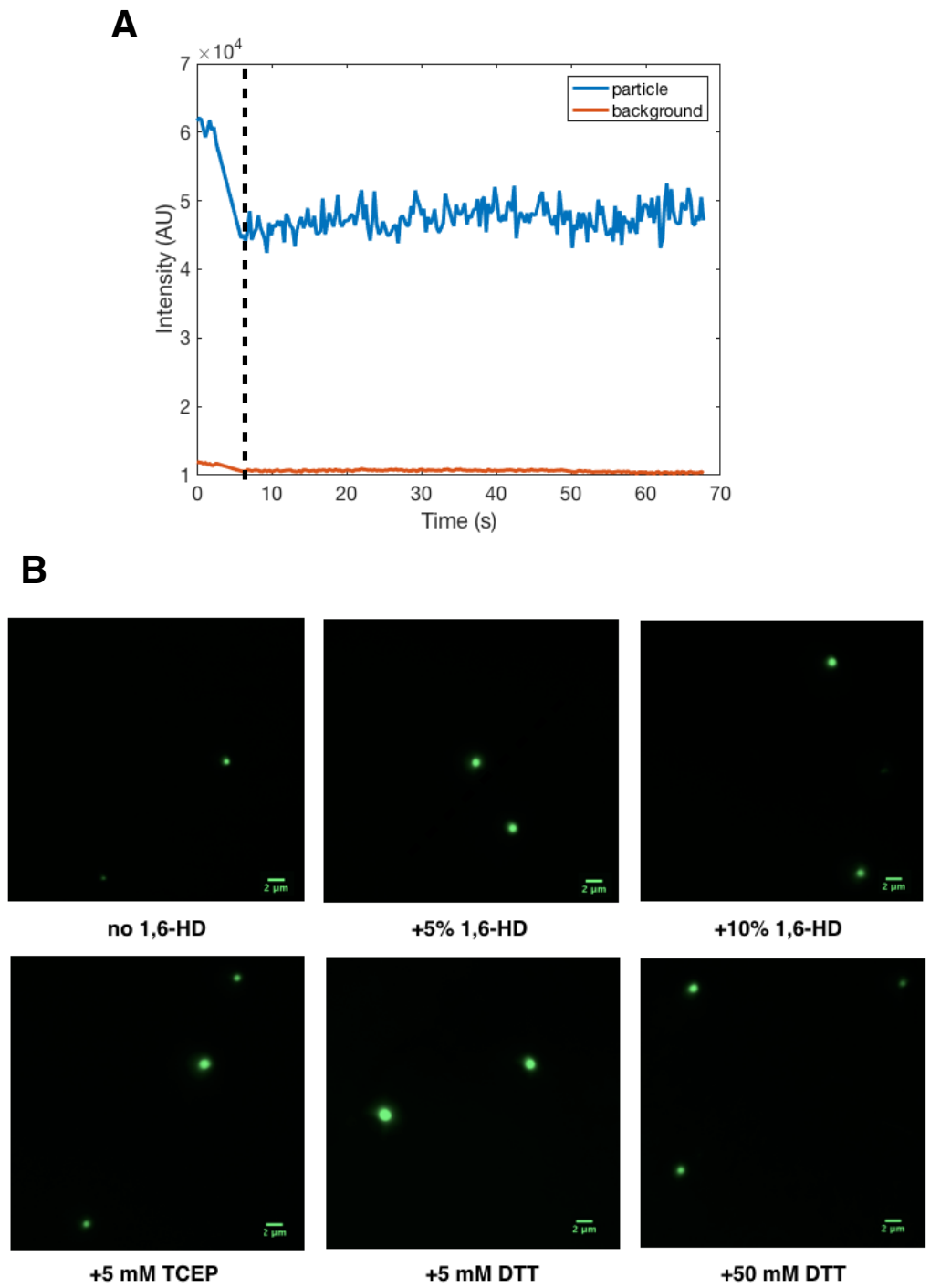

Figure 3-19: Assessment of the liquid-like behavior of CsoS2B-sfGFP
(A) Fluorescent intensity of the regions of interest over the course of the FRAP experiment. Bleaching was started at t = 0 and ended at the broken line. (B) Fluorescent microscopic images of CsoS2B-sfGFP in the presence of 1,6-hexanediol (1,6-HD, 5 and 10%) or reducing agents (5 mM TCEP, 5 mM DTT, and 50 mM DTT)

3.4 Discussion

In this chapter, we have shown that CsoS2 can assemble RuBisCO into paracrystalline aggregates. The N-terminal domain of CsoS2 was responsible for RuBisCO nucleation, which explains why both CsoS2A and B were able to give rise to RuBisCO aggregates. The binding of NTD to RuBisCO was shown to be multivalent and likely mediated by its four short amphiphatic helices. Complete removal of NTD from CsoS2 abolished carboxysome formation, whereas omission of the first two repeats yielded morphologically abnormal particles. Lastly, we found that CsoS2B purified in the presence of a reducing agent displayed spherical droplet-like particles. While we initially hypothesized that these particles were the product of liquid-liquid phase transition, the results from the FRAP experiment and chemical disruption demonstrate otherwise.

3.4.1 Soluble Expression and Purification of Recombinant RuBisCO and CsoS2wt

Preparation of proteins of interest is a requisite for *in vitro* biochemical studies. While expression of proteins can be straightforward and does not deserve detailed elaboration, such is not the case for *H. neapolitanus* RuBisCO and CsoS2wt. RuBisCO homologs from various organisms suffered from insoluble expression when overexpressed in *E. coli* (Georgiou and Valax, 1996; Liu et al., 2010; Wheatley et al., 2014). We found that coexpression of GroEL/ES was sufficient to solubilize *H. neapolitanus* CbbLS, which agrees with the findings that the large subunits of RuBisCO rely on the chaperonin binding to properly fold (Liu et al., 2010). In addition, a protein called RuBisCO activase was found to help release the large subunits of RuBisCO from GroEL/ES and promote proper assembly with small subunits into a hexadecamer (Hauser et al., 2015). Shortly after we settled on the current RuBisCO expression protocol, a study by Wheatley *et al.* demonstrated that a pterin dehydratase-like protein called acRAF from *H. neapolitanus* can increase the solubility of *H. neapolitanus* RuBisCO even further(Wheatley et al., 2014). This protein may function as a RuBisCO activase in *H. neapolitanus*. We were able to acquire soluble and functional RuBisCO with GroEL/ES coexpression alone in the absence of acRAF. However, we did observe that a slight amount of GroEL/ES did copurify with RuBisCO even when an exhaustive wash step was included. Co-expression of acRAF may help improve the purity by releasing the residual GroEL/ES from RuBisCO.

It was previously reported that *H. neapolitanus* CsoS2wt was insoluble *in vitro* and soluble expression was possible only after the removal of 250 aa from the N-terminus (Dou, 2009). Fortunately, our purification of CsoS2wt yielded completely soluble protein, which could be the result of a different choice of plasmid and *E. coli* expression strain. Because of CsoS2's solubility, we were not required to remove the 250 aa on the N-terminus, which turned out to be critical to CsoS2's function.

As anticipated from an intrinsically-disordered protein, the purified CsoS2wt displayed many second-order bands on SDS-PAGE, likely corresponding to the

products of degradation by protease. A recommended strategy for minimizing protease degradation of an IDP is to denature the harvested cells with either urea or boiling, which will inactivate proteases while not affecting the IDP. Unfortunately, this strategy did not reduce the degradation, suggesting that CsoS2wt might have already been degraded *in vivo*. As an alternative, we purified CsoS2wt directly from denatured *H. neapolitanus* carboxysomes and achieved satisfactory yield and purity. Purification of CsoS2wt variants using this strategy, however, would only be possible if the CsoS2 variant of interest still allows the formation of carboxysomes. In the case where direct purification from carboxysomes is not possible, the IDP of interest can be fused to a fusion parter that localizes to inclusion bodies, which help shield the IDP from proteases (Hwang et al., 2014). The protein can then be recovered by purification under a denaturing condition, and then cleaved to remove the insoluble tag.

