

CRENARCHAEAL VIRUS-HOST SYSTEMS: STRUCTURE-FUNCTION STUDIES
OF CRENARCHAEAL VIRUSES AND PROKARYOTIC ADAPTIVE IMMUNITY

by

Nathanael Greenleaf Lintner

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

Doctor of Philosophy

in

Biochemistry

MONTANA STATE UNIVERSITY
Bozeman, MT

March 2011

©COPYRIGHT

by

Nathanael Greenleaf Lintner

2011

All Rights Reserved

APPROVAL

of a dissertation submitted by

Nathanael Greenleaf Lintner

This dissertation has been read by each member of the dissertation committee and has been found to be satisfactory regarding content, English usage, format, citations, bibliographic style and consistency and is ready for submission to the Division of Graduate Education.

Dr. C. Martin Lawrence

Approved for the Department of Chemistry and Biochemistry

Dr. David J. Singel

Approved for the Division of Graduate Education

Dr. Carl A. Fox

STATEMENT OF PERMISSION TO USE

In presenting this dissertation in partial fulfillment of the requirements for a doctoral degree at Montana State University, I agree that the library shall make it available to borrowers under the rules of the library. I further agree that copying of this dissertation is allowable only for scholarly purposes, consistent with "fair use" as prescribe in the U.S. Copyright Law. Requests for extensive copying or reproduction of this dissertation should be referred to ProQuest Information and Learning, 300 North Zeeb Road, Ann Arbor, Michigan 48106, to whom I have granted "the exclusive right to reproduce and distribute my dissertation in and from microform along with the non-exclusive right to reproduce and distribute my abstract in any format in whole or in part."

Nathanael Greenleaf Lintner

March 2011

ACKNOWLEDGEMENTS

The achievement represented by this thesis would not have been possible without the help and support of numerous mentors, teachers, co-workers, collaborators and family members. Particular thanks to my advisor, Dr. Martin Lawrence, Dr. Valérie Copié, and to the rest of my doctoral committee. I thank all present and former members of the Lawrence research group with particular thanks to Dr. Eric Larson, Dr. George Gauss and Brian Eilers. I thank my scientific collaborators on these projects including Melina Kerou, Dr. Malcolm F. White, Dr. Kenneth A. Frankel, Dr. Susan E. Tsutakawa and Dr. John A. Tainer. I also thank my many supportive family members including my mother, Laurie Lintner, my Uncles Bruce and Edwin Selyem, my Aunts Barbara Selyem, Elizabeth Bustamante and Holly Lane, and my father, William Lintner who always firmly believed in the value of education.

TABLE OF CONTENTS

1. INTRODUCTION	1
Archaea: The Third Domain of Life	1
Crenarchaeal Viruses	3
Sulfolobus Turreted Icosahedral Viruses (STIV)	5
The Host Response: The Prokaryotic Adaptive Immune System	9
CRISPR-associated Genes	10
The Adaptation Phase of CRISPR-Cas	11
The Interference Phase of CRISPR-Cas	12
Regulation of CRISPR-Cas	17
The <i>Sulfolobus solfataricus</i> CRISPR-Cas System	18
Research Goals	19
2. MATERIALS AND METHODS	20
Introduction	20
Materials and Methods for STIV-A81	20
Expression, Purification and Crystallization of Native A81	20
Generation of Leucine to Methionine Mutants for Phasing	21
Crystallization and Data Collection	21
Structure Determination and Refinement	22
Screening of <i>S. solfataricus</i> CRISPR-Associated Proteins for Suitable	
Structural Targets	24
Selection of Targets for Structural Studies	25
Cloning	26
Small-scale Screening for Soluble Protein Expression	30
Crystallization Trials	33
Identification of Protein and Nucleic Acid Binding Partners	34
Transformation and Protein Expression in <i>S. solfataricus</i>	36
Co-purification of Cas Proteins and Putative Binding Partners	38
Specific Methods for Csa3	40
Expression and Purification	40
Crystallization and Data Collection	41
Structure Determination and Refinement	41
Small-Angle X-ray Scattering Data Collection and Processing	44
Specific Methods for aCASCADE and Csa2	46
Expression of Csa2 in <i>S. solfataricus</i> and Isolation of A-CASCADE	46
Cloning and Sequencing of Small RNA's	46
Expression and Purification of Recombinant Csa2	47
Crystallization and Data Collection	49
Structure Determination and Refinement	50

TABLE OF CONTENTS – CONTINUED

Conclusions and Future Directions	52
3. THE STRUCTURE OF STIV-A81	54
Introduction	54
Results	55
Discussion, Conclusions and Future Directions	66
4. INSIGHT INTO CRISPR-CAS REGULATION FROM THE STRUCTURE OF CSA3	64
Introduction	64
Results	64
Structure of the Csa3 Protomer	65
Small-Angle X-ray Scattering Identifies the Csa3 Solution-state Dimer ..	66
Structure of the Csa3 Dimer	70
Potential Csa3/DNA Interactions	72
A Putative Ligand Binding Site on the N-terminal Domain	75
Structural Homology to the Csx1 Family of CRISPR-Associated Proteins	80
Discussion and Concluding Remarks	87
5. ISOLATION AND CHARACTERIZATION OF AN ARCHAEAL CASCADE	90
Introduction	90
Results	90
Isolation of a CASCADE-like Protein-RNA Complex from <i>S. solfataricus</i>	90
Recombinant <i>S. solfataricus</i> Cas6 Generates RNA Fragments Identical to Those in aCASCADE	96
Investigation of the Core CASCADE subunits	97
The Csa2-Cas5a Complex Binds crRNA and Forms Ternary Complexes with Target DNA	99
Structural Studies of aCASCADE	101
The Structure of Csa2	103
Discussion	113
<i>S. solfataricus</i> Cas6	113
Structural Models for A-CASCADE	114
CRISPR-Mediated Viral Defense in the Archaea	120

TABLE OF CONTENTS – CONTINUED

6. CONCLUDING REMARKS AND FUTURE DIRECTIONS.....	122
7. PRESENTATION OF STRUCTURAL DATA USING THE 3-D PDF FORMAT	129
Generating Input Files with Pymol	129
Setting up 3D Content with Adobe 3D Reviewer.....	130
Incorporating 3D Content into Documents with Adobe 9 Pro Extended	131
Discussion	
REFERENCES CITED.....	135
APPENDIX A: Pymol Script for the Generation of 3D pdf's	148

LIST OF TABLES

Table	Page
2-1 Mutagenic Primers for A81 Leu to Met Mutants.....	21
2-2 STIV-A81 Data Collection	23
2-3 STIV-A81 Model Refinement	24
2-4 Cas Proteins Selected for Small-Scale Expression Trials.....	26
2-5 Cloning the <i>S. solfataricus</i> <i>cas</i> Genes for Expression in <i>E. coli</i>	29
2-6 Results of Small-Scale Expression and Purification Trials	32
2-7 Primers Used to Generate Constructs for Expression in <i>S. solfataricus</i>	36
2-8 Csa3 Data Collection Statistics	43
2-9 Csa3 Model Refinement Statistics	44
2-10 Csa2 Data Collection Statistics	51
2-11 Csa2 Model Refinement Statistics	52
5-1 Identification of High-Abundance Proteins in aCASCADE	96
5-2 Proteins Co-Purifying with Csa2 Identified by LC-MS/MS	96
5-3 Oligonucleotides Used in Csa2/Cas5a EMSAs	100

LIST OF FIGURES

Figure	Page
1-1	Historic and Modern View of Biological Diversity.....2
1-2	Morphological Diversity of Crenarchaeal Viruses.....4
1-3	Cryo-EM Reconstruction of STIV.....5
1-4	Genome Map of STIV17
1-5	Various CRISPR-Processing Endonucleases.....14
1-6	Proposed Subunit Architecture of <i>E. coli</i> CASCADE.....15
1-7	CasC- and CasD-like Orthologs.....16
1-8	Overview of the CRISPR-Cas System.....18
1-9	The <i>S. solfataricus</i> CRISPR-Cas System.....19
2-1	Crystals of <i>S. solfataricus</i> Cas Proteins34
2-2	Post-transfection Growth of <i>S. solfataricus</i> PH1-16.....37
3-1	Self-Rotation Functions of A8156
3-2	The Structure of STIV-A8158
3-3	Other Examples of Oligomeric Ring Structures61
3-4	Surface Electrostatics of Sliding-Clamp Proteins.....62
4-1	The Structure of Csa367
4-2	Surface Electrostatics of Csa368
4-3	Guinier Plot of Csa3 SAXS68
4-4	Validation of the Csa3 Biological Unit by SAXS70
4-5	The Dimer Interface of Csa3.....72
4-6	Superpositional Docking of DNA to Csa3.....74
4-7	Multiple-Sequence Alignment of Csa3 Orthologs.....77
4-8	Putative Ligand-Binding Pocket on the Csa3 N-terminal Domain.....79
4-9	Structural Alignments of Csa3 with VC1899 and Csx181
4-10	Domain Architectures of Csa3, VC1899 and Csx183
4-11	Multiple Sequence Alignment of Csx1 Sequences.....85

LIST OF FIGURES – CONTINUED

Figure	Page
4-12 Conserved Clefts on Csx1	86
5-1 Isolation and Initial Characterization of A-CASCADE.....	94
5-2 Behavior of aCASCADE on a Superose-6 Column	96
5-3 <i>S. solfataricus</i> Cas6 Generates crRNA	98
5-4 Target-DNA Binding by Recombinant Csa2/Cas5a/crRNA Complex....	101
5-5 The Structure of Csa2	107
5-6 The Csa2 Fold	108
5-7 Multiple Sequence Alignment of Csa2 Orthologues	109
5-8 Conformational Flexibility of the Csa2 1-3 Domain	110
5-9 Structural Comparison of Csa2 and Cas6	112
5-10 Hypothetical Csa2 Hexamer	116
5-11 Preliminary Structural Model for aCASCADE	118
5-12 Distribution of CRISPR Spacer Lengths	120
5-13 The Two Arms of CRISPR-Mediated Viral Defense in Archaea.....	121
7-1 Interactive 3D PDF of the Csa3 Structure	134

ABSTRACT

Virus-host interactions are one of the most important drivers of microbial ecology and evolution. Among the three domains of life, the least is known about viruses infecting members of the domain Archaea. This work combines a traditional hypothesis-driven molecular and biochemical approach with a structural genomics-like approach of pursuing a large number of structural targets to gain a detailed molecular understanding of a model Crenarchaeal virus-host system, *Sulfolobus* turreted icosahedral virus (STIV) and its *Sulfolobus* host. This work specifically focuses on the viral protein, A81, a putative transcriptional regulator associated with the prokaryotic adaptive immune system, CRISPR-Cas, and a large protein-RNA complex involved in CRISPR-mediated DNA interference.

A81 is a protein-of-unknown-function encoded by the STIV genome. The structure of STIV-A81 reveals a unique ring-shaped octameric assembly. While structural homology-based searches fail to reveal a function for A81, the central pore has a size and charge consistent with a potential interaction with single-stranded nucleic acid.

The structure of the CRISPR-associated protein, Csa3 reveals a putative 2-domain transcription factor. The N-terminal domain is a variation on the di-nucleotide binding-domain that orchestrates dimer formation. There is a conserved 2-fold symmetric pocket on the dimer axis that likely represents a regulatory ligand-binding site implying a small-molecule regulator of CRISPR/Cas. The C-terminal domain is a winged helix-turn-helix common among transcription factors. The domain architecture of Csa3 suggests a small molecule regulator of CRISPR/Cas in the Crenarchaea.

The CRISPR-associated complex for antiviral defense (CASCADE) is predicted to be central to CRISPR-mediated DNA-interference in many bacteria and Archaea. We isolated components of an archaeal CASCADE from *Sulfolobus solfataricus*. Csa2 expressed in *S. solfataricus* co-purifies with Cas5a, Cas6, Csa5 and Cas6-processed CRISPR-RNA (crRNA). Csa2, the dominant protein, forms a stable complex with Cas5a. A recombinant Csa2-Cas5a-complex is sufficient to bind crRNA and complementary ssDNA. Transmission electron microscopy reveals an extended helical complex of variable length, perhaps due to substoichiometric amounts of capping factors. Csa2 displays a crescent-shaped fold including a modified RNA-recognition motif (RRM) plus two additional domains present as insertions into the RRM. A preliminary model for this and other CASCADEs is proposed.

INTRODUCTION

Virus-host interactions are one of the most important drivers of evolution on the planet. Viruses are thought to infect every known cellular organism and are the most abundant biological entity on the planet. Viruses, thus, have a profound effect on the size and diversity of microbial populations, the transfer of genetic material and nutrient cycling (1). Among the three domains of life, viruses that infect members of the Archaea are the least well-studied.

The work presented herein presents a structure-function based approach to understanding the molecular details of a model Crenarchaeal virus-host system, *Sulfolobus* Turreted Icosahedral Virus (STIV) and its host is *Sulfolobus solfataricus*. In the structure-function approach, structural models of key components are obtained and used to inform hypothesis-driven experiments. This approach is applied to both virally-encoded components of the system and a key host response system, the recently discovered prokaryotic adaptive immune system.

Archaea: The Third Domain of Life.

The advances in molecular biology in the latter half of the 20th century have fundamentally changed the prevailing view of the diversity of life on earth and led to an abandonment of the centuries-old “Kingdom-based” classification (2). Specifically, molecular techniques have revealed the great diversity among prokaryotes. Phylogenies based on the genes encoding ribosomal RNA (*rDNA* genes) reveal that life is more appropriately classified into three broad categories called “Domains”. These domains

include the Eukarya, the Bacteria, and a poorly-studied group of organisms called the Archaea (3) (Figure 1-1).

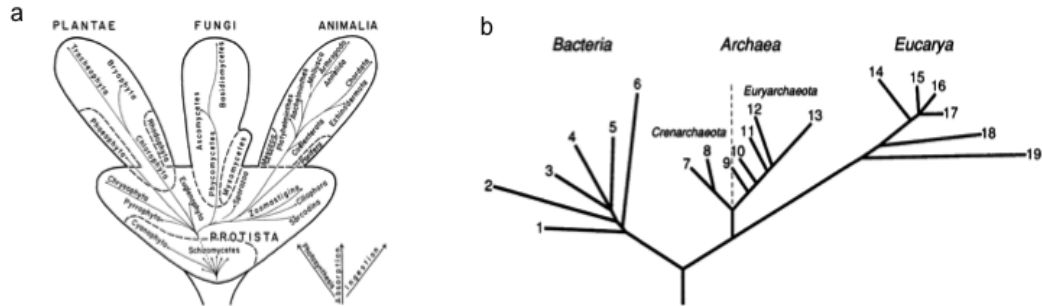


Figure 1-1: Historic and modern views of biological diversity. **(a)** Historical view of biological diversity dividing life into several kingdoms based on easily observed physiological characteristics. **(b)** Modern view dividing life into three domains based on molecular phylogeny of rDNA sequences. [Figure compiled from refs. (2, 3)]

Like Bacteria, Archaea are small-unicellular organisms with genomes that are not found in a membrane-bound nucleus. However, *rDNA*-based phylogenies place the Archaea closer to the Eukarya than to the Bacteria. Archaea also harbor unique versions of the machinery used for key biological processes such as transcription and translation (3).

Archaea were initially classified into two phyla, the Euryarchaeota and the Crenarchaeota (3). Additional discoveries of unique Archaea have led to two additional phyla, the Korarchaeota (4) and the Nanoarchaeota (5). The Euryarchaeota include methanogens, halophiles and thermophiles. The second original phyla, the Crenarchaeota is highly represented in hot acidic environments around the world and is expected to be deeply rooted in the universal tree-of-life (3). Of the Crenarchaeota, species belonging to

the genus *Sulfolobales* have emerged as an important model system due to their ease of culturing (6-19).

Crenarchaeal Viruses

There is a growing interest in the understanding of viruses outside of traditional medical and agricultural fields. One reason is the growing appreciation for the impact of viruses on the biosphere as a whole. The Earth's oceans are estimated to contain about 10^{31} virus particles where they induce the turnover of about 20% of the biomass per day (20, 21). Viruses also play an important role in cellular evolution through horizontal gene transfer (20, 22, 23). Furthermore, the study of viruses often yields genetic tools for the elucidation of host biochemistry and genetics (6, 8, 15, 24-26). Of the 5100 characterized viruses only about 50 infect Archaeal hosts with approximately equal numbers affecting euryarchaeotes and crenarchaeotes (15).

The first Archaeal viruses identified by Wolfram Zillig *et al.* infected members of the Euryarcheota and principally displayed the head-tail morphologies of common bacteriophages (20, 27). However, initial studies of Crenarchaeal viruses identified a surprising diversity both on a molecular (28) and morphological level. The morphological families include the spherical *Globuloviridae*, droplet-shaped *Guttaviridae*, particles with two-tails called *Bicaudaviridae*, linear dsDNA viruses which are long flexible (*Lipothrixviridae*), or stiff (*Rudiviridae*) filaments, lemon-shaped viruses called *Fusseloviridae*, and a turreted icosahedral virus not yet assigned to a family. (Figure 1-2) [reviewed in ref. (20)]. This wide diversity of viruses represents only four genera of

hosts, *Acidianus*, *Sulfolobus*, *Theroproteus* and *Pyrobaculum*. Our research group has focused on the *Fusseloviridae* (29-33) and the unclassified turreted icosahedral virus (34-36) infecting members of the *Sulfolobales*.

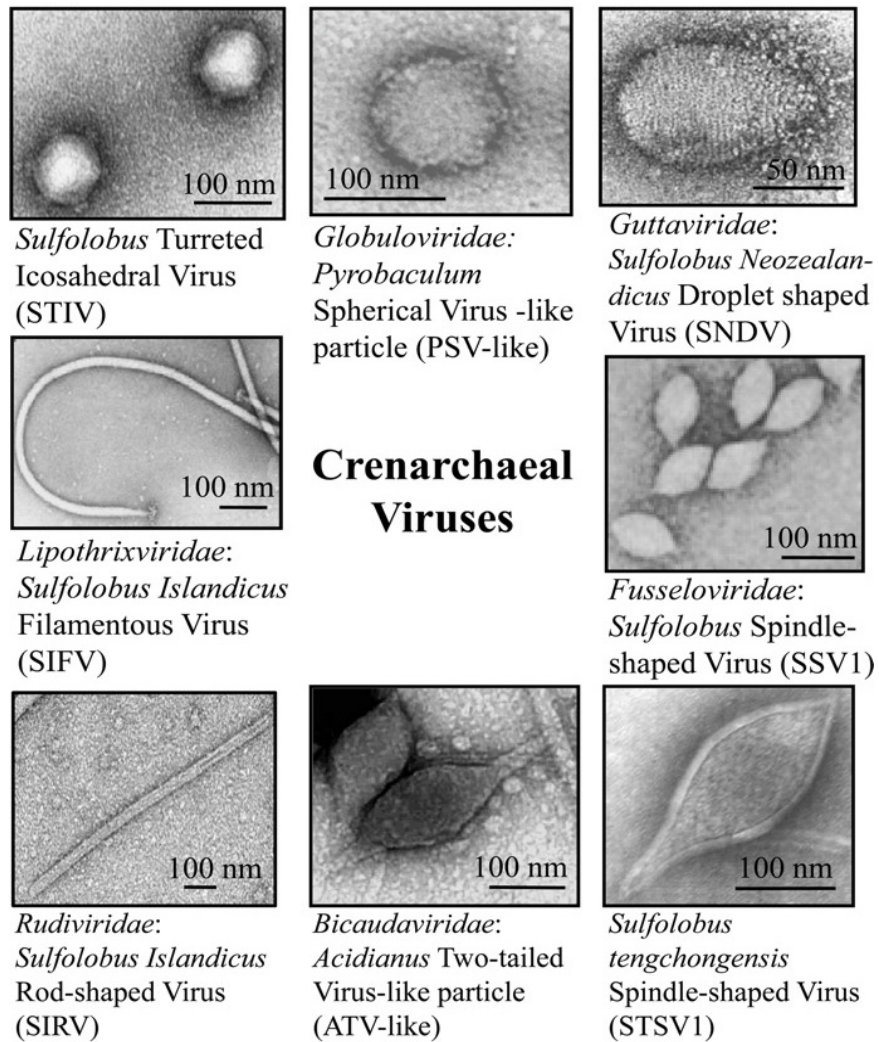


Figure 1-2: A collection of electron micrographs illustrating the morphological diversity of Crenarchaeal viruses. [Figure from ref. (20)].

Sulfolobus Turreted Icosahedral Virus (STIV)

STIV1 and STIV2 are pseudo-T=31 icosahedral viruses. They both display large turret-like projections extending from each of the 5-fold vertices. In STIV1, each of these turrets extend 13 nm from the surface of the virus and are 24 nm in diameter at their broadest point (Figure 1-3) (37). STIV1 was isolated from an acidic, high temperature hot spring (pH=2.2, 80°C) hot spring in Yellowstone National Park, WY, USA from an enrichment culture of an unsequenced *Sulfolobus* isolate, YNPRC179 (37). STIV2 was isolated from hot spring IceG4 (88.3 °C, pH=3.5) in the Hverakjalki valley, Iceland using an enrichment culture with *Sulfolobus* strain G4ST-2 as the host (10).

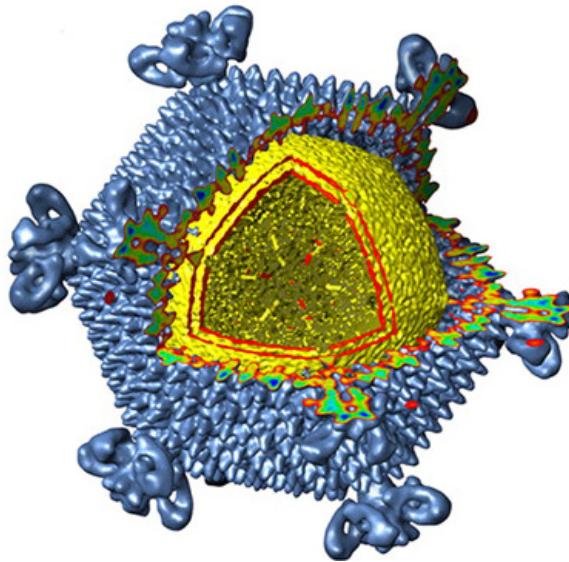


Figure 1-3: Cryo-EM reconstruction at 2.7 nm resolution of STIV1. The figure is false-colored with the capsid and turret-like projections in blue and the internal lipid layer in yellow [Figure from ref (37).]

The STIV1 virus capsid encloses a 17,663 bp circular double-stranded DNA genome with 37 predicted open reading frames (ORF's) (37, 38) and STIV2 has a smaller 16,622 bp circular dsDNA genome with 34 predicted ORF's (10). Comparative genomic analysis found that STIV1 and STIV2 had 25 ORFs in common based on the most recent STIV1 annotation at the time (10). The same study predicted an additional 7 ORF's for STIV1 based on homology to STIV2 ORF's (10). Aside from the similarity between STIV1 and STIV2, Various bioinformatics approaches have revealed few or no homologues to most of the STIV ORF's. For STIV1, ORF B164 is predicted to be an ATPase and ORFs C557, B116 and C92 are homologous only to other hypothetical proteins (38). Similarly, the STIV2 genome encodes two ORFs which are similar to other Crenarchaeal proteins and one ORF, B60, which is predicted to be a ribbon-helix-helix transcriptional regulator (10). Proteomic analysis of purified STIV1 reveals 9 STIV proteins and two host proteins that are present in purified virus. In addition to the major capsid protein, these include A223, C381 and C557, which are predicted to be components of the turret. The predicted ATPase, B164, is also a capsid component and may be involved in DNA packing. B130, B109, A78, A55 and the host proteins Sso7D and Sso0881 are also found in purified capsid but are of unknown function (Figure 1-4)(12).

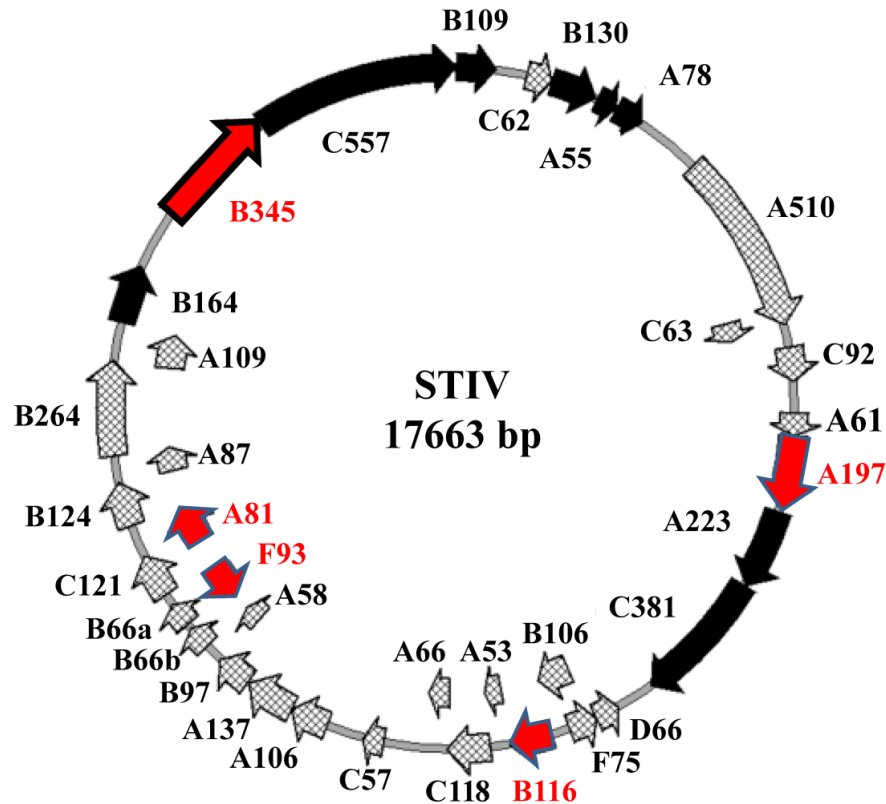


Figure 1-4: Genome map of STIV1. ORF's with products associated with the purified virions are shown in Black and proteins for which a structure is available are red. [Figure modified from ref. (12)]

Because protein tertiary structure evolves at a slower rate than the primary sequence, structural comparison can reveal more distant relationships than sequence comparison. Thus, our research group and others have initiated structural studies of the STIV1 proteome. Prior to the work presented here, our group has solved the structures of F93 (34), B116 (35), and A197 (36). Based on their structures, they appear to be a dimeric winged helix-turn-helix DNA-binding protein (34), a novel DNA-binding protein, which is highly conserved among Crenarchaeal viruses (35), and a glycosyltransferase (36) which may glycosylate the major capsid protein (12). The structure of the major capsid

protein has also been solved and reveals a double beta-barrel fold that is nearly identical to the major capsid proteins of the eukaryotic Adenovirus and *Paramecium bursaria* *Chlorella* virus and the bacteriophage PRD1 (39). The similarity in overall capsid structure and fold of the major capsid proteins suggests that STIV, Adenovirus, *Paramecium bursaria* *Chlorella* virus and PRD1 may belong to an ancient viral lineage that spans all three domains of life (37-39).

STIV1 infection of *Sulfolobus solfataricus* P2 is highly inefficient with only about 10% of the cells susceptible to infection. Thus, for transcriptomics studies, a highly susceptible substrain, P2-2-12, was isolated from the P2 population. Most of the viral genes, including those encoding the capsid-associated proteins, are up-regulated at 16-hours post-transfection. Some intracellular protein-encoding genes are transcribed 8 hours post-transfection but on the whole there are not clear groupings of early and late genes. 177 host genes were differentially regulated during infection (14).

One potential reason for the high resistance of *S. solfataricus* strain P2 to STIV infection may be the discovered prokaryotic adaptive immune system (40-44). The *S. solfataricus* genome encodes six loci called Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs) (11, 45-47), which incorporate short fragments of virus-derived DNA called spacers and utilize them to target the virus in subsequent infection. *S. solfataricus* P2 harbors three such 3 spacers with partial matches to STIV1. CRISPR-B spacer 27 targets C557 (27 out of 40 identical nt are identical), CRISPR F spacer 30 targets between A109 (27 out of 40 nt are identical), CRISPR-F spacer 32 targets C557 (24 of 41 nt are identical) and CRISPR-F spacer 33 targets C557 (23 of 40 nt are identical)

although it has been suggested that CRISPR-F is inactive (11). While these are not perfect matches, it has been shown that perfect matches are not necessary in *S. solfataricus* (48). Although strain P2-2-12 still harbors these spacers targeting STIV1 (38), the transcriptomic analysis no up-regulation of the protein machinery associated with CRISPR-Cas upon STIV1 infection(14). While the reason for the difference between strains P2 and P2-2-12 in susceptibility to STIV1 infection remains unknown, CRISPR-Cas is clearly an important aspect of Bacterial and Archaeal virus-host systems. These studies were thus, expanded to consider the *S. solfataricus* CRISPR-Cas system.

The Host Response: The Prokaryotic Adaptive Immune System

All organisms are targets for infection by selfish genetic elements. Accordingly, they have evolved defensive systems to counteract these pathogens. In higher eukaryotes, defenses include both the innate and adaptive immune systems (44). Systems of innate immunity are also found in prokaryotes, including processes leading to abortive infections and the well-known restriction-modification system. Recently, however, heritable prokaryotic adaptive immune systems have also been recognized; for recent reviews see (42-44, 49, 50). In response to viral infection many prokaryotes incorporate short fragments of virus-derived DNA into genomic loci called Clustered Regularly Interspaced Short Palindromic Repeats (CRISPRs) (44). CRISPR loci consist of a series of small direct repeats (24-47 base pairs) separated by short virus-derived sequences called “spacers” (20-50 base pairs) (45, 51-56). *CRISPR* loci are transcribed and processed into short RNAs, called crRNAs when they include a complete repeat-spacer unit or

prokaryotic silencing RNAs (psiRNA) when they are further shortened. crRNAs and psiRNAs direct the silencing of invading nucleic acids (40, 41) by the CRISPR-associated protein machinery. Recent surveys have identified *CRISPR* loci in 39% of the sequenced bacterial genomes and 88% of the sequenced archaeal genomes (45).

CRISPR-Associated Genes

A significant number of gene families associated with *CRISPR* loci have been identified (57-59). Some of these *CRISPR*-associated (*cas*) genes encode the machinery used to process the *CRISPR*-transcripts, while others are thought to function in the recognition and neutralization of foreign genetic elements or the incorporation of new spacers (44). While the process is reminiscent of eukaryotic RNAi, there are significant differences. The prokaryotic CRISPR/Cas machinery lacks apparent homology to the RNAi protein machinery in both primary sequence and three dimensional structure (7, 60, 61), and CRISPR/Cas appears to target both DNA and RNA (62, 63). *Cas* genes are generally organized in putative operons, which are physically close to the *CRISPR* loci. Co-occurrence patterns for *cas* genes within genomes and gene clusters suggest the Cas machinery takes several different forms. These are alternatively called subtypes by *Haft et al.* (58) or CRISPR-associated systems (CASS) by *Makarova et al.* (59).

A comprehensive study by *Haft et al.* identified 45 CRISPR-associated gene families (58). These include a set of “core” *cas* genes, (*cas1-6*) and 8 groups of subtype-specific gene families (*cse1-4*, *csy1-4*, *csn1-2*, *csd1-2*, *cst1-2*, *csh1-2*, *csa1-5* and *csm1-5*). A given CRISPR/Cas system will encode several of the core Cas proteins plus at least one of

these eight subtypes. In addition, several CRISPR/Cas systems include a third cluster of genes that belong to the Repeat Associated Mysterious Protein (RAMP) superfamily and are named *cmr1-6*. Finally, Haft *et al.* also identify seven additional CRISPR-associated gene families (*csx1-7*), that lack an identified contextual pattern (58).

Similarly, Makarova *et al.* have categorized the CRISPR-associated proteins into 25 families with names loosely based on clusters of orthologous groups (COGs). Subsets of these protein families are then grouped into 7 CRISPR-associated systems (CASS1-7) based on the presence and genomic organization of these COG-based families (59). Functions were predicted for several CRISPR-associated proteins; these include nucleases, helicases and nucleic acid-binding proteins (58, 59). Nuclease activities have now been experimentally confirmed for Cas1, Cas2, Cas3b, Cas6 and Cse3 (7, 9, 41, 60, 61).

The Adaptation Phase of CRISPR-CAS

The adaptation phase of CRISPR-Cas system refers to the process by which viral nucleic acids are processed and incorporated into CRISPRs. This is an active area in CRISPR-Cas research but specific mechanisms have yet to be elucidated. Spacer incorporation appears to be an orderly, polar, process; new spacers are incorporated adjacent to the leader sequence (11, 40, 50, 56). While Cas1 and Cas2 are the most highly conserved Cas proteins (58, 59), in *E. coli* K12, they are unnecessary for viral resistance given a CRISPR that already harbors a targeting spacer (41). These observations thus suggest that Cas1 and Cas2 are involved in the adaptation phase. *Pseudomonas*

aeruginosa Cas1 non-specifically cleaves double and single-stranded DNA (61), *E. coli* Cas1 was shown to have a Holiday junction resolvase activity (64) and *Sulfolobus solfataricus* Cas1 was shown to non-specifically bind single-stranded RNA and DNA (65). *Sulfolobus solfataricus* Cas2 is an endoribonuclease with a preference for U-rich RNA (7). It is difficult, however, to imagine how these reported activities could account for the recognition, generation and incorporation of new CRISPR spacers. Thus, Cas1 and Cas2 must either have undiscovered activities, perhaps in the context of a larger complex, or spacer incorporation must require additional proteins. Many CRISPR-Cas systems include a predicted RecB-like nuclease (59) designated Cas4 or Csa1 by Haft *et al.* (58), which does not appear to be part of the interference machinery and thus a strong candidate for an additional component of the adaptation machinery. Adaptation may also utilize subtype-specific proteins, the Csx proteins, undiscovered CRISPR-associated proteins or may recruit other cellular proteins normally involved in other processes, perhaps DNA repair. In fact, *E. coli* Cas1 has recently been shown to interact with RecB, RecC and RuvB and to promote general DNA repair (64).

The Interference Phase of CRISPR-CAS

In contrast to the adaptation phase of CRISPR-Cas, the general mechanisms of the interference phase are better understood. For all studied systems, the CRISPR locus is transcribed then cleaved once in each repeat sequence by an endoribonuclease to generate small RNAs containing a single virus-targeting “spacer” that is flanked by residual portions of the “repeat” yielding crRNAs. Structures and enzymatic activities have been

reported for the transcript-processing endoribonuclease from three systems. *Thermus thermophilus* Cse3 (66) and *Pyrococcus furiosus* Cas6 (60) both display duplicated ferredoxin-like or RNA-recognition-motif-like (RRM-like) folds. The structure of *Pseudomonas aeruginosa* Csy4 was solved in complex with product RNA and consists of a single RRM-like fold plus a C-terminal domain (Figure 1-5). Surprisingly, the C-terminal domain was principally responsible for the observed protein-RNA interaction (67). *E. coli* Cse3 (41), *Pyrococcus furiosus* Cas6 (60), and *Pseudomonas aeruginosa* Csy4 (67) all display high specificity and cleave their substrate CRISPR-transcripts eight nucleotides upstream from the 3' end of the direct repeat. Among the CRISPR-Cas systems, there appear to be at least three basic system types, which utilize processed crRNA to target invading nucleic acids. Versions of the most common system appear to be found in at least six of the eight subtypes (59), and utilize a large protein-RNA complex called the CRISPR-associated complex for antiviral defense (CASCADE) plus Cas3, a predicted helicase-nuclease fusion (59), to target viral DNA (68, 69). *E. coli* CASCADE is composed of the CasA(Cse1), CasB(Cse2), CasC(Cse4), CasD(Cas5e), CasE(Cse3), and crRNA and each component, along with Cas3, is required for viral resistance *in vivo*. Transmission electron microscopy, small-angle X-ray scattering and non-covalent mass spectrometry reveal that *E. coli* CASCADE has an unusual quaternary structure with six copies of Cse4 forming the structural core of the complex (Figure 1-6) (69).