3.4.2 RuBisCO-Nucleating Ability of CsoS2A and B was Attributed to NTD

Using TEM, we have shown that CsoS2wt and RuBisCO assemble into paracrystalline aggregates that resemble the interior of the carboxysome. Biochemical experiments such as native gel electrophoresis and *in vivo* pull-down also confirmed the interaction between the two proteins. While it is tempting to speculate that the observed CsoS2wt-RuBisCO aggregates represent the intermediate of the assembly pathway, it is more likely that the biologically relevant intermediate also involves the shell facet that serves as a platform to seed CsoS2-RuBisCO network and allow it to fill inward until the assembly is complete. Indeed, an electron microscopic study on *H. neapolitanus* carboxysomes detected a shell fragment lined with RuBisCO particles (Iancu et al., 2010), which may resemble the intermediate described thus far.

According to the putative model we proposed in Chapter 2, both CsoS2A and CsoS2B can bind RuBisCO, whereas only CsoS2B can link RuBisCO to the shell owing to its unique C-terminus. Here, we have demonstrated that CsoS2A and B were capable of forming extensive aggregation *in vitro* in the presence of RuBisCO. In addition, we were able to single out NTD as the domain responsible for RuBisCO nucleation. The ability of both CsoS2A and B to seed RuBisCO aggregation via their N-terminal domain is reminiscent of the role CcmM35 and CcmM58 in β-carboxysome assembly, in which both isoforms are capable of linking RuBisCO via the RbcS-like domain but only the long isoform associates to the shell by the carbonic anhydrase-like domain. This similarity may represent a common strategy used by cell to save the resources; since only cargo molecules on the periphery are linked to the shell protein, biosynthesis of the shell-binding domain in every molecule of the scaffold protein is potentially wasteful. Unfortunately, in our attempt to investigate the morphology of these aggregates, we could only successfully employ fluorescent microscopy, not Negative-Staining TEM. The detection limit of fluorescent microscopy prevented us from determining the actual morphology of the complexes when a CsoS2 was absent. Nonetheless, the native agarose electrophoresis experiment to detect mobility shift of RuBisCO in the presence of CsoS2A- or

CsoS2B-sfGFP suggests that these proteins formed aggregates to a lesser extent, possibly explaining the lack of aggregates under EM.

During our attempt to test the copurification of CsoS2A or B with RuBisCO, we observed that the isoform distribution in the cell lysates was significantly biased towards CsoS2A; even the strain that was supposed to express CsoS2B alone yielded equal amount of CsoS2A and B. For this reason, we were not able to use this assay to determine whether CsoS2B can truly interact with RuBisCO *in vivo*. The mechanism responsible for this aberration, however, deserves further investigation as it may lend us an insight into the finer details of how CsoS2's function is regulated. For example, RuBisCO may co-translationally bind to the nascent CsoS2 peptide (Duncan and Mata, 2011), causing the ribosome to stall and giving more opportunity for the frameshifting to occur. This phenomenon was not observed when the full cso operon was expressed, suggesting that the binding of the nascent CsoS2 peptide to other proteins from the cso operon may help negate the RuBisCO's binding via an unknown mechanism. To confirm this scenario, it may be helpful to compare the ability of RuBisCO to pull down *csoS2* mRNA in the presence and absence of the functional start codon on *csoS2* ORF. If RuBisCO-mRNA copurification depends on the presence of an active start codon, it would suggest that RuBisCO binds to the nascent CsoS2 peptide, which in turn is associated to the ribosome-mRNA complex and allows the mRNA copurification (Duncan and Mata, 2011).

3.4.3 Implications of NTD-mediated Multivalent Binding of RuBisCO

We demonstrated that NTD of CsoS2 was responsible for binding RuBisCO using *in vivo* pull-down and BLI assay. The use of synthetic bivalent NTD in BLI informed us that multiple molecules of NTD that localize near each other on the probe surface may cooperatively interact with RuBisCO, resulting in a fast and high-affinity interaction. In the context of the carboxysome formation, this scenario may portray a possible strategy to switch on the assembly. When CsoS2 freely diffuses in cytosol, the molecules are far apart such that cooperative binding is not possible. However, once a facet of the shell has partially formed, the lumenal side may serve as a platform where CsoS2 can densely localize, allowing RuBisCO binding, and consequently the assembly, to proceed rapidly. The two-stage model of interaction may explain how the BLI signal from CsoS2wt/NTD includes both reversible and irreversible binding components.

Functionally, the reliance on multivalent interaction for strong binding between CsoS2 and their partners may be beneficial for the cells. In nature, cognate expression of IDPs in the cytosol is kept low due to their tendency to self-aggregate or participate in non-specific interaction with house-keeping proteins at high concentration (Babu et al., 2011). Formation of dense CsoS2 patches on a shell facet obviates the need to express a large amount of CsoS2 in the cytosol. In addition, multivalent interaction mediated by IDP is often accompanied by switch-like transition, with significant interaction only occurring after a certain concentration threshold is reached (Li et al., 2012). This may prevent CsoS2 from forming off-

target intermediates with other carboxysomal proteins, until the assembly is "green-lighted" by the formation of shell facets. It would be of special interest to determine whether CcmM35/58 exhibit display a similar two-stage interaction with RuBisCO in β-carboxysomes, but studies thus far have not investigated CcmM-RuBisCO interaction in microscopic details. Considering that the RbcS-like repeats in CcmM were proposed to be integrated into RuBisCO hexadecamers directly (Long et al., 2011), it is possible that CcmM's association to RuBisCO is not reversible to begin with.

3.4.4 Interaction Surface of α-Helices in NTD

From circular dichroism spectroscopy and bioinformatic prediction of secondary structure, we found that NTD contained four amphipathic helices with multiple conserved hydrophobic and charged residues. Mutating lysine or arginine on position 2 and 9 on each helix to alanine completely abolished RuBisCO binding, suggesting that these residues are likely involved in the interaction. This finding also implies that electrostatic interaction is essential to NTD's binding competency, possibly in the formation of the encounter complex which is generally dominated by electrostatic interactions (Ubbink, 2009). In certain cases, the electrostatic interaction that promotes the encounter of two binding partners is subsequently dominated by hydrophobic interaction that provides a better fit (McLendon, 1991). We found that mutating the nearest leucine (L12) to R/K2 significantly altered the irreversible binding to RuBisCO, but not the weaker reversible interaction. Therefore, it is possible that this leucine is among the residues that participate in the post-encounter binding step, although the true necessity of L12 might have been masked by other hydrophobic residues nearby. Additional mutations of other hydrophobic residues may display a more dramatic impact on the binding capability of NTD.

The fact that the NTD's binding to RuBisCO appears to rely on its structured portion raises the question of what role intrinsic disorder, which constitutes 75% of NTD, plays in this domain. The simplest scenario would be that the long disordered regions serve as spacers that set the helices sufficiently apart from each other, allowing them to cover a large area or wrap around large protomers like RuBisCO. Given the usual high functional density of IDPs, we also cannot rule out the possibility that NTD can bind to other proteins besides RuBisCO. For instance, as NTD was found to slightly self-aggregate, the disordered linkers may contain self-association motifs, which may help cluster multiple NTDs and poise them for multivalent interaction. This would be supported by our finding that the complete or partial removal of NTD from CsoS2 had a dramatic effect on the formation of carboxysome. even though the presence of RuBisCO was previously reported to be unnecessary for this process (Menon et al., 2008).