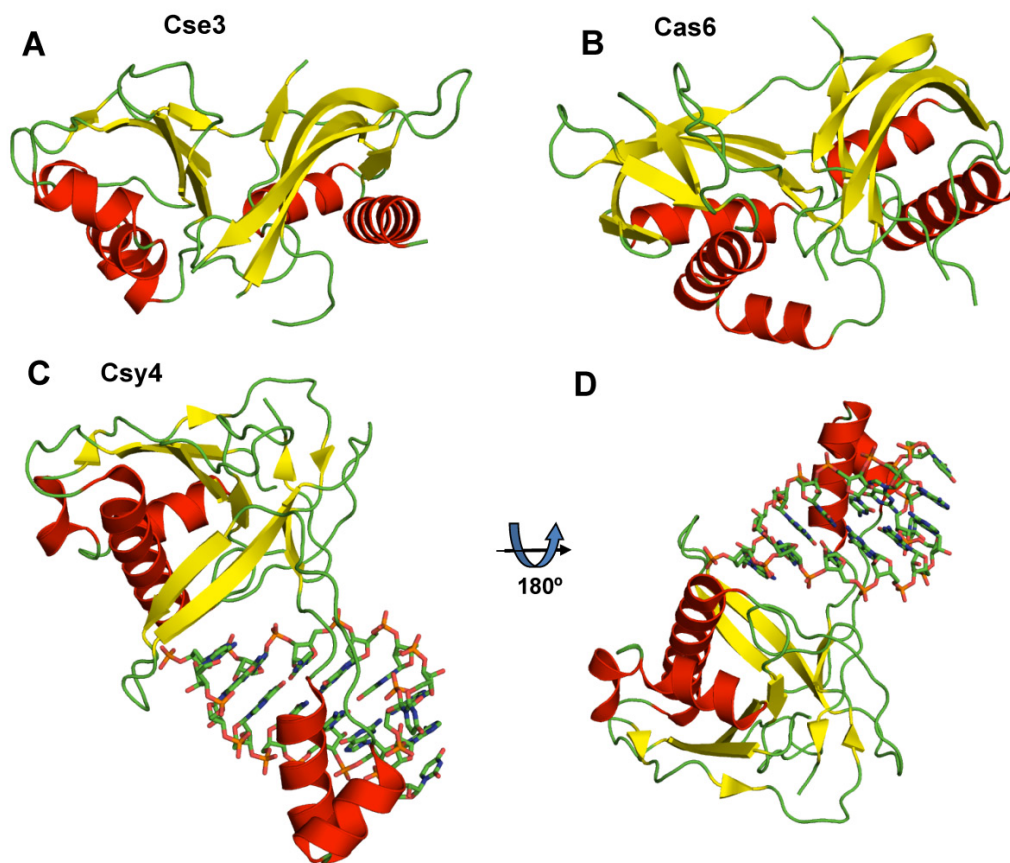


Figure 1-5: Various CRISPR-Processing endoribonucleases shown in equivalent orientations based on SSM structural alignments. **(A)** *Thermus thermophilus* Cse3 displays a duplicated RRM-like fold (66) (PDBID: 3I4H). **(B)** *Pyrococcus furiosus* Cas6 also displays a duplicated RRM-like fold (60) (PDBID: 1WJ9). **(C)** *Pseudomonas aeruginosa* Csy4 in complex with product RNA (67) (PDBID: 2XLK). Unlike Cse3 and Cas6, Csy4 displays only a single RRM-like fold. **(D)** Csy4 rotated to show RNA-interactions.

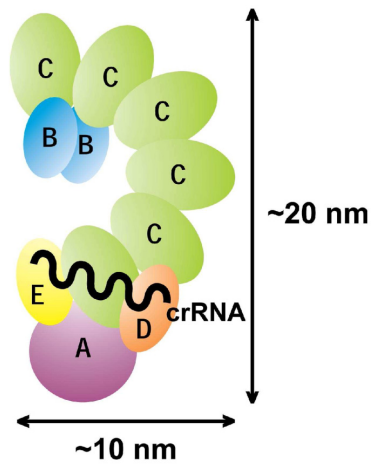


Figure 1-6: Proposed architecture of *E. coli* CASCADE. Six CasC(COG1857) subunits arranged with an open symmetry form the structural core of the complex. [Figure from ref. (69)]

Of the *E. Coli* CASCADE protein components, only the COG1857(Cse4) and COG1688(Cas5e) protein families are found in other Cas subtypes (Cse, Csa, Cst, Csd, Csy and Csh) (58) or CAS Systems (CASS 1,2,3,5,7, and 7a) (41, 59) with recognizable CASCADE components. COG1857 and COG1688 also display conserved gene synteny, and thus, are likely to represent both the structural and the evolutionarily conserved core of CASCADE (Figure 1-7).

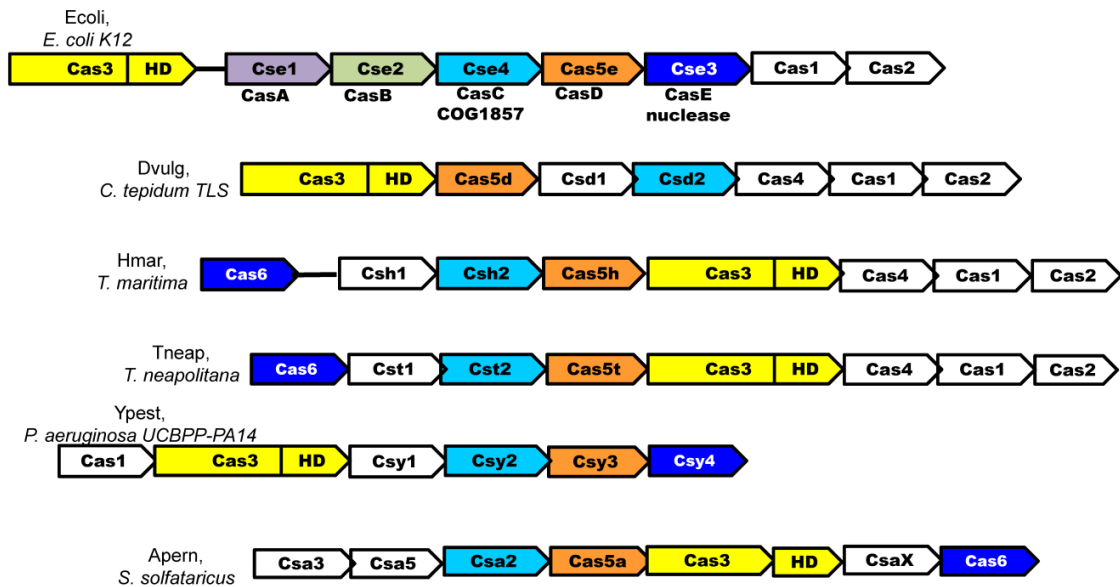


Figure 1-7: CasC-like and CasD-like proteins are the most phylogenetically widespread of the *E. coli* CASCADE components and are found in six of eight subtypes. Example Cas loci from *E. coli*, *D. vulgaris*, *H. marismortis*, *T. neapolitana*, *Y. pestis* and *A. pernix solfataricus* subtypes are shown. Predicted CasC-like proteins are shown in Cyan, CasD-like proteins are orange. The CASCADE components are also found adjacent to Cas3 proteins (yellow) and CRISPR-transcript-processing endonucleases (blue).

The most well-studied example of the second type of system is found in *Streptococcus thermophilus*. This type of CRISPR-Cas system apparently lacks CASCADE or Cas3 but still confers resistance by targeting viral DNA (58, 59, 62). Instead of CASCADE/Cas3, these systems use Csn1 or similar proteins (40), a predicted HNH-type nuclease. Csn1 is called Cas5 by Barrangou *et al.* (40), but is unrelated to Cas5 orthologs as designated by Haft *et al.* (58).

The third type of system is called the RAMP module (Cmr1-6) because it includes a several proteins related to the Repeat Associated Mysterious Protein (RAMP) superfamily (58, 59). Thus far, this system is always found encoded in the same genome

as one of the other two systems. *Pyrococcus furiosus* Cmr1-6 form a complex with psiRNA, a crRNA which has had the 3' handle and part of the 3' end of the spacer cleaved by an unknown nuclease (70). This "CMR" complex specifically degrades RNA that is complementary to the spacer portion of the psiRNA, an activity analogous to the RNA-Induced Silencing Complex (RISC) in eukaryotic RNAi (63). Thus the RAMP module adds an RNAi-like functionality to CRISPR-Cas systems (Figure 1-8).

Regulation of CRISPR-Cas

Transcriptional regulation of the CRISPR-Cas system has been reported for two organisms. In *Thermus thermophilus*, the CRISPR-Cas system is partially regulated by the cAMP-receptor protein (71). In *E. coli* the CRISPR-Cas system is regulated by the H-NS/LeuO system. H-NS binds to and suppresses the *cas* gene promoters. Because H-NS also binds viral DNA, the *cas* genes are likely to be de-repressed when viral DNA is present to compete with the Cas promoter for H-NS binding (72, 73). The results presented in this thesis, provide the first insight into the possible regulation of an Archaeal CRISPR-Cas system (Figure 1-8).

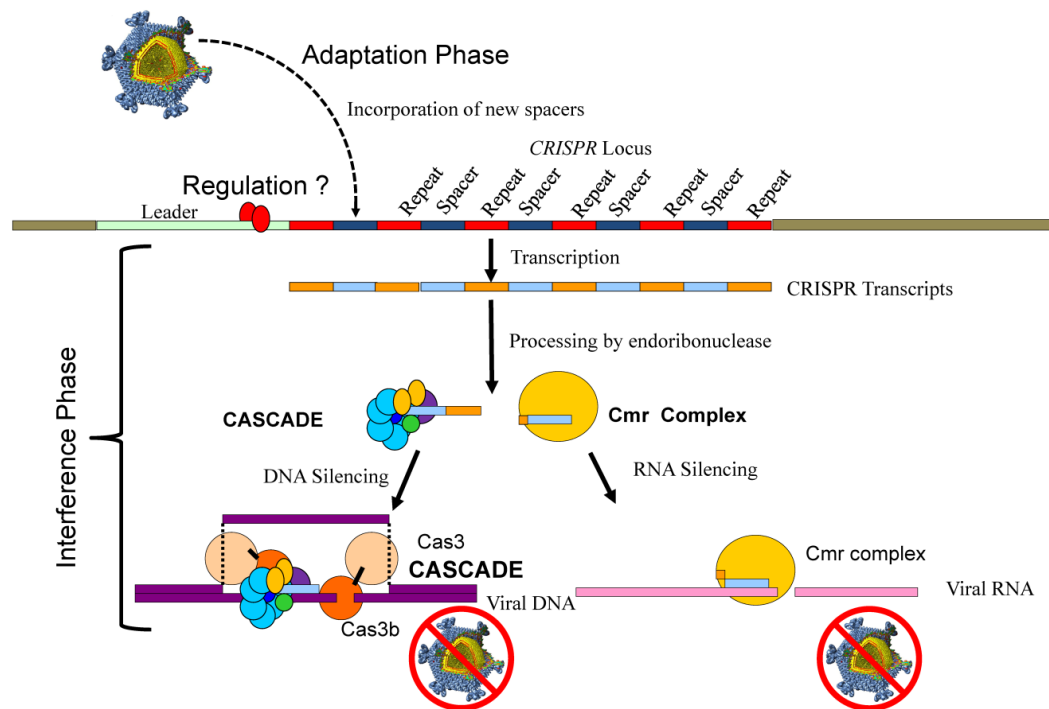


Figure 1-8: Outline of CRISPR-Cas function. In the interference phase, new spacers derived from invading DNA are incorporated into CRISPR loci adjacent to the leader sequence by an unknown process. In the interference phase, the CRISPR is transcribed and the transcript is processed into short RNA molecules which are used to target invading DNA and RNA. CRISPR-Cas also appears to be regulated in several systems.

The *Sulfolobus solfataricus* CRISPR-Cas System

Sulfolobus solfataricus, a model system for the Crenarchaeota, has played a significant role in studies of CRISPR/Cas (7, 9, 11, 16, 59, 65, 74). The *S. solfataricus* genome encodes the *Aeropyrum pernix* or Aperi CRISPR/Cas subtype, designated *csa*, plus the RAMP module of Haft *et al.* (58, 59). *S. solfataricus* strain P2 contains six CRISPR loci designated *CRISPRs A-F*, though E and F may be nonfunctional (11). *CRISPRs A-D* are localized to an 80 kbp region that, except for *csa4*, contains one or more copies of each of the *cas* and *csa* genes (Figure 1-9)(58).

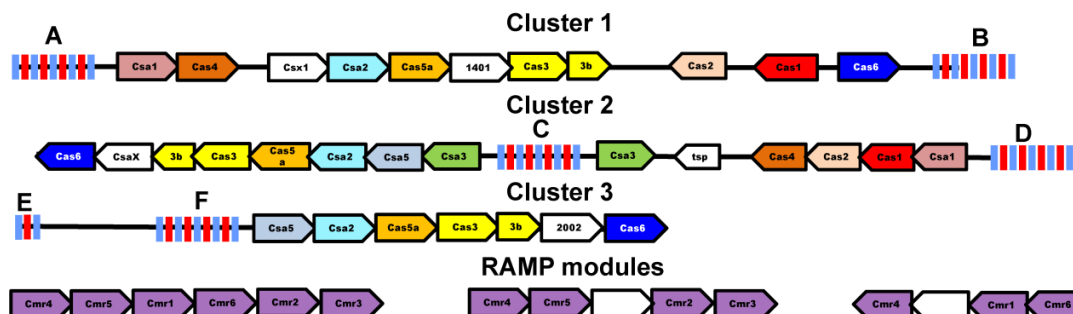


Figure 1-9: The *S. solfataricus* CRISPR-Cas system. The *S. solfataricus* genome encodes six CRISPR loci in three clusters plus three clusters of genes from the RAMP module.

Research Goals

The work presented herein is an expansion and continuation of an ongoing project to understand a model Crenarchaeal virus-host system through a structure-function approach. The project to structurally characterize the STIV proteome was initiated by Dr. Eric Larson and has resulted in the structures of F93 (34), B116 (35) and A197 (36) and the major capsid protein (39). Similar projects focusing on *Sulfolobus*-infecting *Fuselloviridae* were initiated by Drs. Paul Kraft and Smita Menon (29-33). The work described in this thesis initially started as a continuation of the structural characterization of the STIV proteome. It was later expanded to include a key host response system, the CRISPR-Cas system. This work focuses on structural determination of viral and host proteins, the identification and characterization of protein-nucleic acid complexes, and the development of specific functional hypothesis. Specifically, the work focuses on the structural determination of the STIV protein A81, the structural determination of the host CRISPR-associated protein, Csa3, and the isolation and characterization of the first Archaeal CASCADE.

MATERIALS AND METHODS

Introduction

This chapter will outline the materials and methods for all the studies described in this thesis. The first section describes the structural determination of STIV-A81. The chapter then describes the cloning and expression trials for 26 *S. solfataricus* CRISPR-associated proteins. This chapter will outline the procedures used to identify protein-protein interactions and ribo-protein complexes in *S. solfataricus*. The chapter also outlines the materials and methods specific to Csa3, *S. solfataricus* CASCADE and Csa2.

Materials and Methods for STIV-A81

Expression Purification and Crystallization of Native A81

Most of the predicted STIV open reading frames, including *a81* were cloned with C-terminal His₆ tags into pDest14 by Dr. Eric Larson using the Gateway system (36). Native A81 was expressed, purified and subjected to crystallization trials using the standard protocols described below for the screening of *S. solfataricus* CRISPR-associated proteins. The exception was that aliquots of A81 purified through the Ni-NTA affinity step were frozen at -80° for later use. The aliquots were later be thawed at room temperature then incubated at room temperature for an additional 20 minutes then separated on a Superdex-200 column equilibrated in 5 mM Tris-Cl pH 8.0, 25 mM NaCl. Fresh A81 behaved as a 65 kDa. particle on a calibrated Superdex-200 column. The freeze-thaw routine caused A81 to behave as 91 kDa. particle and was found to be

necessary for subsequent crystallization of the protein. A81 was concentrated to 15 mg/ml with 5,000 Da MWCO Amicon Ultra™ spin concentrators (Millipore). Well-diffracting crystals (1.95 Å maximum resolution) of native A81 in the C2 space group were obtained from 2 µL + 2 µL hanging-drop vapor diffusion trials set up at 12 mg/mL over a well-solution containing 18% (v/v) 5-methyl-1,2-pentandiol, 4% tertiary butanol, 0.1 M trisodium citrate pH 5.7 and 0.2 M ammonium acetate.

Generation of Leucine to Methionine Mutants for Phasing

STIV-A81 has no internal methionine residues for selenomethionine-incorporation and numerous heavy-atom soaking procedures failed to yield derivatives suitable for phase determination by isomorphous replacement or anomalous scattering. Thus, three Leucine to Methionine mutants (Leu25, Leu60 and Leu75), were generated by site-directed mutagenesis using the Gateway entry vector as a template. Primers used for are summarized in table 2.1. Mutant clones were sequences verified (Nevada Genomics).

Table 2-1: Mutagenic primers for A81 Leu to Met mutants			
Mutation	Primers	Entry Clone	Expression Clone
<i>L25M</i>	CAAAGAATTTGTGCTAAGTAAT ATG CAAGAAGATATAAAAAATTTG CAAATTTTATATCTTCTTG CAT ATTACTTAGCACAAATTCTTTG	pEntr.A81L25M	pExpA81L25M
<i>L60M</i>	GTAGATACAGAATTATATAAT ATG TTACACGAGTATGCAATATATC GATATATTGCATACTCGTGTA CAT ATTATATAATTCTGTATCTAC	pEntr.A81L25M	pExp.A81L25M
<i>L74M</i>	CTAGCTGGAGCCACATCT ATG TTATTGATTCTGAAAATAAG CTTATTTTCAGAAATCAATAAA CAT AGATGTGGCTCCAGCTAG	pEntr.A81L74M	pExpA81L74M

Crystallization, Data-Collection and Structure Determination

The Leucine to methionine mutants of A81 were expressed and purified using the same protocols as for the native protein. The L60M and L74M variants were found to crystallize under similar conditions as native A81. Thus, selenomethionine-incorporated

L60M and L74M variants were expressed using B834 methionine auxotroph *E. coli* strain in Vogel-Bonner selenomethionine incorporating media (75) and purified using the same protocol as for Native A81. Drops of selenomethionine-incorporated A81L74M were setup at 12 mg/ml using 2 μ L + 2 μ L hanging-drop vapor diffusing over a well solution containing 18% (v/v) 5-methyl-1,2-pentandiol, 4% tertiary butanol, 0.1 M trisodium citrate pH 5.75-5.85 and 0.2 M ammonium acetate. The crystals were directly plunge-frozen in liquid nitrogen.

A three-wavelength anomalous diffraction dataset centered on the Se-K edge was collected at the Stanford Synchrotron Radiation Laboratory (SSRL beamline 9-2). Data were indexed, integrated and scaled in space group C2 using the HKL2000 software package (76). Crystal parameters and data quality are presented in Table 2-2.

Structure Determination and Refinement

SOLVE (77) was used to determine the positions of the 4 selenium atom substructure and to calculate initial phases. RESOLVE (78) was used for density modification and initial model-building. The asymmetric unit was found to contain four chains of STIV-A81 with a 45% solvent content. Iterative model building with Coot (79) and refinement with REFMAC5 (80, 81) against the remote data set ($\lambda = 0.91162 \text{ \AA}$) led to the final model. Temperature/Libration/Screw (TLS) parameters (82) were included in the refinement, with each of the four STIV-A81 chains divided into 4 TLS groups (Chain A: 1:2-11, 2:12-22, 3:23-49, 4:50-81, Chain B: 5:2-22, 6:23-48, 7:49-56, 8:57-81, Chain C: 9:2-22, 10:23-48, 11:49-56, 12:57-81, Chain D: 13:2-26, 14:27-48, 15:49-67, 16:68-81). The final R factors for the model were 17.4% and 20.3% ($R_{\text{work}}/R_{\text{free}}$). Molprobit

(83) was used for model validation, indicating 97.1 % of the residues fall in the most favored regions of the Ramachandran plot and none in disallowed regions. Residue numbers in the model are consistent with the native, non-tagged protein sequence. Residue 1 and the C-terminal His tags of all chains were not modelled due to the lack of interpretable electron density. MALDI-TOF MS indicates that the N-terminal methionine is not present. Additional details on model refinement and model quality are presented in Table 2-3. Three dimensional structural homology searches were carried out using the DALI (84), VAST (85) and SSM (86) servers, and structural figures were generated with PYMOL (87).

Table 2-2. STIV-A81 Data collection				
Data Set	Se-Edge	Se-Peak	Se-Remote	Refinement
Wavelength (Å)	0.97927	0.97913	0.91162	0.91162
Space Group	C2			
Cell Constants (a,b,c; Å)	80.854, 88.895, 55.283			
$\alpha = \gamma = 90.00^\circ$	$\beta = 123.467$			
Resolution Range ^a (Å)	50-1.83 (1.86-1.83)	50-1.83 (1.86-1.83)	50-1.70 (1.74-1.70)	54-1.70 (1.74-1.70)
Unique Reflections ^a	28,285	28,318	35,239	35,400
Average Redundancy ^a	3.9(3.9)	7.8(7.6)	3.9(3.8)	4.6(3.9)
I/σ^a	35.8 (12.1)	38.0 (19.0)	28.7 (6.7)	31.4 (6.7)
Completeness (%)	98.1 (96.9)	98.1 (96.9)	98.1 (94.6)	97.9 (86.2)
$R_{\text{sym}}^{a,b}$ (%)	4.0 (12.7)	4.3 (11.6)	2.9 (23.9)	4.0 (23.4)
^a Numbers in parenthesis refer to the highest resolution shell.				
^b $R_{\text{sym}} = 100 * \sum_h \sum_i I_i(h) - \langle I(h) \rangle / \sum_h I(h)$ where $I_i(h)$ is the i^{th} measurement of reflection h and $\langle I(h) \rangle$ is the average value of the reflection intensity.				

Table 2-3. STIV-A81 Model Refinement	
R_{work}^c (%)	17.4 (24.0)
R_{free}^c (%)	20.3 (22.0)
Real Space CC^d (%)	96.4
Mean B Value (overall; $\text{\AA}^2$)	20.5
Coordinate Error (base on maximum likelihood, $\text{\AA}$)	0.057
RMSD from ideality:	
Bonds ($\text{\AA}$)	0.020
Angles ($^\circ$)	1.606
Ramachandran Plot ^e :	
Most Favored (%)	97.1
Additional Allowed (%)	2.9
PDB Accession Code	N/A
^c $R_{\text{work}} = \Sigma F_o - F_c / \Sigma F_o$ where F_o and F_c are the observed and calculated structure factor amplitudes used in refinement. R_{free} is calculated as R_{cryst} , but using the “test” set of structure factor amplitudes that were withheld from refinement (4.9%).	
^d Correlation coefficient (CC) is agreement between the model and $2Mf_o - DF_c$ density map.	
^e Calculated using Molprobit (83)	

Screening of *S. solfataricus* Cas Proteins for Structural Targets

In our effort to determine the specific functions of the *Sulfolobus* CRISPR-associated proteins, we initiated systematic structural and protein-protein interaction studies of the predicted CRISPR-associated genes. Initially, 26 proteins were cloned, screened for soluble expression in *E. coli*. Well-behaved proteins were subjected to high-throughput sparse matrix crystallization screening. The structures of two proteins, Csa2 and Csa3, have been solved to date. We also initiated pilot *S. solfataricus* protein-protein interaction studies using Csa2 and Csa3.

Selection of Targets for Structural Studies

The *S. solfataricus* P2 genome encodes 6 *CRISPR* loci which are designated *CRISPR*A-F respectively (46) and multiple paralogues of the Cas, Csa and Cmr proteins. Four *CRISPR* loci (A-D) are encoded in close proximity to each other on the genome. Many of the intervening genes can be identified as belonging to CRISPR-associated families (58, 88, 89). We initially undertook small-scale expression and purification trials to select proteins suitable for structural and biochemical study. The selected genes were cloned and subjected to small-scale expression trials in *E. coli*. Constructs yielding soluble, non-aggregated protein were subjected to robotic high-throughput sparse-matrix crystallization screening.

Twenty-five genes encoded within and between clusters 1 and 2 were selected for study. Most of these genes belong to the core or Aperi subtype CRISPR-associated gene families (58). Several additional genes between clusters 1 and 2 were also included. HHMsearch reveals that several of these belong to the (RAMP) (59) superfamily. One protein (Sso0454) is not encoded near the CRISPRs but had previously been shown to bind to the *S. solfataricus* CRISPR loci and was named the SRSR-binding protein (16). The selected proteins are listed and summarized in Table 2-4.

Table 2-4: CRISPR associated proteins selected for small-scale expression trials				
Gene	Gene Identifier	Family (ref. (58))	Family (ref. (59))	Function ^p predicted, ^e experimental
<i>sso0454</i>	13813609	N/A	N/A	Direct repeat-binding protein ^e (16)
<i>sso1391</i>	15898232	Csa1	RecB-like	RecB-family exonuclease ^p (59)
<i>sso1392</i>	13814612	Cas4	RecB-like	RecB-family exonuclease ^p (59)
<i>sso1399</i>	15898239	Csa2	COG1857	CASCADE component ^e (this work)
<i>sso1400</i>	15898240	Cas5a	RAMP	CASCADE component ^e (this work)
<i>sso1402</i>	13814623	Cas3	COG1203	DNA Helicase ^p (58, 59)
<i>sso1405</i>	15898245	Cas1	COG1518	DNase ^e (61)
<i>sso1406</i>	13814627	Cas6	RAMP	CRISPR transcript-processing endoribonuclease ^e (60)
<i>sso1422</i>	13814646	Cas6	RAMP	CRISPR transcript-processing endoribonuclease ^e (60)
<i>sso1425</i>	15898264	TIGR02581	RAMP	
<i>sso1426</i>	13814650	TIGR02581	RAMP	
<i>sso1427</i>	13814651	PF03787	RAMP	
<i>sso1428</i>	13814654	N/A	N/A	
<i>sso1429</i>	13814655	N/A	N/A	
<i>sso1430</i>	13814656	TIGR02581	RAMP	
<i>sso1431</i>	13814657	TIGR01903	RAMP	
<i>sso1432</i>	13814658	PF03787	RAMP	
<i>sso1440</i>	13814667	Cas3	COG1203	DNA Helicase ^p (58, 59)
<i>sso1441</i>	15898279	Cas5a	RAMP	CASCADE component ^e (this work)
<i>sso1442</i>	15898280	Csa2	COG1857	CASCADE component ^e (this work)
<i>sso1443</i>	13814670	Csa5	AF1870	CASCADE interacting ^e (this work)
<i>sso1444</i>	15898282	Csa3	1517/HTH	DNA-binding, ligand binding ^p (this work)
<i>sso1445</i>	15898283	Csa3	1517/HTH	DNA-binding, ligand binding ^p (this work)
<i>sso1449</i>	13814676	Cas4	RecB-like	RecB-family exonuclease ^p (59)
<i>sso1450</i>	15898286	Cas1	COG1518	DNase ^e (61)
<i>sso1451</i>	15898287	Csa1	RecB-like	RecB-family exonuclease ^p (59)

Cloning

All cloning was done using gateway technology (Invitrogen) to streamline the process. Gateway is a ligase-free cloning method the utilized site-specific recombination between compatible sites called Att sites. In Gateway, a gene-of-interest is amplified using PCR. The primers are designed to incorporate an AttB1 site on the 5' end of the PCR product and an AttB2 site on the 3' end. The gene product is then cloned into pDonr201 using the BP-clonease II enzyme (Invitrogen) to generate an entry vector. The

entry vector insert is then sequenced. After the sequence is confirmed, the gene can be transferred to any number of expression vectors using the LR clonease recombination reaction.

The primers were designed to PCR amplify the *S. solfataricus* CRISPR-associated genes while at the same time adding an AttB1 site, a Shine-Dalgarno site, and a noncleavable His₆ tag to the 5' end of the product and an AttB2 site to the 3' end.

Because 67 additional nucleotides needed to be added to the 5' end of the PCR product, a nested-PCR scheme was designed to amplify the PCR products (29). The gene was first amplified with a set of primers (designated the “inside primers”), which add the nucleotides encoding the N-terminal 6x-his tag to the 5' end of the product and most of the AttB2 site to the 3' end of the PCR product. The resulting product was used as a template for a second round of PCR using a second set of primers named the outside primers. The outside primers added an AttB1 site and a Shine-Dalgarno sequence to the 5' end and the remaining portion of the AttB2 site to the 3' end. The gene-specific “inside” primers are listed in Table 2-5. The universal “outside” primers were 5'-GGGGACAAGTTTGTACAAAAAAGCAGGCTTCGAAGGAGATAGAACCATGCATCACCATCATCAC-3' and 5'-GGGGACCACTTTGTACAAGAAAGCTGGGTCCTA-3'. The final PCR products were generally purified using Agarose gel electrophoresis and the Qiagen gel extraction kit.

The PCR products were cloned into pDonr201 (Invitrogen) using BP Clonease II. The protocol followed manufacturer's instructions with the following changes: The enzymatic reaction was incubated from four hours to overnight instead of one hour and

the protease K reaction was allowed to go for 30 minutes instead of ten minutes. The recombinant products were transformed into the 10G Elite cloning strain of *E. coli*. and plated on LB-agar plates containing 50 µg/ml Kanamycin. Clones were screened for the correct insert using colony PCR with the AttB site-specific primers (Invitrogen). Positive clones were grown overnight in 5 ml LB plus 50 µg/mL Kanamycin. Plasmid DNA was purified using the Qiagen miniprep kit according to manufacturer's instructions. The sequence of the insert was verified using dye terminator sequencing at the Nevada Genomics facility. For *E. coli* expression, the genes were transferred from the pDonr201 vector to the pDest14 vector using LR Clonease II (Invitrogen). The protocol followed manufacturer's instructions with the following changes: The enzymatic reaction was incubated from four hours to overnight instead of one hour and the protease K reaction was allowed to go for 30 minutes instead of ten minutes. The recombinant products were transformed into the DH5α cloning strain of *E. coli* and plated on LB-agar plates containing 100 µg/ml Ampicilin. Clones were screened for the correct insert using colony PCR. Positive clones were selected and plasmids were purified using the Qiagen miniprep kit according to manufacturer's instructions. The genes cloned, primers used, and names of the resulting constructs are summarized in table 2-5.

Table 2-5: Cloning of the *S. solfataricus* CRISPR-Associated genes for expression in *E. coli*. The portion of the primer sequences written in upper-case are the gene-specific portion and the lower-case sequence is the “adapter” portion, which overlaps with the universal primers.

Gene	Inside Primers Sequences	Entry Clone	Expression Clone
<i>ssol0454</i>	ccatgcatcaccatcaccatcacaTGAGCGAGGAAGAAAACATTGAAAAAGTTA cactttgtacaagaaagctgggtcCTAAGCAGATGTGGGAGAAGATTAC	pEntr.A.21	pExp14.A.1
<i>ssol1391</i>	ccatgcatcaccatcaccatcacaTGTTCTTTACTCATTAGATATGCTGTTGC cactttgtacaagaaagctgggtcTTAAGGGTGACAAACCTTATAAAACATGCA	pEntr.B.22	pExp14.B.1
<i>ssol1392</i>	ccatgcatcaccatcaccatcacATGACAATAAGTGGGATTACGATAAAGCAC cactttgtacaagaaagctgggtcTCAATCAACATAAGCTAAATTTCCCTTTAT	pEntr.C.01	pExp14.C.1
<i>ssol1399</i>	ccatgcatcaccatcaccatcacATGTGGATATCCTTCTCCGTTAGGTATTTG cactttgtacaagaaagctgggtcTTACATTACCGTAAACGCATCTATTGGCGT	pEntr.D.08	pExp14.D.1
<i>ssol1400</i>	ccatgcatcaccatcaccatcacATGAAAAGGTGTTGTTATTTTCGGTATTAC cactttgtacaagaaagctgggtcTTATTTTGAATACCATTCTCTAATCCCTC	pEntr.E.04	pExp14.E.1
<i>ssol1402</i>	ccatgcatcaccatcaccatcacATGGGATGTGAAGAAAATAAGAGAGGCA cactttgtacaagaaagctgggtcTCATTTCCTCAACCTCAACCCAGT	Not cloned	
<i>ssol1405</i>	ccatgcatcaccatcaccatcacATGGATAAGAAAATAGCGTTTCGTTAAGGAT cactttgtacaagaaagctgggtcTCACTTCGCTAGGTATGGTCTGTATTTC	pEntr.G.31	N/A
<i>ssol1406</i>	ccatgcatcaccatcaccatcacaTGATTTACACACTTACATTTTCGATTAAAAC cactttgtacaagaaagctgggtcTTAACCATTAAAGTGATTTTACTTCTACCTC	pEntr.H.21	pExp14.H.1
<i>ssol1422</i>	ccatgcatcaccatcaccatcacATGCAGAATACAAAATGCGTGATTTTCGG cactttgtacaagaaagctgggtcTTATCGTAACACAGCTTGCCTCCA	pEntr.I.21	pExp14.I.1
<i>ssol1425</i>	ccatgcatcaccatcaccatcacATGTCTTGTATGGATCTTGATGTTATAACG cactttgtacaagaaagctgggtcTTAACCATTAGTACCAATCACCTCTTCCAA	pEntr.J.21	pExp14.J.1
<i>ssol1426</i>	ccatgcatcaccatcaccatcacATGGTTAACTATACATTTATTGATAAGAGA cactttgtacaagaaagctgggtcTTACTTCTTCGGATAGGGGATCTTTGC	pEntr.K.21	pExp14.K.1
<i>ssol1427</i>	ccatgcatcaccatcaccatcacATGCAAAGATCCCTATCCGAAGAAG cactttgtacaagaaagctgggtcCTACCCTCTCATTAATATATCTAGGATT	pEntr.L.01	pExp14.L.1
<i>ssol1428</i>	ccatgcatcaccatcaccatcacATGATAAGAGGTAAGTTGTTGTTACCAGAA cactttgtacaagaaagctgggtcTCAAGGCAATATACCTACTGTCAAGTAG	Not cloned	
<i>ssol1429</i>	ccatgcatcaccatcaccatcacATGGAGATTATTGCGGGTTATCGTACTAC ccatgcatcaccatcaccatcacATGGAGATTATTGCGGGTTATCGTACTAC	pEntr.N.01	pExp14.N.1
<i>ssol1430</i>	ccatgcatcaccatcaccatcacATGATACCGATTAAAGGTGAGTATGAGGAAT cactttgtacaagaaagctgggtcTCATTTATTCTCAGCCTCCCACAAATATCT	pEntr.O.21	pExp14.O.1
<i>ssol1431</i>	ccatgcatcaccatcaccatcacATGATCTTAATAAGTTAAAGGTTAAAGGC cactttgtacaagaaagctgggtcTTAACTCACCTTATTAAGTATTGAGCTTAA	pEntr.P.32	pExp14.P.1
<i>ssol1432</i>	ccatgcatcaccatcaccatcacATGGTCATTCAAGTTTCAGGTTAAATTTTC cactttgtacaagaaagctgggtcTTAACTCAGATATCTCTTAAGTTCAGAAAA	pEntr.Q.04	pExp14.Q.1
<i>ssol1440</i>	ccatgcatcaccatcaccatcacATGAGGTGCGGCGGGGAATAT cactttgtacaagaaagctgggtcTCAATACACACCACCTATTTCACTATTATA	pEntr.R.32*	pExp14.R.1
<i>ssol1441</i>	ccatgcatcaccatcaccatcacaTGATCTACTCTAAGGTTTTTTTAAACTTC cactttgtacaagaaagctgggtcTCAACTAAAGACGACATATTCCTCCGC	pEntr.S.04**	pExp14.S.1
<i>ssol1442</i>	ccatgcatcaccatcaccatcacATGATAAGCGGTTAGTTAGGTTTTTGGTA cactttgtacaagaaagctgggtcCTACTCTTCTCTAATTTAACTACTAAGTC	pEntr.T.04	pExp14.T.1
<i>ssol1443</i>	ccatgcatcaccatcaccatcacATGGCTCAGAAATCGGAGAAGGA cactttgtacaagaaagctgggtcTCATTTCTTCTACACCTTGCTCA	pEntr.U.21	pExp14.U.1
<i>ssol1444</i>	ccatgcatcaccatcaccatcacATGATATTGATTTTGACACTTGGATTGAT cactttgtacaagaaagctgggtcTTAGAACCATAGATTTTCAATAAAATCTC	pEntr.V.01	pExp14.V.4
<i>ssol1445</i>	ccatgcatcaccatcaccatcacATGAAGTCTATTTTGTACCATGGGA cactttgtacaagaaagctgggtcCTAATAAGGTATGTTAACTTCTTTTCCCTT	pEntr.W.01	pExp14.W.1
<i>ssol1449</i>	ccatgcatcaccatcaccatcacATGGTTAGTGTAAACGGATTTAAAGACTAT cactttgtacaagaaagctgggtcTCATAGCAAAACACTAACTGAGCAAAACC	pEntr.X.31	pExp14.X.2
<i>ssol1450</i>	ccatgcatcaccatcaccatcacaTGATAAGCGTGAGGACTTAGTAATATCTG cactttgtacaagaaagctgggtcTCACATCAACCACTGAAACCCCTATAC	pEntr.Y.01	pExp14.Y.1
<i>ssol1451</i>	ccatgcatcaccatcaccatcacaTGAGGTGCGCAGATAGTTAGACAGC cactttgtacaagaaagctgggtcTCACCTCTTATTATTACAATAGTTTAG	pEntr.Z.21	pExp14.Z.1
*Sequencing revealed that the Sso1440 clone is actually Sso1999, a close paralog. **The Sso1441 clone has a deletion in the Shine-dalgarno sequence and should be re-cloned.			