3.4.5 Potential Roles of MR and CTD in α-Carboxysome Assembly
Aside from NTD, the two other individual domains, MR and CTD, were also biochemically tested for their ability to bind RuBisCO. CTD was confirmed by both *in vivo* pull-down and BLI to possess no interaction towards RuBisCO. MR's ability in binding RuBisCO still remains inconclusive, as it was copurified with RuBisCO in the *in vivo* pull-down assay but did not show any signal in the BLI assay. As discussed in section 3.3.4, this discrepancy may be caused by the false-positive result from *in vivo* pull-down or the false-negative result from BLI.

While determining the binding partners of MR and CTD besides RuBisCO is beyond the scope of this study, we believe that these two domains are likely to mediate some form of interaction due to the abundance of [V/I][S/T]G motifs. Similar short hydrophobic repeats have been found to function in protein-protein interaction in some intrinsically-disordered proteins (Bayliss et al., 2000; Burke et al., 2015; lde and Kehlenbach, 2010)
. Therefore, [V/I][S/T]G motifs may perform similar roles in MR and CTD. It is worth noting that the presence of [V/I][S/T]G motifs does not necessarily indicate that both MR and CTD interact with the same set of proteins. This is due to the fact that intrinsically-disordered regions can display different interaction profiles depending on their conformations, which in turn are determined by local amino acid composition (Cino et al., 2013; Oldfield et al., 2008; Tompa et al., 2005). [V/I][S/T]G motifs in MR and CTD were surrounded by sequences with different lengths and composition, so they could adopt significantly different conformations and perform different roles.

In our study of each individual CsoS2 domain, we made the assumption that each domain functions independently and thus the behavior of each domain by itself would accurately reflect its true function in the full CsoS2. In reality, intrinsically-disordered proteins may exhibit interdependency between different domains. For instance, the intrinsically-disordered hub protein called the adenovirus early region 1A (E1A) oncoprotein binds its partners, CREB binding protein TAZ2 and the retinoblastoma protein (pRb), with the ability to switch between negative and positive cooperativity depending on the presence or absence of the N- or C-terminal domain (Ferreon et al., 2013). In our case, RuBisCO-binding capability of NTD was unlikely to be significantly influenced by MR and CTD, as the binding behavior exhibited by NTD in the BLI assay was similar to full CsoS2. Nonetheless, it is possible that the functions of MR and CTD would be more sensitive to removal of the other domains. Future studies should consider employing higher-resolution analytical techniques that can locally detect the interaction contributed by each residue in full CsoS2 to avoid dissecting the protein, such as protein NMR or H/D exchange coupled with mass spectrometry (Keppel et al., 2011).

3.4.6 CsoS2 Aggregates: Liquid or Solid?
The hallmark of CsoS2 discussed thus far – being an IDP with multiple repeats are also the features that enable liquid-liquid phase separation of macromolecules. This

is the mechanism that drives the formation of eukaryotic membrane-less organelles such as RNA granules (Lin et al., 2015), NWASP-NCK-nephrin (Li et al., 2012), and BugZ-mediated spindle assembly (Jiang et al., 2015). Our fluorescent microscopic investigation of TCEP-treated CsoS2B-sfGFP reveals that the protein can phase-separate into spherical structures, in contrast to the rugged aggregates formed when TCEP was omitted. However, these structures did not appear to behave like a liquid: they did not undergo fast recovery after bleaching; they expanded into branched aggregates in the presence of RuBisCO; and they could not be dissolved by aliphatic alcohols. It is possible that TCEP-treated CsoS2B-sfGFP might have started off as phase-separated liquid, but the core of the particles became solidified while the periphery remained liquid. A similar phenomenon was observed in the RNA binding protein hnRNPA1 which underwent fibrillization at high local protein concentration in the droplet (Molliex et al., 2015). Overexpression of CsoS2B-sfGFP in *E. coli* might have resulted in high concentration of CsoS2B per droplet, and in turn solidification of the particle's interior. In contrast, in *H. neapolitanus* where CsoS2 is likely expressed in a moderate level, liquid droplets may be able to form. Lowering the expression level of CsoS2B-sfGFP in *E. coli* by adjusting level of induction or expressing from a low-copy plasmid may yield CsoS2B-sfGFP that behaves more closely to its native counterpart.

If the assembly of *H. neapolitanus* carboxysome indeed proceeds via liquid-liquid phase separation, what biological advantages does it gain? It has been shown that liquid-liquid phase transition occurs in a switch-like manner and is highly sensitive to the cellular abundance of its composition and change in cellular environment (Banani et al., 2016). Once the conditions are met, the assembly can complete within seconds or minutes (Molliex et al., 2015). It is possible that the organisms with α-carboxysomes have to encounter a situation that necessitates prompt carboxysome formation or property adjustment, such as drastic changes of CO_2 levels or changes in reducing power from light or lithotrophic source. In such scenario, fast carboxysome assembly with suitable stoichiometry of subunits can increase the chance of survival, which would rationalize the widespread use of CsoS2 in α-carboxysome assembly. It remains to be seen whether organisms that carry β-carboxysomes need such spontaneous carboxysome formation or use a similar alternative.

3.5 References

Axen, S.D., Erbilgin, O., and Kerfeld, C.A. (2014). A taxonomy of bacterial microcompartment loci constructed by a novel scoring method. PLoS Comp Biol *10*, e1003898.

Babu, M.M., van der Lee, R., de Groot, N.S., and Gsponer, J. (2011). Intrinsically disordered proteins: regulation and disease. Curr. Opin. Struct. Biol. *21*, 432–440.

Badger, M.R., Andrews, T.J., and Whitney, S.M. (1998). The diversity and coevolution of Rubisco, plastids, pyrenoids, and chloroplast-based CO_2-concentrating mechanisms in algae. Canadian Journal of Botany *76*, 1052–1071.

Baker, S.H., Lorbach, S.C., Rodriguez-Buey, M., Williams, D.S., Aldrich, H.C., and Shively, J.M. (1999). The correlation of the gene *csoS2* of the carboxysome operon with two polypeptides of the carboxysome in *Thiobacillus neapolitanus*. Arch. Microbiol. *172*, 233–239.

Banani, S.F., Rice, A.M., Peeples, W.B., Lin, Y., Jain, S., Parker, R., and Rosen, M.K. (2016). Compositional control of phase-separated cellular bodies. Cell *166*, 651–663.

Barbar, E., and Nyarko, A. (2015). Polybivalency and disordered proteins in ordering macromolecular assemblies. Semin. Cell Dev. Biol. *37*, 20–25.

Bayliss, R., Littlewood, T., and Stewart, M. (2000). Structural basis for the interaction between FxFG nucleoporin repeats and importin-beta in nuclear trafficking. Cell *102*, 99–108.

Burke, K.A., Janke, A.M., Rhine, C.L., and Fawzi, N.L. (2015). Residue-by-residue view of in vitro FUS granules that bind the C-terminal domain of RNA polymerase-II. Molecular Cell *60*, 231–241.

Cai, F., Dou, Z., Bernstein, S., Leverenz, R., Williams, E., Heinhorst, S., Shively, J., Cannon, G., and Kerfeld, C. (2015a). Advances in understanding carboxysome assembly in *Prochlorococcus* and *Synechococcus* implicate CsoS2 as a critical component. Life *5*, 1141–1171.