Small-Scale Screening for Soluble Protein Expression

To identify targets suitable for structural studies, the constructs were screened for soluble expression in *E. coli* using standard protocols. Constructs, which did not yield soluble protein under the conditions tested are likely to require extraction from the inclusions bodies and *in vitro* refolding or additional optimization. Potential optimization includes expressing the protein at reduced temperatures, co-expressing GroEL-GroES to chaperone protein folding or expressing the protein in the Origami (Novagen) *E. coli* strain, which has mutations in the thioredoxin reductase and glutathione reductase genes to promote cytoplasmic disulfide bond formation. Other possibilities include expressing the protein in yeast or insect cell lines or in *S. solfataricus* itself (24).

For initial screening, constructs were transformed into the *E. coli* strain BL21-CodonPlus(DE3)-RIL (Stratagene). Single colonies were used to inoculate 50 ml of LB containing 100 µg/ml ampicillin and 34 µg/ml chloramphenicol. Starter cultures were grown overnight at 37°C. 10 ml of each starter culture was used to inoculate 1 L LB containing 100 µg/ml ampicillin and 34 µg/ml chloramphenicol which was grown at 37°C. When the cultures reached an optical density at 600 nm of 0.6-0.8 (usually 4-6 hours), IPTG was added to a final concentration of 0.5 mM to induce protein expression. The cultures were grown an additional 4-6 hours at 37°C then harvested by centrifugation. Cell pellets were stored at -80°C prior to purification. During the course of the screening, we became aware of a newly developed *E. coli* autoinduction system (90). Two proteins, Sso1442 (Csa2) and Sso1445 (Csa3) were also expressed using the autoinduction system leading to a significant improvement in yield. The improvement suggests that in

the future, it would be worthwhile to try expressing the other constructs using the autoinduction system.

In our experience, 1D SDS-PAGE on the *E. coli* cellular proteome fails to definitively show whether or not the target protein is expressed in suitable quantities for crystallographic studies. Therefore, the constructs were screened by small-scale purification. Typically, a cell pellet grown and induced in 1L of LB was resuspended in lysis buffer (20 mM Tris-Cl pH 8.0, 400 mM NaCl) plus 0.1 mM PMSF. Cells were lysed by two passages through a French Press. Because the target proteins are presumably thermal-stable, the lysate was incubated at 65°C for 15 minutes to denature most of the *E. coli* proteins. The lysate was then clarified by centrifugation at 30,000 x g for 20 minutes. The supernatants were applied to 0.5 ml gravity-flow Ni-NTA columns. The columns were washed four times with wash buffer (20 mM Tris-Cl pH 8.0, 400 mM NaCl, 5 mM Imidazole pH 8.0). Bound proteins were eluted using elution buffer (10 mM Tris-Cl pH 8.0, 50 mM NaCl, 200 mM Imidazole pH 8.0). High pH conditions are necessary for N-NTA affinity purification because protonated His side-chains do not bind to the immobilized Ni. Proteins expressing in high enough quantities further purified using a Superdex-75 column in 10 mM Tris-Cl pH 8.0 and 50 mM NaCl to check for aggregation and estimate the oligomeric state. The results of the small-scale expression and purification trials are summarized in table 2-6.

Table 2-6: Results of small-scale expression and purification trials								
Protein	Residues ¹	M.W. (kDa.) ^{1,2}	pI ^{1,2}	Met content ^{2,3}	Ext. coeff. (M ⁻¹ *cm ⁻¹) ^{1,2}	OD ₂₈₀ at 0.1% ^{1,2}	Soluble expression (mg/L) ⁴	Apparent M.W. by SEC
Sso0454	157	18532	10.1	4	11920	0.634	Yes	77 kDa.
Sso1391	298	34445	6.3	7	40715	1.182	2.8	61 kDa.
Sso1392	192	22624	9.3	7	31080	1.374	No	
Sso1399	319	36114	8.3	7	29465	0.816	0.2	
Sso1400	270	30910	7.7	4	54110	1.751	0.1	
Sso1402	579	66489	6.7	8	56450	0.849	Not cloned	
Sso1405	307	35707	9.7	4	58330	1.634	Not cloned	
Sso1406	268	31375	10.0	4	36330	1.158	No	
Sso1422	270	30408	8.9	2	34380	1.131	No	
Sso1425	255	28633	8.1	6	23755	0.830	3.3	20 kDa.
Sso1426	285	32137	9.1	8	24660	0.760	2.7	18 kDa.
Sso1427	282	31893	9.6	8	23630	0.741	No	
Sso1428	559	64240	6.2	7	66295	1.032	No	
Sso1429	301	34433	6.8	7	27850	0.809	No	
Sso1430	213	23154	6.7	3	24785	1.026	6.7	37 kDa.
Sso1431	311	35830	9.2	7	37945	1.059	No	
Sso1432	259	29871	8.8	3	22920	0.767	0.1	
Sso1440	514	57877	8.3	8	41385	0.715	Not cloned	
Sso1441	247	28350	9.1	1	50435	1.775	0.1	
Sso1442	328	36219	6.6	7	16390	0.453	0.5 ⁴	50 kDa.
Sso1443	157	17441	8.7	3	11920	0.683	1.4	19 kDa.
Sso1444	210	23993	7.9	5	12950	0.540	no	
Sso1445	244	27770	9.2	7	10430	0.376	0.5 ⁴	56 kDa.
Sso1449	182	21486	9.1	2	24410	1.136	No	
Sso1450	314	25849	9.0	6	41830	1.176	No	
Sso1451	289	33198	7.7	7	33725	1.016	No	

¹Number of residues, molecular weight and Pi include the N-terminal His₆ tag. ²Molecular weight (M.W.), Pi, Met content, extinction coefficient, and OD₂₈₀ at 1 mg/ml were calculated using the online Protparam tool (<http://expasy.org/tools/protparam.html> (91)). ³The reported number of methionine residues excludes the two N-terminal methionines. ⁴The reported expression levels are for growth in LB and induction with IPTG; The expression levels of Sso1442(Csa2) and Sso1445(Csa3) were later improved using the autoinduction system (90).

Crystallization Trials

For constructs yielding well-behaved soluble protein, the expression and purification were scaled up to generate enough protein for crystallization trials. Pure protein was concentrated to approximately 10 mg/ml and initial crystallization screens were done using 400nl+400nl sitting drop vapor diffusion with commercial screens in 96-well format. The screens used were Crystal Screens 1 and 2, Crystal Screen Lite, Crystal Screen Cryo, PEG/Ion Screen and Natrix Screen from Hampton Research and Cryo Screens I and II from Emerald Biosciences. Trays were setup using the Honeybee crystallization robot (Genomic solutions). Initial crystallizations conditions were optimized in by hand in 24 well 2 μ L+2 μ L hanging-drop vapor diffusion format. Sso0454 (SRSR binding protein), Sso1426, Sso1391 (Csa1), Sso1442 (Csa2), Sso1443 (Csa5) and Sso1445 (Csa3) were subjected to crystallization trials and Sso0454, Sso1426, Csa2, Csa5 and Csa3 crystallized (Figure 2-1). Upon completion of the small-scale screening, we considered the results of the screening, the importance of the proteins to CRISPR-Cas and the desire to avoid direct competition with other research groups and chose to initially focus on Csa2 (Sso1442) and Csa3 (Sso1445) for further study.

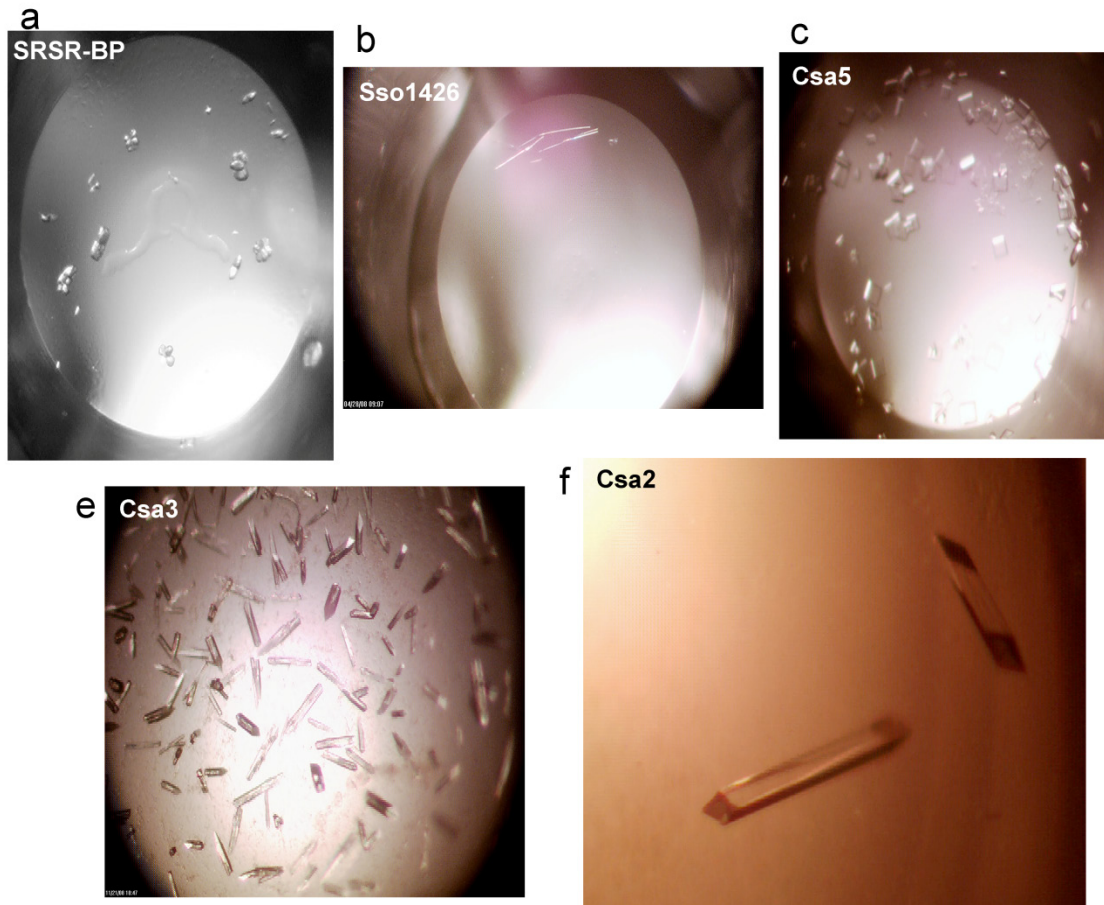


Figure 2-1: Crystals of Cas proteins. Initial conditions identified by robotic sparse-matrix screening for (a) SRSR-binding protein in 0.2 M ammonium dihydrogen phosphate, 0.1 M Tris-Cl pH 8.5 and 50% MPD. (b) Sso1426 (RAMP) in 0.1 M Na/K Phosphate pH 6.2, 40% ethylene glycol. (c) Sso1443 (Csa5) in 0.2 M MgCl_2 , 0.1 M Citrate pH 5.5 and 40% PEG-400. Optimized crystals of (e) Selenomethionine-incorporated Csa3 and (f) native Csa2.

Identification of Protein and Nucleic-Acid Binding Partners

During the course of these studies, we received a new *S. solfataricus* expression vector, pSeSD1 from Dr. Qunxin She. This new vector allowed us to express recombinant tagged protein in its native host and identify co-purifying binding partners. pSeSD1 is based on the *S. icelandicus* pRN2 replicon (18) fused to *E. coli* pUC replicon for cloning

and propagation. The vector also encodes the *S. solfataricus* *pyrE-pyrF* gene cassette for complementation of uracil auxotrophs and a multiple cloning site downstream from an *araS* promoter.

Because of the high growth temperature of *S. solfataricus*, affinity tags based on whole-proteins such as glutathione-S-transferase, chitin-binding protein and maltose-binding protein are unsuitable. Thus, for expression in *S. solfataricus* we chose tandem Strep II and His affinity tags, which are short peptides. Sso1442 (Csa2) and Sso1445 (Csa3) with the tags at both the N- and C- termini were cloned into the vector using the *NdeI* and *NotI* restriction sites. The genes were amplified from the original entry vectors using a nested PCR reaction. The primers were designed to add StrepII and His tags and restriction sites. For constructs with C-terminal tags, the His tag already present in the pSeSD1 vector was utilized. For constructs with N-terminal tags, the His tags already present in the template were utilized. The PCR products and vector were digested with *NdeI* and *NotI* (NEB) and the digested vector was gel-purified. The PCR product was ligated into the vector using T4 ligase (Promega). Clones were screened for the correct insert in the correct orientation using colony PCR. The final vectors were all sequence-confirmed. The constructs made and primers used are summarized in Table 2-7.

Table 2-7: Primers used to generate constructs for expression in <i>S. solfataricus</i>		
Construct	Primers	
pSeSD_sso 1442_ NTAP	Inside	TGGAGCCACCCGAGTTCGAAAAACTCGAGCATCACCATCACCATCACCATCACATG AAATTGCGGCCGCTTACTCTTCCTCTAATTTAACTACTAAGTCC
	Outside	GGAACCCATATGGCGAGCGGTGGAGCCACCCGAGTTC AAATTGCGGCCGCTTACTCTTCCTCTAATTTAACTACTAAGTCC
pSeSD_sso 1442_ CTAP	Inside	GGAACCCATATGAAGTCCTATTTTGTCAACCATGG TCGAACTGCGGGTGGCTCCACTCTCCTCTAATTTAACTACTAAGTCCTC
	Outside	GGAACCCATATGAAGTCCTATTTTGTCAACCATGG ATGATGTGCGGCCGCTTTTCGAACTGCGGGTGGCTCCA
pSeSD_sso 1445_ NTAP	Inside	TGGAGCCACCCGAGTTCGAAAAACTCGAGCATCACCATCACCATCACCATCACATG AAATTGCGGCCGCTTAAATAAGGTATGTTAACTTC
	Outside	AAATTGCGGCCGCTTACTCTTCCTCTAATTTAACTACTAAGTCC AAATTGCGGCCGCTTAAATAAGGTATGTTAACTTC
pSeSD_sso 1445_ CTAP	Inside	GGAACCCATATGAAGTCCTATTTTGTCAACCATGG ATGATGTGCGGCCGCTTTTCGAACTGCGGGTGGCTCCACGCGCTTGCATAAGGTATGTT
	Outside	GGAACCCATATGAAGTCCTATTTTGTCAACCATGG ATGATGTGCGGCCGCTTTTCGAACTGCGGGTGGCTCCA

Transformation and Protein Expression in *S. solfataricus*

The *S. solfataricus* PH1-16 uracil auxotroph strain (13) was grown at 80°C in Brock's defined media (92) plus 10 µg/ml uracil and 0.2% sucrose with long-necked shaker-flasks. For transformation, a 50 ml culture was grown to an OD₆₀₀=0.2-0.5 and chilled on ice for 15 minutes. The cells were pelleted for 10 minutes at 3000 x G and gently resuspended in 20 mM ice cold sterile 20 mM sucrose. The cells were repeatedly pelleted and resuspended in 25 ml, 10 ml, 1 ml and 0.4 ml ice cold 20 mM sucrose. For electroporation, 1 µg of the expression vector was combined with 50 µL electrocompetent cells in a 1 mm cuvette. Electroporation was done using a Biorad GenePulser Xcell (1.5 kV, 400 Ω, 25 µF). Time constants were generally 9-10 ms. Transformed cells were immediately diluted into 1 mL ice cold water and incubated on ice for 10 minutes. The cells were then diluted into 25 ml 80 °C Brock's defined media plus 0.2% sucrose. Growth rates of transfected and mock-transfected *S. solfataricus* were monitored using OD₆₀₀. A growth advantage for transfected cells were generally observed 50-60 hours post-transfection. Example growth curves are shown in (Figure 2-2). The presence of the

expression vectors were confirmed using PCR. For long-term storage, a 25 ml culture of transformant ($OD_{600}=0.2-0.8$) were cooled on ice for 10 minutes then centrifuged at 3000 x g for 10 minutes. The pellets were gently resuspended in 1 ml Brock's media with 0.2 % sucrose and 25% glycerol. The cells were aliquoted, flash-frozen in liquid nitrogen and stored at -80°C .

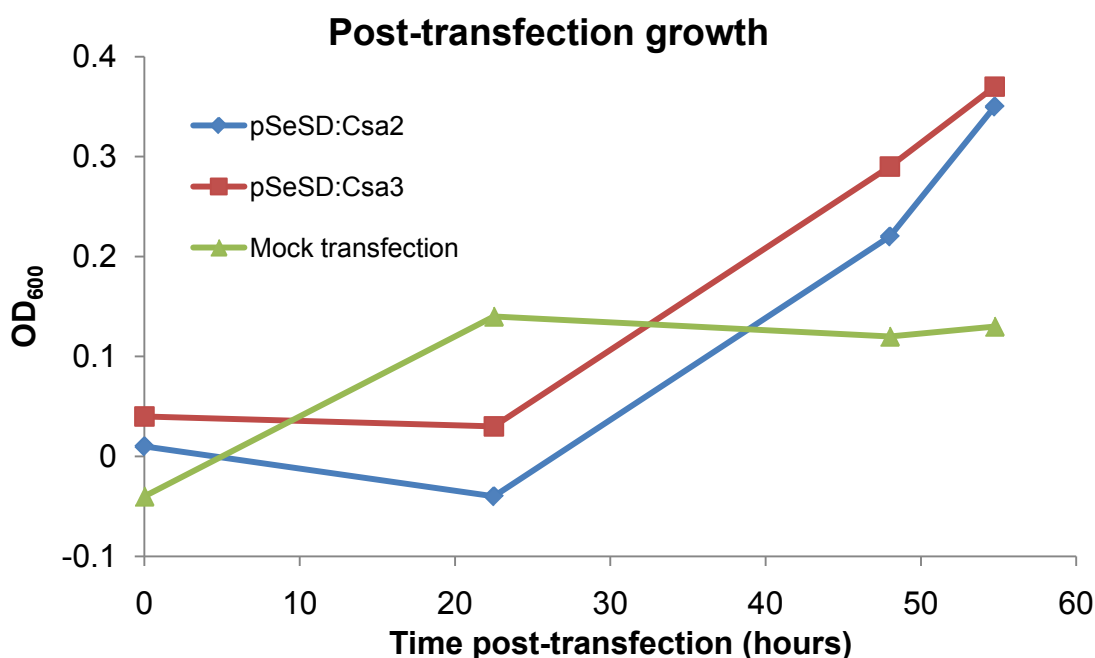


Figure 2-2: Post-transfection growth of *S. solfataricus* PH1-16. Example growth curves of cells transformed with pSeSD1:Csa2 (red), pSeSD1:Csa3(blue) and mock transfected cells (green) grown in uracil-free media are shown. A growth advantage for cells transformed with the pSeSD1 is observed at 50-60 hours post-transfection.

PH1-16 containing the vectors were typically grown at 80°C to an $OD_{600}=0.8-1.2$ in Brock's minimal media supplemented with 0.1% tryptone and 0.2% sucrose. 50 ml of the resulting starter culture was used to inoculate 1 L Brocks minimal media supplemented with 0.1% tryptone and 0.2% arabinose to induce expression. The cultures were initially

grown at 80°C for two days (the final OD₆₀₀ ranged from 0.8 to 2.0) and harvested by centrifugation and stored at -80°C. For expression of generation of the CASCADE-like complex, we found that higher yields were obtained when the cultures were induced for three days.

Co-Purification of Cas Proteins and Putative Binding Partners

The *S. solfataricus* PH1-16 cell pellets were resuspended in lysis buffer [20 mM NaH₂PO₄, pH 7.5, 150 mM NaCl, 0.1% NP-40, 1 mM EDTA, 0.1 mM PMSF and protease inhibitor cocktail set III (Calbiochem)] and homogenized. The homogenate was centrifuged 30 min. at 30,000 xG. The supernatant was combined with 100-500 µL StrepTactin resin (Sigma) and inverted end-over-end for 4 hours at 4°C. Beads were washed four times with 2 bed volumes 20 mM NaH₂PO₄, pH 7.5, 150 mM NaCl, 0.1% NP-40, 1 mM EDTA and twice with two bed volumes 20 mM NaH₂PO₄, pH 7.5, 150 mM NaCl, 0.1% NP-40. Bound proteins were eluted from the StrepTactin resin by washing four times with StrepTactin Elution buffer [20 mM NaH₂PO₄ pH 7.5, 150 mM NaCl, 2.5 mM desthiobiotin]. Elution fractions were pooled. 50-200 µL Ni-NTA-agarose (Qiagen) was added to pooled elution fractions and sample was rotated end-over-end for 30 min at 4°C. The beads were washed four times with two bed volumes 20 mM NaH₂PO₄, pH 7.5, 150 mM NaCl, 0.1% NP-40 and twice with two bed volumes 20 mM NaH₂PO₄, pH 7.5, 150 mM NaCl. The proteins were eluted from the remaining 40 µL of beads by washing four times with one bed volume Ni-NTA elution buffer [10 mM Tris-Cl pH 8.0, 50 mM NaCl, 200 mM Imidazole].

For in-gel trypsin digests, samples eluted from the Ni-NTA resin were concentrated when necessary and separated on a 12% SDS-PAGE gel. The proteins were visualized using a colloidal-Coomassie stain (93), and the bands were excised and washed. Samples were reduced with 10 mM DTT followed by S-alkylation with 55 mM Iodoacetamide. The samples were subjected to overnight in-gel digestion with sequencing-grade trypsin (Promega). The peptides were extracted from the gel slices by vortexing and sonication in ultra-pure water followed by 5% TFA/50% Acetonitrile. Extracted peptides were pooled, concentrated to 10 μ L in a Speed-Vac and brought up to a final volume of 25 μ L with 0.5% TFA/5% Acetonitrile.

LC-MS/MS was performed with an Agilent XCT-Ultra 6330 ion trap mass spectrometer with an Agilent 1100 CapLC and ChipCube (Agilent Technologies) as described previously (94). Samples were trapped and desalted using a Zorbax 300SB-C18 Agilent HPLC-Chip enrichment column (40 nl) in 5% acetonitrile, 0.1% formic acid delivered by an auxiliary CapLC pump at 4 μ L/min. The desalted peptides were reverse-eluted and loaded onto an analytical capillary column (43 mm by 75 μ m, packed with 5 μ m Zorbax 300SB-C18 particles) connected in-line to the nanospray ion source with a flow of 600 nl/min. The peptides were eluted using a 16 minute 5-95% acetonitrile gradient. Data-dependent acquisition of collision induced dissociation tandem mass spectrometry (MS/MS) was utilized. Parent ion scans were run over the 200-2,200 m/z range at a rate of 24,300 $m/z*s^{-1}$. The data were analyzed using Bruker Daltonics Data analysis software. The compound lists were stored in MASCOT generic file (MGF) format and used in MASCOT (Matrix Science, London, UK) queries of the NCBI nr

database of protein sequences limited to the Archaea with MS and MS/MS mass tolerances of 1.2 and 0.6 Da. respectively.

Specific Methods for Csa3

Expression and Purification

Cloning and initial expression trials were performed as described above. Yields of Csa3 were subsequently increased when expression was switched to an autoinducing system (90). For Sso1445 protein expression, BL21(DE3)-pLysS *E.coli* (Stratagene) were transformed with pExp14-6xHis-Sso1445. Typically a single colony was used to inoculate 25 ml of ZYP-0.8G (90) with 100 µg/ml ampicillin and 34 µg/ml chloramphenicol and grown for 6-8 hours. The resulting starter culture was stored at 4 °C for up to two months. Native protein was expressed in 500-750 ml of ZYP-5025 autoinducing media (90) containing 100 µg/ml ampicillin and 34 µg/ml chloramphenicol, inoculated with starter culture at a 1,000-fold dilution, and grown at 37 °C for 16-20 hours. For expression of selenomethionine-incorporated protein, 500 ml of PASM-5025 medium (90) was inoculated with starter culture at a 1,000-fold dilution and grown at 37 °C for 30 hours. Cells were harvested by centrifugation at $5,500 \times g$ (Sorvall Superspeed RC-2) for 10 minutes and the pellets were stored at -80 °C.

Cell pellets were thawed and resuspended at 5 ml/g of cell pellet in lysis buffer (20 mM Tris, 400 mM NaCl, pH 8.0). Phenylmethylsulfonyl fluoride (PMSF, 0.1 mM) was added to the cell suspension and cells were lysed by passage through a French Press (American Instrument Co., Inc., Silver Springs, MD). The lysate was incubated at 65 °C

for 20 minutes to denature *E. coli* proteins, and clarified by centrifugation at $22,000 \times g$ for 30 minutes. The supernatant was then applied to a gravity-flow column containing a 1-3 ml bed volume of Ni-NTA Agarose (Qiagen). The column was washed with 8 column volumes of wash buffer (20 mM Tris, 400 mM NaCl, 10 mM imidazole, pH 8.0) and Sso1445 was eluted in 10 mM Tris (pH 8.0), 50 mM NaCl and 200 mM imidazole. Sso1445 was then applied to a calibrated Superdex 75 (GE Healthcare) column equilibrated with 10 mM Tris (pH 8.0) and 50 mM NaCl. Protein concentrations were determined by Bradford assay (95) using Protein Assay Reagent (Bio-Rad) and BSA as a standard. The purity and molecular weight of Sso1445 were confirmed by SDS-PAGE.

Crystallization and Data Collection

Purified selenomethionine-incorporated Sso1445 was concentrated to 15 mg/ml with 5,000 Da MWCO Amicon UltraTM spin concentrators (Millipore). The protein was crystallized using sitting drop vapor diffusion. Drops were setup at 22 °C using 4 µl of selenomethionine-incorporated Sso1445 and 4 µl of well solution consisting of 37.5-42.5% MPD, 5% PEG-8000, and 0.1 M sodium cacodylate at pH 7.0-7.5. Crystals up to 0.20 x 0.05 x 0.05 mm in size were obtained in 5-7 days time. A three-wavelength anomalous diffraction dataset centered on the Se-K edge was collected at the Stanford Synchrotron Radiation Laboratory (SSRL beamline 9-1). Data were indexed, integrated and scaled in space group $P2_12_12_1$ using the HKL2000 software package (76). Crystal parameters and data quality are presented in Table 2-8.

Structure Determination and Refinement

SOLVE (77) was used to determine the positions of the 16 selenium atom substructure and to calculate initial phases. RESOLVE (78) was used for density modification and initial model-building. The asymmetric unit was found to contain two chains of Sso1445 with a 43% solvent content. Iterative model building with Coot (79) and refinement with REFMAC5 (80, 81) against the remote data set ($\lambda = 0.91737 \text{ \AA}$) led to the final model. Temperature/Libration/Screw (TLS) parameters (82) were included in the refinement, with each of the two Sso1445 chains divided into 9 TLS groups (Chain A: 1:1-24, 2:25-49, 3:50-68, 4:69-93, 5:94-133, 6:134-143, 7:144-163, 8:164-187, 9:188-212, Chain B: 10:1-24, 11:25-51, 12:52-68, 13:69-93, 14:94-133, 15:134-143, 16:144-160, 17:161-186, 18:187-215). The final R factors for the model were 18.1% and 22.0% ($R_{\text{work}}/R_{\text{free}}$). Molprobity (83) was used for model validation, indicating 99.5% of the residues fall in the most favored regions of the Ramachandran plot and none in disallowed regions. The overall Molprobity score (83) was 1.52, placing Csa3 in the 93rd percentile for overall geometric quality among protein crystal structures of similar resolution (1.55-2.05 $\text{\AA}$). Residue numbers in the model are consistent with the native, non-tagged protein sequence. Residues 213-237 of chain A, and residues 192-196 and 216-237 of chain B were not modelled due to lack of interpretable electron density. Coordinates and structure factors have been deposited in the Protein Data Bank under accession code 2WTE. Additional details on model refinement and model quality are presented in Table 2-9. Three dimensional structural homology searches were carried out using the DALI (84),

VAST (85) and SSM (86) servers, and structural figures were generated with PYMOL (87).

Table 2-8. Csa3 Data collection			
Data Set	Se-Edge	Se-Peak	Se-Remote
Wavelength (Å)	0.97925	0.97869	0.91737
Space Group		P2 ₁ 2 ₁ 2 ₁	
Cell Constants (a,b,c; Å)	53.400, 89.810 102.501		
$\alpha=\beta=\gamma=90.00^\circ$			
Resolution Range ^a (Å)	50-1.90 (1.94-1.90)	50-1.90 (1.94-1.90)	50-1.80 (1.84-1.80)
Unique Reflections ^a	39,389	74,809	46,357
Average Redundancy ^a	4.0(4.0)	4.2(4.1)	4.0(4.1)
I/ σ ^a	30.0(3.2)	34.0(4.0)	27.5(1.9)
Completeness (%)	99.1(95.0)	99.6(98.0)	99.5 (100)
R _{sym} ^{a,b} (%)	3.6 (25.8)	3.6 (25.8)	4.4 (45.7)
^a Numbers in parenthesis refer to the highest resolution shell.			
^b R _{sym} =100* $\sum_h \sum_i I_i(h) - \langle I(h) \rangle / \sum_h I(h)$ where I _i (h) is the i th measurement of reflection h and $\langle I(h) \rangle$ is the average value of the reflection intensity.			

Table 2-9. Csa3 Model Refinement	
R_{work}^c (%)	18.0 (23.4)
R_{free}^c (%)	22.0 (32.9)
Real Space CC ^d (%)	96.1
Mean B Value (overall; Å ²)	16.3
Coordinate Error (base on maximum likelihood, Å)	0.074
RMSD from ideality:	
Bonds (Å)	0.015
Angles (°)	1.287
Ramachandran Plot ^e :	
Most Favored (%)	99.5
Additional Allowed (%)	0.5
PDB Accession Code	2WTE
^c $R_{\text{work}} = \Sigma F_o - F_c / \Sigma F_o$ where F_o and F_c are the observed and calculated structure factor amplitudes used in refinement. R_{free} is calculated as R_{cryst} , but using the “test” set of structure factor amplitudes that were withheld from refinement (4.9%).	
^d Correlation coefficient (CC) is agreement between the model and $2Mf_o - DF_c$ density map.	
^e Calculated using Molprobity (83)	

Small Angle X-ray Scattering Data Collection and Processing

SAXS data were collected on purified Csa3 at the SIBYLS beamline of the Advanced Light Source (Lawrence Berkeley National Laboratory) with a Mar165 CCD detector. Scattering data for consecutive 0.5, 0.5, 5 and 0.5 sec exposures were collected with 1.03320 Å wavelength X-ray radiation at room temperature, and processed with the ATSAS 2.1 suite (96). Prior to data collection Csa3 was repurified using a 24 ml Superdex 200 column equilibrated with 10 mM Tris pH 8.0 and 50 mM NaCl. Sample volumes were 15 µl. Data were collected from both the peak and tail fractions from the

Superdex 200 column, and on samples concentrated from each fraction. The peak fraction and concentrated peak samples were at 2.5, 4.7, 6.2, and 7.2 mg/ml respectively, while the tail and concentrated tail samples were 1.8, 3.1, 6.3, and 7.6 mg/ml. Protein concentrations were calculated from UV absorbance at 280 nm. Data were also collected on buffer blanks consisting of column fractions prior to the void volume for the unconcentrated samples. Flow-through from the spin concentrators was used for background subtraction from the concentrated fractions. Samples were radiation sensitive as monitored by an increase in slope at low scattering angle. However, data from the first and second 0.5 second exposures overlapped, indicating that radiation damage was minimal, at least in the first 0.5 seconds.

The data were processed with PRIMUS of the ATSAS 2.1 suite (96) for the Guinier and Porod analyses. In the Guinier plot, the data were linear and no aggregation was detected in the gel filtration fractions or any of the concentrated samples. Data from the most concentrated sample was further processed using the regularization technique in GNOM, in which the pairwise electron distribution function (Patterson function) was obtained from the transform of the scattering function with a D_{max} of 115 Å. The molecular mass in Da. was calculated by dividing the Porod volume by 1.67. *Ab initio* shape predictions using the programs DAMMIN and GASBOR were run on the scattering data, and 10 independent runs were averaged. Although the averaged envelope was similar to the crystallographic dimer, the envelope contour did not follow the sharp contours of the crystal structure (data not shown). While this discrepancy could be due to the presence of the His-tag which can contribute to the scattering (97), it is more likely

due to the irregular shape, as control shape predictions run on small angle scattering data calculated from the crystallographic dimer also failed to match the crystal structure.

Specific Methods for A-CASCADE and Csa2

Expression of Csa2 in *S. solfataricus* and the Isolation of an Archaeal CASCADE

Sso1442 was cloned with an N-terminal tandem-affinity (Strep-II and His₈) tag and expressed and purified from *S. solfataricus* PH1-16 using the general protocol described above with the following modifications: Cells were grown in 3 L batches to yield 8-10 g wet cell pellet. The cells were lysed with two passages through a French press. The resin volumes were increased to 0.4-0.5 ml for both the Strep-tactin and Ni-NTA purification steps. The material was further purified using a calibrated Superose-6 column equilibrated in 20 mM NaH₂PO₄, 150 mM NaCl. In the absence of detergent, the material eluted in a broad peak corresponding to 350-500 kDa. plus a “tail” of higher molecular weight material extending to the void volume. For TEM analysis, the material was instead passed over the Superose-6 column in 10 mM Na-HEPES pH 7.5, 150 mM NaCl.