Cai, F., Sutter, M., Bernstein, S.L., Kinney, J.N., and Kerfeld, C.A. (2015b). Engineering bacterial microcompartment shells: chimeric shell proteins and chimeric carboxysome shells. ACS Synth. Biol. *4*, 444–453.

Cameron, J.C., Wilson, S.C., Bernstein, S.L., and Kerfeld, C.A. (2013). Biogenesis of a bacterial organelle: the carboxysome assembly pathway. Cell *155*, 1131–1140.

Chen, A.H., Robinson-Mosher, A., Savage, D.F., Silver, P.A., and Polka, J.K. (2013). The bacterial carbon-fixing organelle is formed by shell envelopment of preassembled cargo. PLoS ONE *8*, e76127.

Choudhary, S., Quin, M.B., Sanders, M.A., Johnson, E.T., and al, E. (2012). Engineered protein nano-compartments for targeted enzyme localization. PLoS ONE *7*.

Cino, E.A., Killoran, R.C., Karttunen, M., and Choy, W.-Y. (2013). Binding of disordered proteins to a protein hub. Sci Rep *3*, 2305.

Dou, Z. (2009). Functional characterization and assembly studies of carboxysomes in *Halothiobacillus neapolitanus*. 1–176.

Drozdetskiy, A., Cole, C., Procter, J., and Barton, G.J. (2015). JPred4: a protein secondary structure prediction server. Nucleic Acids Res. *43*, W389–W394.

Duncan, C.D.S., and Mata, J. (2011). Widespread cotranslational formation of protein complexes. PLoS Genet *7*, e1002398.

Ekman, D., Light, S., Björklund, Å.K., and Elofsson, A. (2006). What properties characterize the hub proteins of the protein-protein interaction network of *Saccharomyces cerevisiae*? Genome Biol *7*, R45.

Fan, C., and Bobik, T.A. (2011). The N-terminal region of the medium subunit (PduD) packages adenosylcobalamin-dependent diol dehydratase (PduCDE) into the Pdu microcompartment. J. Bacteriol. *193*, 5623–5628.

Fan, C., Cheng, S., Liu, Y., Escobar, C.M., Crowley, C.S., Jefferson, R.E., Yeates, T.O., and Bobik, T.A. (2010). Short N-terminal sequences package proteins into bacterial microcompartments. Proc Natl Acad Sci USA *107*, 7509–7514.

Fan, C., Cheng, S., Sinha, S., and Bobik, T.A. (2012). Interactions between the termini of lumen enzymes and shell proteins mediate enzyme encapsulation into bacterial microcompartments. Proc Natl Acad Sci USA *109*, 14995–15000.

Ferreon, A.C.M., Ferreon, J.C., Wright, P.E., and Deniz, A.A. (2013). Modulation of allostery by protein intrinsic disorder. Nature *498*, 390–394.

Georgiou, G., and Valax, P. (1996). Expression of correctly folded proteins in *Escherichia coli*. Curr. Opin. Biotechnol. *7*, 190–197.

Gonzales, A.D., Light, Y.K., and Zhang, Z. (2005). Proteomic analysis of the CO_2-concentrating mechanism in the open-ocean cyanobacterium *Synechococcus* WH8102. Canadian Journal of Botany *83*, 735–745.

Greenfield, N.J. (2006). Using circular dichroism spectra to estimate protein secondary structure. Nat Protoc *1*, 2876–2890.

Harbury, P.B., Zhang, T., Kim, P.S., and Alber, T. (1993). A switch between two-, three-, and four-stranded coiled coils in GCN4 leucine zipper mutants. Science *262*, 1401–1407.

Hauser, T., Popilka, L., Hartl, F.U., and Hayer-Hartl, M. (2015). Role of auxiliary proteins in Rubisco biogenesis and function. Nplants *1*, 15065.

Heinhorst, S., Cannon, G.C., and Shively, J.M. (2006). Carboxysomes and carboxysome-like inclusions. Complex Intracellular Structures in Prokaryotes 142–165.

Held, M., Kolb, A., Perdue, S., Hsu, S., Bloch, S.E., Quin, M.B., and Schmidt-Dannert, C.

(2016). Engineering formation of multiple recombinant Eut protein nanocompartments in *E. coli*. Sci Rep *6*, 24359.

Huang, Y., and Liu, Z. (2009). Kinetic advantage of intrinsically disordered proteins in coupled folding-binding process: a critical assessment of the "fly-casting" mechanism. J. Mol. Biol. *393*, 1143–1159.

Hwang, P.M., Pan, J.S., and Sykes, B.D. (2014). Targeted expression, purification, and cleavage of fusion proteins from inclusion bodies in *Escherichia coli*. FEBS Lett. *588*, 247–252.

Iancu, C.V., Morris, D.M., Dou, Z., Heinhorst, S., Cannon, G.C., and Jensen, G.J. (2010). Organization, structure, and assembly of alpha-carboxysomes determined by electron cryotomography of intact cells. J. Mol. Biol. *396*, 105–117.

Jerabek-Willemsen, M., André, T., Wanner, R., Roth, H.M., Duhr, S., Baaske, P., and Breitsprecher, D. (2014). MicroScale Thermophoresis: Interaction analysis and beyond. J. Mol. Struct *1077*, 101–113.

Jiang, H., Wang, S., Huang, Y., He, X., Cui, H., Zhu, X., and Zheng, Y. (2015). Phase transition of spindle-associated protein regulate spindle apparatus assembly. Cell *163*, 108–122.

Karush, F. (1976). Multivalent binding and functional affinity. In Contemporary Topics in Molecular Immunology, (Boston, MA: Springer US), pp. 217–228.

Keppel, T.R., Howard, B.A., and Weis, D.D. (2011). Mapping unstructured regions and synergistic folding in intrinsically disordered proteins with amide H/D exchange mass spectrometry. Biochemistry *50*, 8722–8732.

Kerfeld, C.A., and Erbilgin, O. (2015). Bacterial microcompartments and the modular construction of microbial metabolism. Trends Microbiol. *23*, 22–34.

Kinney, J.N., Salmeen, A., Cai, F., and Kerfeld, C.A. (2012). Elucidating essential role of conserved carboxysomal protein CcmN reveals common feature of bacterial microcompartment assembly. J. Biol. Chem. *287*, 17729–17736.

Lassila, J.K., Bernstein, S.L., Kinney, J.N., Axen, S.D., and Kerfeld, C.A. (2014). Assembly of robust bacterial microcompartment shells using building blocks from an organelle of unknown function. J. Mol. Biol. *426*, 2217–2228.

Lawrence, A.D., Frank, S., Newnham, S., Lee, M.J., Brown, I.R., Xue, W., Rowe, M.L., Mulvihill, D.P., Prentice, M.B., Howard, M.J., et al. (2014). Solution structure of a bacterial microcompartment targeting peptide and its application in the construction of an ethanol bioreactor. ACS Synth. Biol.

lde, S.W., and Kehlenbach, R.H. (2010). The Part and the Whole: functions of

nucleoporins in nucleocytoplasmic transport. Trends Cell Biol. *20*, 461–469.

Li, P., Banjade, S., Cheng, H., Kim, S., Chen, B., Guo, L., Llaguno, M., Hollingsworth, J.V., King, D.S., Banani, S.F., et al. (2012). Phase transitions in the assembly of multivalent signalling proteins. Nature *483*, 336–340.

Lin, M.T., Occhialini, A., Andralojc, P.J., Devonshire, J., Hines, K.M., Parry, M.A.J., and Hanson, M.R. (2014). β-Carboxysomal proteins assemble into highly organized structures in Nicotiana chloroplasts. Plant J *79*, 1–12.

Lin, Y., Protter, D.S.W., Rosen, M.K., and Parker, R. (2015). Formation and maturation of phase-separated liquid droplets by RNA-binding proteins. Molecular Cell *60*, 208–219.

Liu, C., Young, A.L., Starling-Windhof, A., Bracher, A., Saschenbrecker, S., Rao, B.V., Rao, K.V., Berninghausen, O., Mielke, T., Hartl, F.U., et al. (2010). Coupled chaperone action in folding and assembly of hexadecameric Rubisco. Nature *463*, 197–202.

Long, B.M., Badger, M.R., Whitney, S.M., and Price, G.D. (2007). Analysis of carboxysomes from *Synechococcus* PCC7942 reveals multiple RuBisCO complexes with carboxysomal proteins CcmM and CcaA. J. Biol. Chem. *282*, 29323–29335.

Long, B.M., Rae, B.D., Badger, M.R., and Price, G.D. (2011). Over-expression of the β-carboxysomal CcmM protein in Synechococcus PCC7942 reveals a tight co-regulation of carboxysomal carbonic anhydrase (CcaA) and M58 content. Photosyn. Res. *109*, 33–45.

Long, B.M., Tucker, L., Badger, M.R., and Price, G.D. (2010). Functional cyanobacterial beta-carboxysomes have an absolute requirement for both long and short forms of the CcmM protein. Plant Physiol. *153*, 285–293.

Mackinder, L.C.M., Meyer, M.T., Mettler-Altmann, T., Chen, V.K., Mitchell, M.C., Caspari, O., Freeman Rosenzweig, E.S., Pallesen, L., Reeves, G., Itakura, A., et al. (2016). A repeat protein links Rubisco to form the eukaryotic carbon-concentrating organelle. Proc. Natl. Acad. Sci. U.S.A *113*, 5958–5963.

Marshall, W.F. (2004). Cellular length control systems. Annu. Rev. Cell Dev. Biol. *20*, 677–693.

McLendon, G. (1991). Control of biological electron transport via molecular recognition and binding: The “velcro” model. 159–174.

Menon, B.B., Dou, Z., Heinhorst, S., Shively, J.M., and Cannon, G.C. (2008). *Halothiobacillus neapolitanus* carboxysomes sequester heterologous and chimeric RuBisCO Species. PLoS ONE *3*, e3570.

Micsonai, A., Wien, F., Kernya, L., Lee, Y.-H., Goto, Y., Réfrégiers, M., and Kardos, J.

(2015). Accurate secondary structure prediction and fold recognition for circular dichroism spectroscopy. Proc. Natl. Acad. Sci. U.S.A *112*, E3095–E3103.

Minten, I.J., Hendriks, L.J.A., Nolte, R.J.M., and Cornelissen, J.J.L.M. (2009). Controlled encapsulation of multiple proteins in virus capsids. J. Am. Chem. Soc. *131*, 17771–17773.

Molliex, A., Temirov, J., Lee, J., Coughlin, M., Kanagaraj, A.P., Kim, H.J., Mittag, T., and Taylor, J.P. (2015). Phase separation by low complexity domains promotes stress granule assembly and drives pathological fibrillization. Cell *163*, 123–133.

Nott, T.J., Petsalaki, E., Farber, P., Jervis, D., Fussner, E., Plochowietz, A., Craggs, T.D., Bazett-Jones, D.P., Pawson, T., Forman-Kay, J.D., et al. (2015). Phase transition of a disordered nuage protein generates environmentally responsive membraneless organelles. Molecular Cell *57*, 936–947.

Nyarko, A., Song, Y., Nováček, J., Žídek, L., and Barbar, E. (2013). Multiple recognition motifs in nucleoporin Nup159 provide a stable and rigid Nup159-Dyn2 assembly. J. Biol. Chem. *288*, 2614–2622.

Oldfield, C.J., Cheng, Y., Cortese, M.S., Romero, P., Uversky, V.N., and Dunker, A.K. (2005). Coupled folding and binding with alpha-helix-forming molecular recognition elements. Biochemistry *44*, 12454–12470.

Oldfield, C.J., Meng, J., Yang, J.Y., Yang, M.Q., Uversky, V.N., and Dunker, A.K. (2008). Flexible nets: disorder and induced fit in the associations of p53 and 14-3-3 with their partners. BMC Genomics *9 Suppl 1*, S1.

Peña, K.L., Castel, S.E., de Araujo, C., Espie, G.S., and Kimber, M.S. (2010). Structural basis of the oxidative activation of the carboxysomal gamma-carbonic anhydrase, CcmM. Proc. Natl. Acad. Sci. U.S.A *107*, 2455–2460.

Rae, B.D., Long, B.M., Badger, M.R., and Price, G.D. (2013). Functions, compositions, and evolution of the two types of carboxysomes: polyhedral microcompartments that facilitate CO_2 fixation in cyanobacteria and some proteobacteria. Microbiol. Mol. Biol. Rev. *77*, 357–379.

Schmid, M.F., Paredes, A.M., Khant, H.A., Soyer, F., Aldrich, H.C., Chiu, W., and Shively, J.M. (2006). Structure of *Halothiobacillus neapolitanus* carboxysomes by cryo-electron tomography. J. Mol. Biol. *364*, 526–535.

Segrest, J.P., De Loof, H., Dohlman, J.G., Brouillette, C.G., and Anantharamaiah, G.M. (1990). Amphipathic helix motif: classes and properties. Proteins *8*, 103–117.

Serebriiskii, I.G., and Golemis, E.A. (2001). Two-hybrid system and false positives. Approaches to detection and elimination. Methods Mol. Biol. *177*, 123–134.

Shah, N.B., and Duncan, T.M. (2014). Bio-layer Interferometry for Measuring Kinetics of Protein-protein Interactions and Allosteric Ligand Effects. J Vis Exp.

Shulga, N., and Goldfarb, D.S. (2003). Binding dynamics of structural nucleoporins govern nuclear pore complex permeability and may mediate channel gating. Mol. Cell. Biol. *23*, 534–542.

Smock, R.G., and Gierasch, L.M. (2009). Sending signals dynamically. Science *324*, 198–203.

Sugase, K., Dyson, H.J., and Wright, P.E. (2007). Mechanism of coupled folding and binding of an intrinsically disordered protein. Nature *447*, 1021–1025.

Sutter, M., Boehringer, D., Gutmann, S., Günther, S., Prangishvili, D., Loessner, M.J., Stetter, K.O., Weber-Ban, E., and Ban, N. (2008). Structural basis of enzyme encapsulation into a bacterial nanocompartment. Nat. Struct. Mol. Biol. *15*, 939–947.

Tompa, P. (2012). Intrinsically disordered proteins: a 10-year recap. Trends Biochem. Sci. *37*, 509–516.

Tompa, P., Szász, C., and Buday, L. (2005). Structural disorder throws new light on moonlighting. Trends Biochem. Sci. *30*, 484–489.

Ubbink, M. (2009). The courtship of proteins: Understanding the encounter complex. FEBS Lett. *583*, 1060–1066.

Uversky, V.N., and Dunker, A.K. (2010). Understanding protein non-folding. Biochim. Biophys. Acta *1804*, 1231–1264.

Uversky, V.N., Gillespie, J.R., and Fink, A.L. (2000). Why are “natively unfolded” proteins unstructured under physiologic conditions? Proteins *41*, 415–427.

Vacic, V., Oldfield, C.J., Mohan, A., Radivojac, P., Cortese, M.S., Uversky, V.N., and Dunker, A.K. (2007). Characterization of molecular recognition features, MoRFs, and their binding partners. J. Proteome Res. *6*, 2351–2366.