Cloning and Sequencing of Small RNAs from the Purified Complex

Small RNAs co-purifying with Csa2 were cloned and sequenced using previously described methods (63). RNA was extracted from purified CASCADE complex using Trizol LS (Sigma, following manufacturer’s instructions). Isolated RNA was treated with 1 U calf intestinal alkaline phosphatase (Promega) for one hour at 37° C then extracted

with phenol:chloroform:isoamyl alcohol (25:24:1, pH 5.2, Fisher) (PCI). RNAs were ethanol precipitated with 100 µg/ml RNase-free glycogen (Ambion) as a co-precipitant. RNA's separated on an 18% polyacrylamide TBE-UREA gel and stained using Sybr-gold (Invitrogen). The visible band was excised and passively eluted overnight at 4° C in 0.5 M Ammonium acetate, 0.1% SDS, 0.5 mM EDTA and ethanol precipitated. A 5'-phosphorylated, 3'-capped DNA oligonucleotide adapter (5'pCTCGAGATCTGGATCCGGG-ddC3', IDT) was ligated to the 3' end of the RNA's using T4 RNA ligase I (NEB). The ligated RNA's were PCI extracted, ethanol precipitated, gel purified and again ethanol precipitated. The RNA's were subjected to reverse transcription using Superscript III (Invitrogen) using the following primer: 5'-CCCGGATCCAGATCTCGAG-3'. cDNA was subjected to extraction with basic phenol:chloroform (1:1, pH 8.3) followed by ethanol precipitation, gel purification and a second ethanol precipitation. The cDNA was poly-Adenylated using terminal transferase (NEB) and PCR amplified using Taq polymerase (NEB) and the following primers: 5'-CCCGGATCCAGATCTCGAG-3' and 5'-GCGAATTCTGCAGT₃₀-3'. The PCR product was gel purified (2% Agarose, TAE), cloned into TOPO-pCRII (Invitrogen), and transformed into TOP10 cells. Cells were plated on X-Gal containing media. White and light blue colonies were chosen for plasmid DNA preparation. Sequencing was performed at the Nevada Genomics sequencing facility.

Expression and Purification of Recombinant Csa2

Cloning and initial expression screening *csa2* (*sso1442*) was carried out as described above. Autoinduction media (90) was subsequently found to provide higher

yields. Thus, for Sso1442 protein expression, BL21(DE3)-pLysS *E.coli* cells (Stratagene) were transformed with pExp14-6xHis-Sso1442. Typically a single colony was used to inoculate 25 ml of ZYP-0.8G media (90) with 100 µg/ml ampicillin and 34 µg/ml chloramphenicol and grown for 6-8 hours. The resulting starter culture could be stored at 4°C for up to two months. For expression, 500-750 mL of ZYP-5025 autoinduction media (90) with 100 µg/ml ampicillin and 34 µg/ml chloramphenicol were inoculated with the starter culture at a 1,000-fold dilution and grown at 37° C for 16-20 hours. Cells were harvested by centrifugation at $5,500 \times g$ (Sorvall Superspeed RC-2) for 10 minutes and the pellets were stored at -80 °C.

For Sso1442 purification, cell pellets were thawed and resuspended at 5 ml/gram of cell pellet in lysis buffer (20 mM Tris, 400 mM NaCl, pH 8.0). PMSF (0.1 mM) was added to the cell suspension and cells were lysed by passage through a French Press (American Instrument Co., Inc., Silver Springs, MD). The cell lysate was incubated at 65 °C for 20 minutes to denature contaminating *E. coli* proteins, and clarified by centrifugation at $22,000 \times g$ for 30 minutes. The supernatant was applied to a gravity-flow column containing a 1-3 ml bed volume of Ni-NTA Agarose (Qiagen). The column was washed with 8 column volumes of wash buffer (20 mM Tris, 400 mM NaCl, pH 8.0) and Sso1442 was eluted in 10 mM Tris, pH 8.0, 200 mM NaCl and 200 mM Imidazole. Sso1442 was then applied to a calibrated Superdex S-75 column (GE Healthcare Life Sciences) equilibrated with 10 mM Tris (pH 8.0) and 200 mM NaCl. Protein concentrations were determined using the Bradford assay (95), Protein Assay Reagent

(Bio-Rad), and BSA as a standard. The purity and molecular weight of Sso1442 were confirmed by SDS-PAGE.

Crystallization and Data Collection

Purified Sso1442 was concentrated to 11 mg/ml with 5,000 Da MWCO Amicon UltraTM spin concentrators (Millipore). The protein was crystallized using hanging drop vapour diffusion. Drops were setup at 22 °C using 2 µl of Sso1442 and 2 µl of well solution consisting of 1.9-2.1 M sodium formate, 0.1 M sodium acetate pH 3.8-4.1 and 0.2 M sodium thiocyanate. Crystals up to 0.40 x 0.06 x 0.06 mm in size were obtained in 3 weeks time. Although single crystals were occasionally observed, Sso1442 crystals usually grew in clusters. These clusters were manually separated to yield several single, diffraction-quality crystals each. For derivatization, 0.3 µL of 100 mM KAu(CN)₂ was added directly to the drop (final concentration ~ 15 mM) and incubated 22 hours prior to mounting and freezing. The crystals were transferred to synthetic mother liquor containing 12.5% glycerol as a cryoprotectant for 30 seconds, and then flash-frozen in liquid nitrogen. A three-wavelength anomalous diffraction dataset centred on the Au-L3 edge was collected on a KAu(CN)₂-soaked crystal and a single-wavelength dataset was collected on a native crystal at the Stanford Synchrotron Radiation Laboratory (SSRL beamline 9-2). Data were indexed, integrated and scaled in space group P2₁2₁2₁ using the HKL2000 software package (76). Crystal parameters and data quality are presented in Table 2-10.

Structure Determination and Refinement

SOLVE (77) was used to determine the positions of the gold-substructure and to calculate initial phases. Two gold sites per asymmetric unit were identified using SOLVE (77). RESOLVE (78) was used for density modification and initial model-building. The asymmetric unit was found to contain four Sso1442 chains and a 50% solvent content. Iterative model building with Coot (79) and refinement with REFMAC5 (80, 81) against the native data set led to the final model. The final model was refined against a native dataset, at 2.0 Å resolution yielding an R_{work} of 18.3% and R_{free} of 22.1%. The geometry of the final model was evaluated using Molprobity (83). 98.1% of the residues fall in the most favored regions of the Ramachandran plot and none fall into the disallowed regions. The overall Molprobity score (83) is 1.34, placing Csa2 in the 99th percentile for overall geometry quality among protein crystal structures of similar resolution (1.75-2.25 Å) in the protein databank (Table 2-11). The coordinates and structure factors have been deposited with the protein databank (PDB ID: 3PS0).

TLS parameters (82, 98) were included in the refinement, wherein the four Sso1442 chains were divided into 7 TLS groups each (Chain A: 1:1-29, 2:30-64, 3:65-105, 4:106-192, 5:193-226, 6:227-272, 7:273-320, Chain B: 8:1-28, 9:29-63, 10:64-135, 11:136-193, 12:194-225, 13:226-296, 14:297-321, Chain C: 15:1-44, 16:45-98, 17:99-132, 18:133-198, 19:199-260, 20:261-295, 21:296-319, Chain D: 22:1-42, 23:43-99, 24:100-185, 25:186-228, 26:229-259, 27:260-295, 28:296-320). The final R factors for the model were 19.8 % / 23.2 % ($R_{\text{work}}/R_{\text{free}}$) respectively. Residue numbers in the model are consistent with the native protein sequence. Residues 18-26, 165-172, 233-239, 303 and 321 of chain

A, 17-25, 164-176, and 233-241 of chain B, 15-22, 149-179, 234-242, 302-303 and 320-321 of chain C, 16-26, 149-177, 234-241, 301-303 and 321 of chain D were not modelled due to lack of interpretable electron density. The model also includes two residues from the N-terminal His tag in chain A and three residues for chains B, C, and D. These N-terminal non-native residues were annotated using negative residue numbers. Coordinates and structure factors have been deposited in the Protein Data Bank under accession code 3PS0. Further details on model refinement and quality are presented in Table 2-11.

Three dimensional structural homology searches were carried out using the DALI (84), VAST (85) and SSM (86) servers, and structural figures were generated with PYMOL (87).

Table 2-10: Csa2 Data Collection

Data Set	Au-L3 edge	Au-L3 peak	Au L3-remote	Native
Wavelength (Å)	1.04013	1.03982	0.96109	0.95369
Space Group	$P2_12_12_1$			
Cell Constants (a,b,c; Å)	74.526, 121.235, 159.163			74.65, 120.876, 158.06
$\alpha=\beta=\gamma=90.00^\circ$				
Resolution	50-2.12	50-2.12	50-1.96	90-2.00
Range ^a (Å)	(2.17-2.12)	(2.17-2.12)	(2.01-1.96)	(2.05-2.00)
Unique Reflections ^a	82,893	158,117	104,196	96,002
Average Redundancy ^a	4.6(4.6)	4.8(4.7)	4.6(4.5)	5.4(5.5)
I/σ^a	32.9(3.4)	39.9(4.9)	33.5(1.8)	34.6(2.7)
Completeness (%)	99.9(99.6)	99.9(99.8)	99.9(99.7)	98.7(100)
$R_{\text{sym}}^{a,b}$ (%)	3.6(34.3)	3.3(22.7)	4.2(63.3)	4.1(45.7)

^aNumbers in parenthesis refer to the highest resolution shell.

^b $R_{\text{sym}}=100*\sum_h\sum_i|I_i(h)-\langle I(h)\rangle|/\sum_h I(h)$ where $I_i(h)$ is the i^{th} measurement of reflection h and $\langle I(h)\rangle$ is the average value of the reflection intensity.

Table 2-11: Csa2 Model Refinement		
R_{work}^c (%)		18.3 (20.0)
R_{free}^c (%)		22.1 (24.6)
Real Space CC ^d (%)		0.960
Mean B Value (overall; Å ²)		22.155
Coordinate Error (based on maximum likelihood, Å)		0.096
RMSD from ideality:		
Bonds (Å)		0.006
Angles (°)		0.930
Ramachandran Plot ^e :		
Most Favored (%)		98.08
Additional Allowed (%)		1.92
PDB Accession Code		3PS0
^c $R_{\text{work}} = \Sigma F_o - F_c / \Sigma F_o$ where F_o and F_c are the observed and calculated structure factor amplitudes used in refinement. R_{free} is calculated as R_{work} , but using the “test” set of structure factor amplitudes that were withheld from refinement (4.9%). Numbers in parentheses refer to the fit to the reflections in the highest resolution bin (2.051 Å – 2.000 Å).		
^d Correlation coefficient (CC) is agreement between the model and $2Mf_o - DF_c$ density map.		
^e Calculated using Molprobit (99).		

Conclusions and Future Directions

This chapter outlines the screening of *S. solfataricus* CRISPR-associated proteins for suitability for structural studies as well as the specific materials and methods for STIV-A81, Csa3 and CASCADE. The screening of constructs for soluble expression in *E. coli* and binding partner identification in *S. solfataricus* suggest several future avenues for research. Csa1 (Sso1391) expressed well but has not crystallized. Csa1 is predicted to be similar to Cas4 and belong to the RecB-like nuclease family of CRISPR-associated

proteins (59). Given that this family is not required for viral resistance given that the appropriate spacer is already present (41) these protein are likely to be involved in the adaptation stage of CRISPR/Cas. Given the lack of understanding of the CRISPR adaptation mechanism (7, 44, 61), a functional and structural understanding of Csa1 is likely to be of great interest. Csa5 has yielded non-reproducible crystals and, as a potential CASCADE-component, is a natural target for further study.

The initial success in binding-partner identification for Csa2 is encouraging and suggests that this approach could be used for other *S. solfataricus* cellular and viral proteins. Potential targets for binding partner identification include the putative adaptation-phase proteins, Cas1, Cas2, Cas4 and Csa1. For viral work, Dr. Narahari Akkaladevi has developed uracil auxotroph of *S. shibate* B16 harboring an inducible SSV1. It may also be worthwhile to investigate the use of cross-linking agents to identify weaker binding partners.

THE STRUCTURE OF STIV-A81

Introduction

A81 is a predicted 81 residue protein encoded in reading frame A of the STIV1 genome. As for most STIV proteins, sequence-based bioinformatic tools have been unsuccessful in predicting the function of A81. There are no sequence homologues to A81 in the NCBI nr sequence database and A81 is not found in STIV2 (10). A81 is the second protein encoded in an apparent polycistron that also includes C121, B124, B264, A109 and B164. Of these, only B164 has a predicted function; it is similar to a Poxvirus ATPase (12). Co-occurrence in a putative polycistron often implicates proteins in a common process, however B164 was detected as a component of the virus particle while the other proteins were not (12). It is possible that all members of the putative operon function in a process such as packaging the genome into the formed capsid and only one of the proteins remains associated with the capsid. However, the real function of these proteins cannot be determined without additional experimental evidence. We hypothesize that the structures of proteins can suggest their function and form the basis for the development of further hypotheses. Our lab has thus, undertaken structural studies of the proteins encoded by the genomes of STIV and other crenarchaeal viruses (34-36). A81 was targeted as part of the on-going effort.

Results

Recombinant C-terminally His₆-tagged A81 was expressed in *E. coli* BL21(DE3)RIL cells and purified using heat denaturation of *E. coli* proteins, Ni-NTA affinity chromatography and Size exclusion chromatography. After the Ni-NTA affinity step, A81 was stored at -80° C and thawed at room temperature. Fresh A81 displayed migration on a calibrated S-200 column consistent with a hexamer (65 kDa) but freeze-thawed A81 migrated as a nonamer (91 kDa). This freeze-thaw cycle was found to be necessary for the subsequent concentration and crystallization of A81.

Fresh A81 was crystallized in the C2 space group and a native dataset was collected to a resolution of 2 Å on our home X-ray source consisting of a Rigaku rotating Cu anode X-ray generator and a Mar245 image plate detector. Self rotation functions (100) (Figure 3-1) revealed non-crystallographic 2- and 4-fold symmetry axes suggesting that four chains were present in the asymmetric unit with 4₂₂ symmetry. This, along with the size-exclusion chromatography data, suggests that A81 is present as an octamer.

Native A81 includes no internal methionine residues for selenomethionine incorporation and numerous heavy atom-soaking procedures failed to produce derivatives suitable for phase determination by isomorphous replacement or anomalous scattering. Therefore three Leucine to methionine variants were produced and assayed for soluble expression, native-like behavior on a Superdex-200 column and crystallizability. Selenomethionine-incorporated A81 L74M was crystallized in the C2 space group and the structure was solved using multi-wavelength anomalous diffraction data collected at

Stanford Synchrotron radiation labs beam line 9-2. Iterative rounds of model-building and refinement yielded a model with $R_{\text{work}}=17.4\%$ and $R_{\text{free}}=20.3\%$.

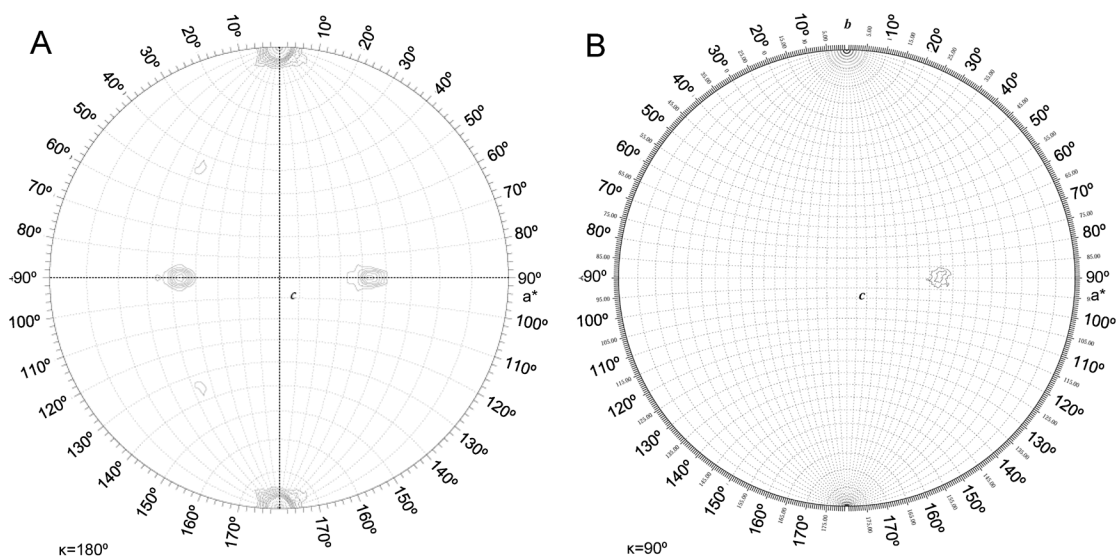


Figure 3-1: Kappa 180° and kappa 90° self-rotation functions. Output of the self-rotation function with (A) $\kappa=180^\circ$ to reveal 2-fold axes and (B) $\kappa=90^\circ$ to reveal a 4-fold rotation axis. The 4-fold axis is perpendicular to both the crystallographic and noncrystallographic two-fold axes, which are themselves, perpendicular to each other. In light of the SEC results, this suggests the presence of an A81 octamer with 422 or D4 symmetry, both in the crystal, and in solution.

The structure of A81 reveals an unusual octameric ring-structure with 422 symmetry. The ring has a central pore lined with α -helices that is 17 Å in diameter. Although most of the A81 surface is negatively charged and the overall pI is 4.5, the interior surface of the pore is positively charged. Residues 1-21 of A81 form three β -strands connected by short loops in an amphipathic 1-2-3 antiparallel β -sheet. Residues 25-81 then form two α -helices with α -helix 1 lining the central pore. In the octameric assembly, there are four-sets of two chains related by 2-fold axes. These four pairs of

chains each share extensive hydrophobic interfaces with $1435 \text{ \AA}^2$ of buried surface area per chain, or 26% of the total surface area and thus could represent 2-chain subcomplexes. In these apparent 2-chain subcomplexes, the three-stranded antiparallel β -sheets from each chain collectively form a six-stranded antiparallel β -sheet with β_3 - β_2 - β_1 - β_1' - β_2' - β_3' topology. This antiparallel β -sheet is amphipathic with one face covered by the helices and the other exposed to the solvent. Additionally, the two α -helices from each chain collectively form an antiparallel four-helix bundle. If the octameric ring is assembled and disassembles these apparent 2-chain subcomplexes could represent a smaller subunit. The octameric assembly is formed from four of the apparent 2-chain subcomplexes. Formation of the octamer buries an additional $980 \text{ \AA}^2$ of surface area per protomer in a hydrophobic interface (Figure 3-2).