Velazquez-Campoy, A., Leavitt, S.A., and Freire, E. (2004). Characterization of protein-protein interactions by isothermal titration calorimetry. Methods Mol. Biol. *261*, 35–54.

Weatheritt, R.J., and Gibson, T.J. (2012). Linear motifs: lost in (pre)translation. Trends Biochem. Sci. *37*, 333–341.

Wheatley, N.M., Sundberg, C.D., Gidaniyan, S.D., Cascio, D., and Yeates, T.O. (2014). Structure and identification of a pterin dehydratase-like protein as a RuBisCO assembly factor in the alpha-carboxysome. J. Biol. Chem.

Williams, E.B. (2006). Identification and Characterization of Protein Interactions in The Carboxysome of *Halothiobacillus Neapolitanus*. Dissertation 1–194.

Xue, B., Dunbrack, R.L., Williams, R.W., Dunker, A.K., and Uversky, V.N. (2010). PONDR-FIT: a meta-predictor of intrinsically disordered amino acids. Biochim. Biophys. Acta *1804*, 996–1010.

Chapter 4 Conclusions

4.1 Summary

The exact molecular mechanism of the α-carboxysome assembly has been an enigma in the study of bacterial microcompartments for the past several decades. At the center of this conundrum is the protein known as CsoS2, whose intractable behaviors defy attempts to biochemically characterize its function. In this work, we have elucidated the co-translational mechanism involved in CsoS2 biogenesis and an important protein-protein interaction that CsoS2 likely employs to organize *H. neapolitanus* carboxysome assembly. In Chapter 2, we demonstrated that -1 programmed ribosomal frameshifting caused approximately 50% of CsoS2 to possess C-terminal truncation. A slippery sequence and a stem-loop were found to be necessary and sufficient for -1PRF, and perturbation of these elements allowed us to express CsoS2A and CsoS2B individually *in vivo*. Additionally, bioinformatics predicted that more than half of the CsoS2 homologs discovered thus far may undergo this cotranslational event, leading us to hypothesize that both isoforms were essential for α-carboxysome assembly. However, CsoS2A was deemed unnecessary for the expression of carboxysome both in *E. coli* expression system and *H. neapolitanus*. Although carboxysomes were smaller in the absence of CsoS2A, they were not functionally compromised and were able to support the growth of *H. neapolitanus* in air. We therefore proposed a model that both isoforms could interact with RuBisCO via their shared N-terminus, while the C-terminal domain that links shell to RuBisCO matrix is unique to CsoS2B.

In Chapter 3, we experimentally confirmed that both CsoS2A and B interact with RuBisCO via their N-terminal domain (NTD). The combination of qualitative and quantitative strategies demonstrated that NTD was capable of strong multivalent interaction and rapid RuBisCO nucleation, and that the binding competency was attributable to four short amphipathic α-helices in NTD. Substitution of a highly conserved leucine and two conserved arginine/lysine residues in each helix with alanine led to compromised RuBisCO binding, indicating that both hydrophobic and electrostatic interactions were essential NTD's function. Notably, we discovered that fluorescently-labeled CsoS2B was able to self-associate into smooth spherical particles in a reductant-dependent fashion. While we initially linked the particle's droplet-like appearance to liquid-liquid phase separation often involved in IDP-mediated assembly of eukaryotic membraneless organelles, slow recovery after photobleaching and resistance to aliphatic alcohols ruled out this possibility.

4.2 Discussion

Our ability to produce CsoS2A and B separately *in vivo* led us to discover that only CsoS2B was required in the assembly of carboxysomes in both native and heterologous hosts. The true necessity of -1 PRF in carboxysome biogenesis, therefore, still remains an open question. It is possible that CsoS2A may be crucial for the integrity or performance of carboxysomes under certain stress conditions, such as sub-atmospheric level of CO_2, high salt, and high oxidative stress that the organism experiences in its natural habitat. Alternatively, as α-carboxysome shares surprisingly many characteristics with viruses, including icosahedral shell, the use of -1 PRF (Brierley, 1995), and the prevalence of IDPs (Xue et al., 2014), the production of CsoS2A via -1 PRF may be the relic from the viral ancestor of α-carboxysomes that no longer serves any important functions. Comparison of the CO_2-fixation efficiency and morphology between α-carboxysomes that consist of only one CsoS2 isoform and those that possess two isoforms may help shed light on the necessity of multiple isoforms

Programmed frameshifting controlled by cis-acting elements generally results in the fixed stoichiometry between two products, which raises this raises the question of how the ratio between CsoS2A and CsoS2B can be adjusted. While this lack of stoichiometric flexibility may suggest that the carboxysome can be compromised by the unbalanced levels of the two isoforms, the fact that we could express carboxysomes without CsoS2A argues against this possibility. In certain systems, frameshifting can also be modulated by trans-acting elements (Li et al., 2014). While we did not find any evidence of trans-regulation of CsoS2 expression in our work, we could not completely rule out the presence of such mechanism. Interestingly, we observed that coexpression of RuBisCO and CsoS2 led to a strong bias towards CsoS2A, and thus RuBisCO may act as a trans-frameshifting regulator in regard. In the context of engineering, should there be a need adjust the ratio of CsoS2A to CsoS2B to alter the property of the nanoreactor, a first strategy to consider is to alter the slippery sequence and the stem-loop. Screening different types of slippery sequence and stem-loop may be necessary, as there has not been a convenient way to predict what kind of changes on the frameshifting elements would yield the desired frameshifting efficiency (Mouzakis et al., 2013).

In Chapter 3, the success of our biochemical efforts to elucidate the function of CsoS2 largely relied on the integrity of this protein, and this required us to develop strategies for dealing with intractable qualities of IDPs. Initially, we believed that an interaction assay that did not involve purification of CsoS2 would allow us to circumvent self-aggregation and proteolysis . However, as observed in our *in vivo* pull-down experiment, CsoS2 still suffered from these problems despite coexpression with RuBisCO. High expression of CsoS2, in addition, might have given rise to non-specific binding with RuBisCO, which would explain the discrepancy observed with MR (Figure 3-7). Therefore, we refocused our effort to improve the behavior of CsoS2 *in vitro*. We successfully minimized proteolysis and aggregation

by employing direct purification of CsoS2 from carboxysomes and renaturing the protein *in situ* during the BLI experiment. The caveat of this strategy is that the CsoS2 variant to be purified must allow the formation of carboxysomes, thus not applicable to variants that lack critical domains such as NTD and CTD. However, smaller domains of CsoS2 appear to suffer significantly less from proteolysis than the full CsoS2, and direct purification from carboxysomes could presumably be omitted. Alternatively, simultaneous inclusion of orthogonal affinity tags on the N- and C- termini of CsoS2 followed by two-step denaturing purification (Tagwerker et al., 2006) should allow only the intact protein to be isolated. Future work with CsoS2 or other IDPs involved in BMC assembly should consider implementing the strategies we described thus far to minimize potential problems in downstream experiments.

Based on the findings from Chapter 3, we hereby propose a putative model that explains the role of CsoS2 in the assembly and maturation of *H. neapolitanus* carboxysome (Figure 4-1A). It should be noted that, while we portray our model as a series of assembly events, the actual pathway may not proceed in the exact sequence shown. Instead, our model should be treated as a conceptual framework for further mechanistic investigations, not an attempt to elaborate the actual microscopic events.

In the first step, shell protomers assemble into a facet onto which CsoS2B can localize. Accumultation of CsoS2B gives rise to NTD clusters (previously portrayed in Figure 3-9B), which can nucleate RuBisCO with high affinity and rapidly initiate cargo accretion. After the first layer of RuBisCO has formed, CsoS2A can be recruited to the growing matrix via the interaction with RuBisCO. As multiple layers of RuBisCO exist, there is a question of how RuBisCO in subsequent layers could be recruited, considering that the NTDs of CsoS2A molecules have already been used in binding the outer layer of RuBisCO. There are two possible ways that the incoming RuBisCO can dock to the already-bound CsoS2A: (i) Bound CsoS2A may associate to another unbound CsoS2A via MR; or (ii) MR can serve as a secondary binding domain for RuBisCO. The absence of BLI response from MR suggests that the second possibility is less likely, but it cannot be ruled out until the discrepancy between the BLI and *in vivo* pull-down results is resolved.

The growing RuBisCO-CsoS2 may be in the form of dynamic phase-separated liquid, which would allow for influx and efflux of carboxysomal components. Our attempt to characterize the CsoS2 spherical particles did not support the existence of this liquid-like state, but further refinement of CsoS2 expression protocol may allow this intermediate to be observed. After the accreted cargo is enveloped by the shell and becomes a closed structure, the interior may be "cured" by disulfide crosslinking between CsoS2 and other carboxysomal proteins. This scenario is possible considering that the lumen of a carboxysome is oxidizing (Chen et al., 2013), which may allow spontaneous oxidation of cysteines on cargo proteins. In fact, cysteine oxidation was shown to be critical for γ-carbonic anhydrase in β-carboxysomes from

Thermosynechococcus elongatus BP-1 (Peña et al., 2010). Similarly, a number of conserved cysteines on CsoS2 (Fig 3-2C) may participate in disulfide bonding with nearby CsoS2 molecules or other carboxysomal proteins. Notably, we also found that the spherical structures of CsoS2B-sfGFP became branched after the reducing agent was removed, and this transformation might be related to the events that take place after the closing of the shell.

Although the benefit of the post-assembly curing is still currently unclear, we hypothesize that it may serve as a fail-safe of the carboxysome assembly. During the course of the assembly, an error may occur such that formation of closed structure is not possible (Figure 4-1B). Without the complete enclosure by the shell, the assembly intermediate is accessible to cellular reductants and thus cannot form disulfide linkages. Therefore, the complex can continue to disassemble and reassemble until a correct intermediate that allows for shell closure is made. Then, upon the complete encapsulation, the lumen becomes oxidizing, which triggers the formation of disulfide network and stabilization of the interior.

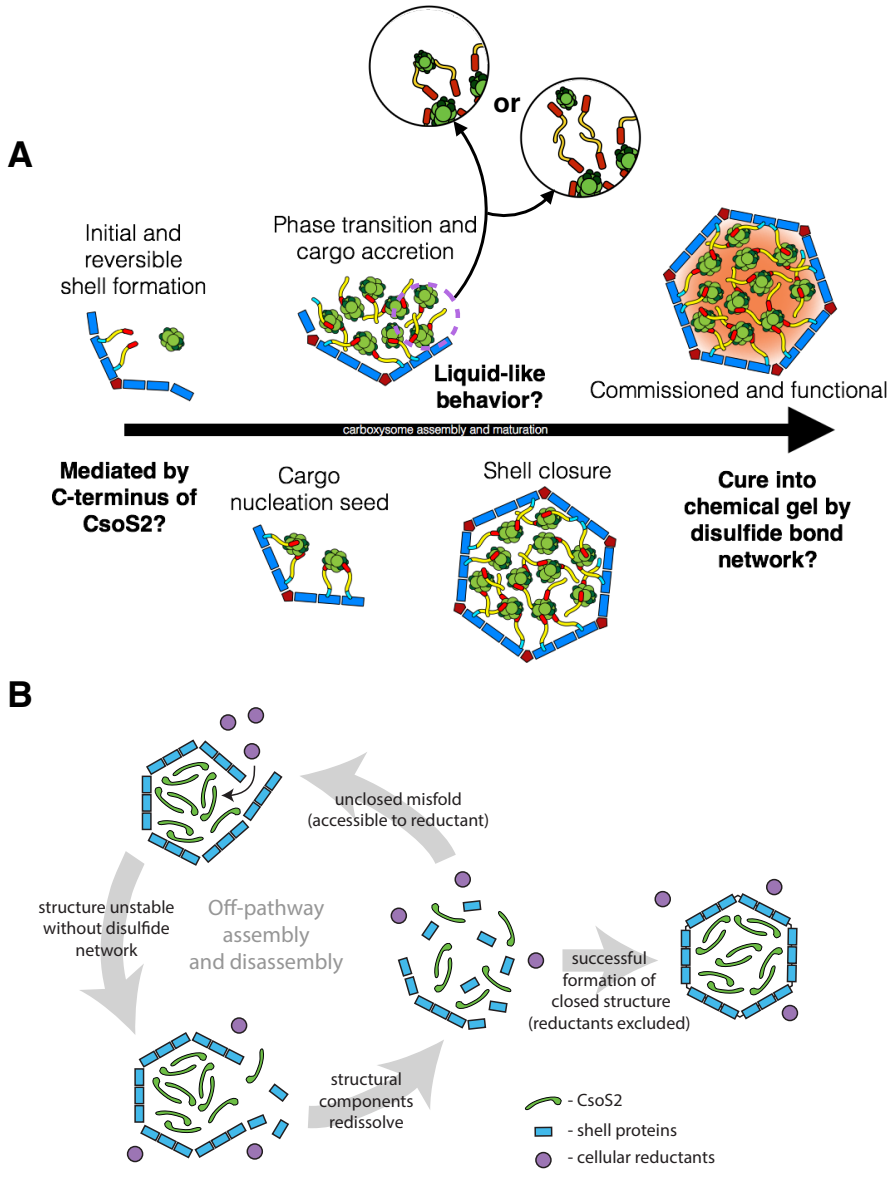

Figure 4-1: A speculative model of the α-carboxysome assembly pathway
(A) Diagram depicting stages of carboxysome assembly and maturation based on the findings reported in this work thus far. First, shell protomers are assembled into facets, allowing CsoS2B to localize on the lumenal surface possibly via the C-terminus. High density of CsoS2B on the surface enables the closely located NTDs to cooperate in seeding RuBisCO nucleation, and subsequent inward filling of RuBisCO is mediated by CsoS2A. This assembly-in-progress may be in the form of phase-separated liquid which is highly capable of protein recruitment and malleable enough to allow compositional adjustment. Once the assembly matrix is completely enveloped by the icosahedral shell, its internal structure may be cured, possibly by disulfide crosslinking, into a final fixed structure. (B) Schematic showing the hypothetical error correction mechanism during the carboxysome assembly. Prior to the complete enclosure by the shell, the assembly matrix is accessible to cellular reductants which prevent intermolecular disulfide crosslinking. Assembly errors that result in incorrect geometry of a carboxysome will forbid the successful formation of closed particle and subsequent disulfide network formation, allowing the assembly intermediate to disassemble and reassemble until the errors are rectified.

4.3 Future Directions

Besides the subject of RuBisCO-CsoS2 interaction, there are also many questions that are beyond the scope of our study, but would be crucial in solving the problem of carboxysome biogenesis. For instance, there are seven other protomers in CB whose ability to interact with CsoS2 has never been evaluated. Our opinion is that the next candidate for future studies should be carbonic anhydrase, as it is the only other enzymatic cargo besides RuBisCO and has been shown to associate to CcmM in *Synechocystis* sp. strain PCC6803 (Cot et al., 2008). In addition, besides pairwise studies, there is value in understanding the level of interdependence between different protein-protein interactions within the CsoS2-centric interaction network. Would the binding of one protein to CsoS2 allosterically affect CsoS2's interaction with another partner (Ferreon et al., 2013)? This aspect is especially worth investigating for an IDP like CsoS2 considering its high conformational flexibility. We are also interested to understand how the size of a carboxysome is controlled. Is the size determinant encoded into the structure of CsoS2, in a similar manner to the so-called "molecular ruler (Marshall, 2004)"? The answer to this question may shed light on the reason why the size of CB varies among organisms and grant us a tool for altering the size of the microcompartment for engineering purposes.

As stated in the introduction, engineering of bacterial microcompartments would immensely benefit from fundamental understanding of their properties. Although we are still far from completely unraveling the assembly mechanism of the carboxysome, we would like to offer a few pointers for engineering applications. The use of minimal nanoreactors, with as few types of building blocks as possible, would help simplify the future engineering applications and reduce the metabolic burden on the host (Gonzalez-Esquer et al., 2015). We have found that CsoS2A was not necessary for the formation of the carboxysome, and getting rid of CsoS2A via perturbation of frameshifting may provide certain benefits, such as more space availability for foreign cargo or less resource consumption to make one particle of carboxysome. However, as removal of CsoS2A decreased the diameter of carboxysomes by almost 30 nm, this strategy may not be appropriate if the goal is to increase the size of the carboxysome, although it could serve as a good strategy for reducing the size if the application demands. In contrast, deletion of the entire NTD or CTD (which is equivalent to removing CsoS2B) would be ill-advised, as we found that the absence of these domains completely abolished carboxysome formation. Partial truncation of these domains may be viable, although in our case the removal of two repeats from NTD dramatically impacted the morphology of the particles.

While the biological encapsulation signal used by carboxysomes to package their cargos has not been delineated, we may take advantage of the interaction between CsoS2 and carboxysomal proteins to develop a non-biological encapsulation tag. Considering that NTD interacts with RuBisCO, peptides derived from this domain may serve as an encapsulation tag that allows a heterologous cargo to be targeted

into carboxysome, similar to that in Pdu and Eut. Alternatively, as NTD contains many positively-charged side chains that have been shown to be important for RuBisCO binding, a synthetic peptide with similar total charges could also be tested. A caveat of the NTD-dependent strategy is that RuBisCO must be present, and thus there may be less capacity for the cargo of interest, or RuBisCO could interfere with the activity of the desired pathway. An alternative would be to use a different domain that targets the shell, which could be the CTD if our putative model is proven correct.

One of the primary goals of studying the assembly of the carboxysome is to help us develop a strategy to assemble it *in vitro* from individual protomers. We have demonstrated that NTD, and in turn CsoS2, is able to engage in strong multivalent binding when tethered to the surface, which is in accordance with the shell-first assembly model. Therefore, a hypothetical way to initiate the assembly *in vitro* is to allow the shell protomers to assemble into a facet first, followed by addition of CsoS2 which will be recruited to the shell facet, and then RuBisCO. To prevent dead-end intermediates, such as two shell facets that are bridged to each other via CsoS2, the first set of shell facets may be anchored to a functionalized surface, and more can be added later after the assembly has proceeded beyond a certain stage. In addition, while we cannot confirm that CsoS2 on its own behaves like a liquid, this possibility still cannot be ruled out. We found that a reducing agent such as TCEP enabled CsoS2 to attain spherical morphology resembling a droplet. While these droplet-like structures displayed no other liquid properties, a reducing condition still likely would be one of the requisites for liquid-liquid separation of CsoS2. Finally, we have demonstrated that CsoS2 can be denatured and renatured without being functionally compromised. This may serve as a strategy to prevent CsoS2 from self-aggregating until it is needed in an appropriate assembly stage. In addition, certain applications, such as encapsulation of metal catalyst, require the compartment to be disassembled and then reconstituted after the functionalization. This method would not work if proteins with a difficult folding pathway like RuBisCO are present in the carboxysome. However, if a minimal carboxysome could be made from only CsoS2 and the shell, it may be possible to disassemble the compartment via denaturation and then reassemble. Therefore, constructing such bare-boned carboxysome should be made the ultimate goal in minimizing the number of carboxysomal components.

In summary, we have unveiled the mechanism that governs CsoS2 isoform generation and took the important first step to shed light on how CsoS2 employs its inherent properties as an IDP to build an entire carboxysome from smaller building blocks. This is the first time that -1 PRF and an IDP, both of which are prevalent in viruses, were found to participate in the construction of a bacterial microcompartment. These unprecedented findings should remind us that, despite many shared characteristics among different types of bacterial microcompartments, assembly mechanisms at play are diverse and a one-size-fits-all approach in engineering different BMCs may not be effective. Therefore, future engineering

efforts would benefit from understanding both the universal and divergent aspects among bacterial microcompartments. While this work may be a small piece in the grand puzzle of α-carboxysome assembly, we hope that the insights from investigating the molecular details of CsoS2 isoform biogenesis and CsoS2-RuBisCO interaction will pave the way for studying the rest of CsoS2-mediated interaction network in the *H. neapolitanus* carboxysome in the future.

4.4 References

Brierley, I. (1995). Ribosomal frameshifting on viral RNAs. J. Gen. Virol. *76*, 1885–1892.

Chen, A.H., Robinson-Mosher, A., Savage, D.F., Silver, P.A., and Polka, J.K. (2013). The bacterial carbon-fixing organelle is formed by shell envelopment of preassembled cargo. PLoS ONE *8*, e76127.

Cot, S.S.-W., So, A.K.-C., and Espie, G.S. (2008). A multiprotein bicarbonate dehydration complex essential to carboxysome function in cyanobacteria. J. Bacteriol. *190*, 936–945.

Ferreon, A.C.M., Ferreon, J.C., Wright, P.E., and Deniz, A.A. (2013). Modulation of allostery by protein intrinsic disorder. Nature *498*, 390–394.

Gonzalez-Esquer, C.R., Shubitowski, T.B., and Kerfeld, C.A. (2015). Streamlined construction of the cyanobacterial CO_2-fixing organelle via protein domain fusions for use in plant synthetic biology. Plant Cell.

Li, Y., Treffers, E.E., Napthine, S., Tas, A., Zhu, L., Sun, Z., Bell, S., Mark, B.L., van Veelen, P.A., van Hemert, M.J., et al. (2014). Transactivation of programmed ribosomal frameshifting by a viral protein. Proc. Natl. Acad. Sci. U.S.A *111*, E2172–E2181.

Marshall, W.F. (2004). Cellular length control systems. Annu. Rev. Cell Dev. Biol. *20*, 677–693.

Mouzakis, K.D., Lang, A.L., Vander Meulen, K.A., Easterday, P.D., and Butcher, S.E. (2013). HIV-1 frameshift efficiency is primarily determined by the stability of base pairs positioned at the mRNA entrance channel of the ribosome. Nucleic Acids Res. *41*, 1901–1913.

Peña, K.L., Castel, S.E., de Araujo, C., Espie, G.S., and Kimber, M.S. (2010). Structural basis of the oxidative activation of the carboxysomal gamma-carbonic anhydrase, CcmM. Proc. Natl. Acad. Sci. U.S.A *107*, 2455–2460.

Tagwerker, C., Flick, K., Cui, M., Guerrero, C., Dou, Y., Auer, B., Baldi, P., Huang, L., and Kaiser, P. (2006). A tandem affinity tag for two-step purification under fully

denaturing conditions. Mol. Cell. Proteomics 737–748.

Xue, B., Blocquel, D., Habchi, J., Uversky, A.V., Kurgan, L., Uversky, V.N., and Longhi, S. (2014). Structural disorder in viral proteins. Chem. Rev. *114*, 6880–6911.