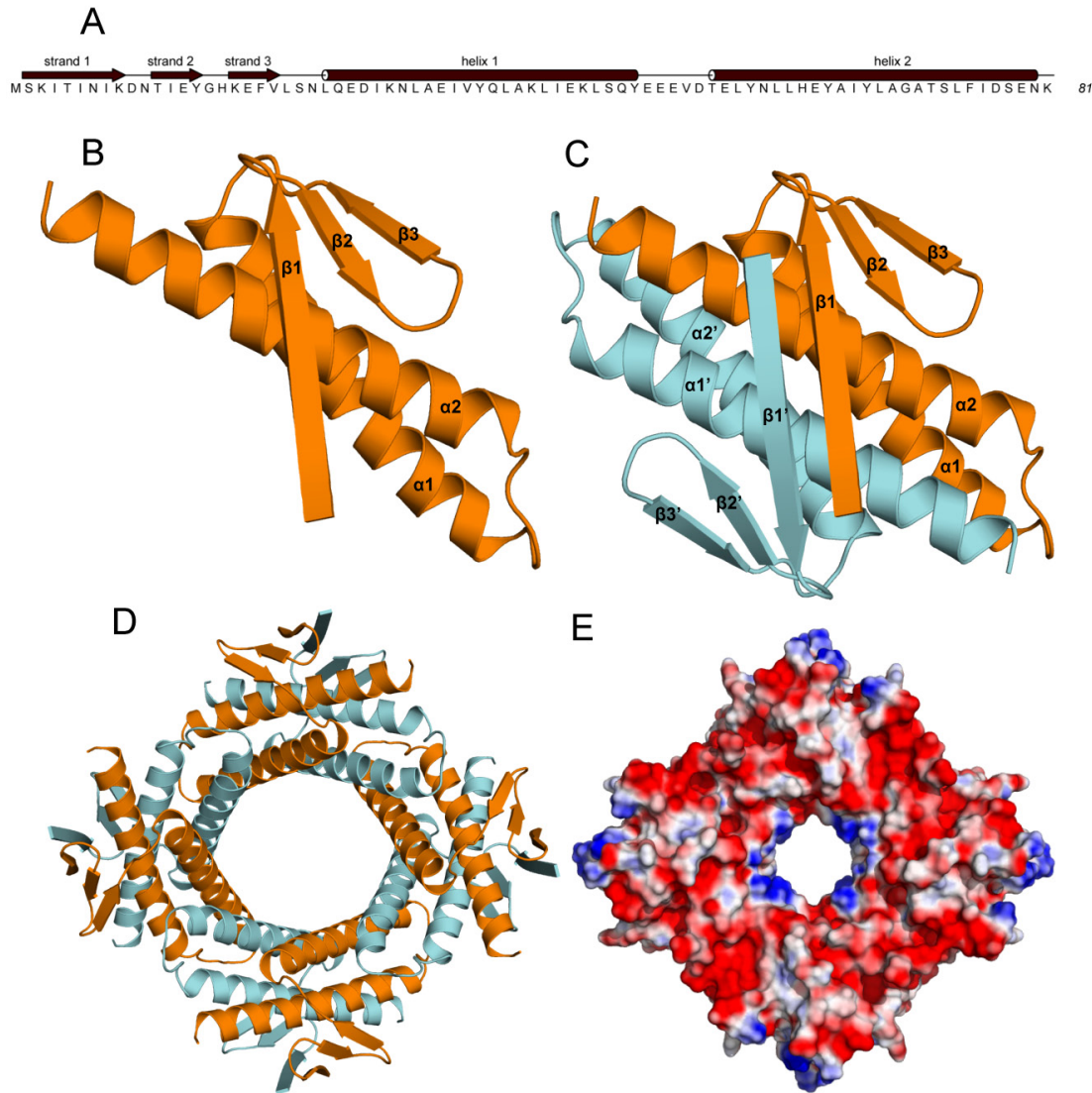


Figure 3-2: The structure of A81. (A) The secondary structure of A81. (B) A single chain of A81 consists of three β -strands followed by two α -helices. (C) The putative two-protomer subcomplex includes a four-helix bundle and a six-stranded β -sheet spanning the inter-protomer interface. (D) The 422-A81 octamer is comprised of four 2-chain subcomplexes and has a 17 Å pore. The octamer is rotated 90° about the horizontal axis as compared to (a) and (b). (E) The central pore is positively charged. Surface rendering of the A81 octamer showing electrostatics. The color ramp is from -15 kt/e (red) to 15 kt/e (blue).

To find structural homologs to A81, the protein databank was queried for similar protein using Dali, VAST, and SSM. A Dali search reveals that the closest structural homolog to a single chain of A81 is *Pyrococcus furiosus* Rubrerythrin, a Fe-bound protein thought to be involved in detoxification of reactive oxygen species (PDB ID's 2HR5 and 1NNQ, Z=6.7, rmsd=2.3 Å over 59 residues (101)). However, A81 lacks the Fe-binding residues found in Rubrerythrin and Rubrerythrin does not form the ring-structure observed in A81. Thus, A81 clearly does not have a Rubrerythrin-like function. Other top Dali matches also did not match the ring-shape of the A81 octamer. One problem with finding proteins similar to A81 using common structural database search tools is that they use a single chain of the query structure to search the protein databank. To overcome this issue, a query representing the A81 octameric assembly was generated by assigning a single chain ID to all eight chains and renumbering the residues to make them unique. VAST and SSM searches fail to yield significant structural matches for A81. While Dali did yield statistically significant matches, they were principally to proteins containing the very common antiparallel four-helix bundle and still did not have the overall ring shape. The preliminary structure of A100 from SSV1 solved by Brian Eilers appears to have a similar fold to STIV1-A81 (SSM structural alignment of a single chain: rmsd=3.0Å over 54 residues). Like A81, A100 oligomerizes in to a large ring but it is an even larger dodecameric ring. The preliminary structure of A100 displays a pore that is 45 Å in diameter and is thus large enough to accommodate dsDNA. (Eilers, *et al.* unpublished results).

Because the central pore appears to be a major feature of the A81 structure, we considered the analogy to other multi-subunit ring-shaped proteins which may have a different overall fold. At 17 Å in diameter, the central pore of A81 is similar in diameter to the 18 Å pore in trimeric RuvBL1. RuvBL1 functions as a helicase and encircles single stranded DNA and/or RNA (102), suggesting that A81 may also encircle single-stranded nucleic acid. A81, however, is not likely to function as a helicase alone because it lacks ATPase domains. Yeast PCNA also forms a ring-like structure, though it has a larger pore diameter (34 Å) that accommodates double-stranded DNA (103). There is also a protein of unknown function from *Bacteroides thetaiotaomicron*, which appears to be a ring-shaped pentamer of four-helix bundles. Similar to RuvBL1 and A81, the pore is about 17 Å in diameter (PDB ID: 2RLD, Joint Center for Structural Genomics, unpublished) (Figure 3-3).

Similar to A81 (pI=4.5), RuvBL1 and PCNA with pIs of 6.0 and 4.4 respectively, have significant areas of negative charge on their surfaces with localized regions of positive charge (Figure 3-4). The high negative charge weakens the interactions of these proteins with single- and double-stranded nucleic acids and allows the proteins to have processivity or the ability to slide along a nucleic acid (102, 103). Thus, if A81 does encircle nucleic acids it is likely to function as part of a processive process.

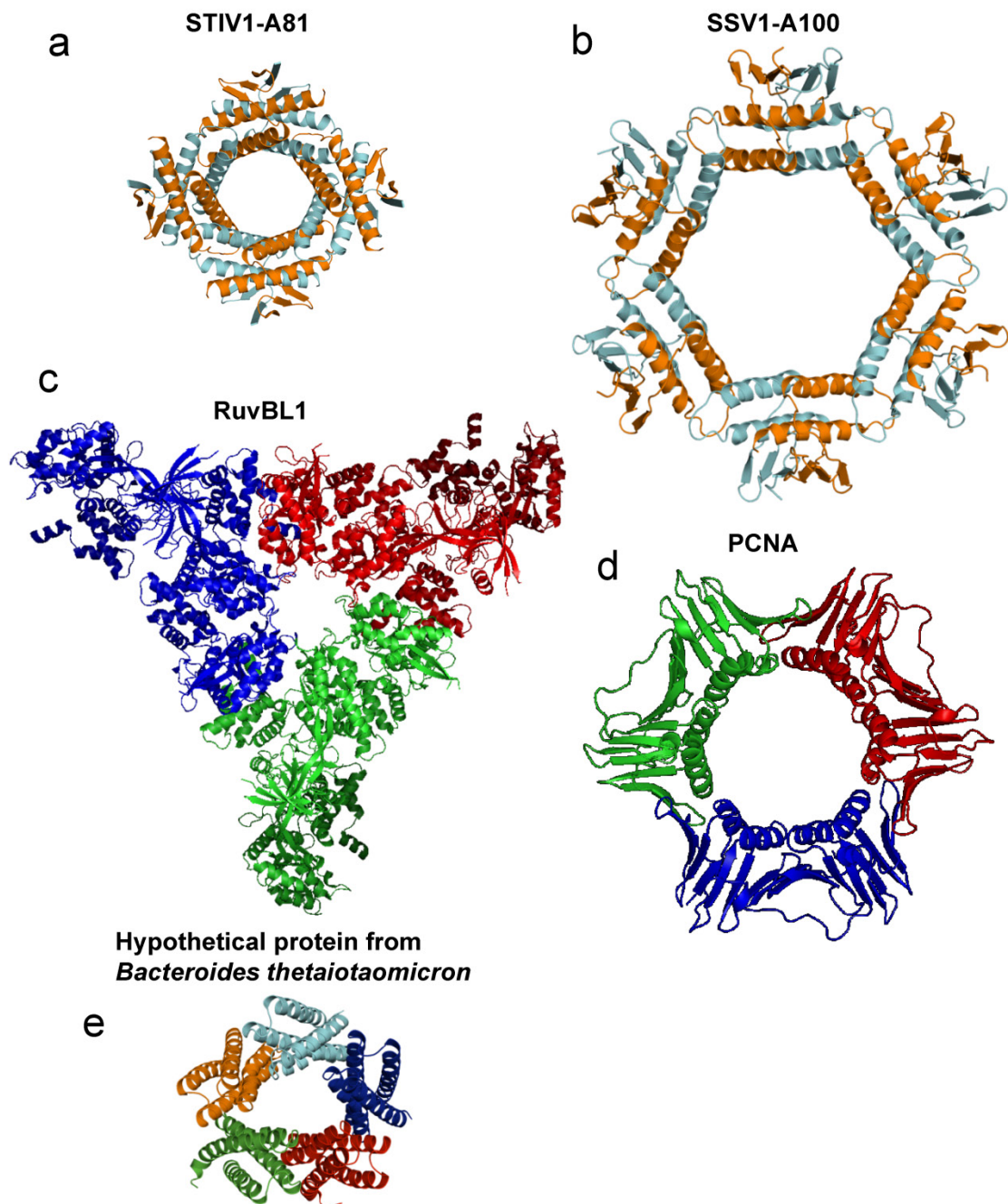


Figure 3-3: Five oligomeric ring structures. (a) STIV1-A81 (b) The preliminary structure of dodecameric A100 from SSV1 (c) Trimeric human RuvBL1 (PDB ID: 2C9O), which encircles single-stranded nucleic acid. (d) Trimeric yeast PCNA (PDB ID: 1PLQ), which encircles double-stranded DNA. (e) Pentameric protein of unknown function from *Bacteroides thetaiotaomicron* (2RLD, Joint Center for Structural Genomics, unpublished).

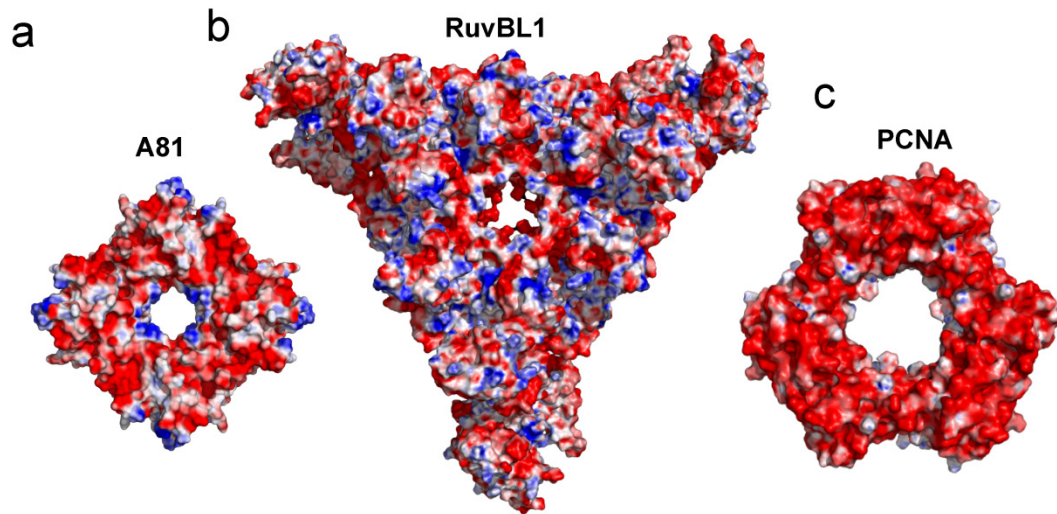


Figure 3-4: A81 has similar surface electrostatics to sliding-clamp proteins. Electrostatic surface maps contoured from -15 kt/e (red) to 15 kt/e (blue) of (a) A81, (b) RuvBL1, and (c) PCNA.

Discussion, Conclusions and Future Directions

A81, with a lack of structural homologs, has proven difficult to definitively assign a function to. Furthermore, A81 has no sequence homologues in the public databases, not even other proteins-of-unknown function. Thus, functionally important sites on the A81 structure cannot be predicted by residue conservation.

What the structure of A81 provides is a basis to develop testable hypothesis about its function. One such hypothesis is that A81 encircles a single-stranded nucleic acid and has a “sliding-clamp functionality”. The central pore is of similar size and charge to the pore in RuvBL1 which encircles single-stranded nucleic acids (102). Furthermore, the

pore is lined with eight solvent-exposed tyrosine side-chains, which are commonly involved in single-stranded nucleic acid binding through aromatic base-stacking interactions (104). This hypothesis could be tested by assaying the binding of A81 to single-stranded and single-strand overhang nucleic acid substrates. A81 lacks any apparent ATPase domain. Thus, if A81 does perform a role similar to RuvBL1, it will need to associate with other proteins which will contribute the ATPase domain. One possibility for the binding partner is B164, which is encoded in the same polycistron as A81 and is predicted to contain an ATPase domain. However, B164 was found as a component of the viral capsid while A81 was not (12). In any case, A81 should thus be expressed in *S. solfataricus* and screened for binding partners in a procedure similar to that described in Chapter 2 for Csa2 and Csa3. Ideally, this would be done in STIV infected cells but a stable uracil auxotroph of P2-2-12 must be isolated first.

Another possibility, which must be considered, is that the crystallized structure of A81 is not the native structure. The behavior of freshly purified A81 during size-exclusion chromatography and the preliminary structure of SSV1-A100 both indicate that this is a possibility. Thus, it is possible that A81 forms a larger ring that can bind double-stranded nucleic acid. To determine the size of A81, lysate from STIV1-infected *S. solfataricus* should be separated by size-exclusion chromatography and the fractions probed for A81 by Western blotting. The migration of native A81 will provide a rough estimate of the true size of the complex it is a component of. While the structure of A81 did not definitively point to a function it revealed a striking and previously unknown quaternary structure and provided a strong starting point for the development of functional hypothesis.

INSIGHT INTO CRISPR-CAS REGULATION FROM THE STRUCTURE OF CSA3

Introduction

Csa3 proteins are predicted to have an N-terminal domain of unknown function fused to a C-terminal winged helix-turn-helix domain, and thus, are predicted to be involved in regulation of CRISPR-Cas (58, 59). Csa3 proteins are members of COG0640 (58) and are predicted to share their domain of unknown function with COG1517-like proteins (59), a large super-family identified by Makarova *et al.*, which includes both CRISPR-associated and non-CRISPR-associated proteins (59). Here, using X-ray crystallography and solution-state small angle X-ray scattering (SAXS), we report the structure of Sso1445, the first structure of a Csa3 protein. The results presented in this chapter have been published in the Journal of Molecular Biology (105).

Results

S. solfataricus Csa3 (Sso1445) was expressed in *E. coli* with a minimal, N-terminal His₆-tag, giving a total calculated mass of 27,769 Da. The purified protein migrates as a 60 kDa. particle on a Superdex 75 column, suggesting a dimer in solution. Csa3 was crystallized in space group P2₁2₁2₁ with two chains per asymmetric unit, and the structure was solved by multi-wavelength anomalous dispersion (MAD) at the selenium edge. Detailed statistics on data collection and structure refinement are presented in Chapter 2. The coordinates and structure factors have been deposited with the protein databank (PDB ID: 2WTE).

Structure of the Csa3 Protomer

The structure of Csa3 reveals dimeric a two-domain protein. The N-terminal domain, comprised of residues 1-132, consists of a six-stranded, doubly wound, mixed β -sheet with flanking α -helices (Figure 4-1A). β -strands N1 through N5 run parallel to each other with β -strand N6 running antiparallel to the other 5. The β -sheet displays 3-2-1-4-5-6 topology with strands β N1 through β N5 connected to each other by right handed helical crossovers. Strand β N5 is then connected to strand β N6 by an 8-residue reverse turn.

The overall fold of the N-terminal domain is thus similar to that of the classic dinucleotide-binding domain composed of six parallel β -stands connected by right handed helical crossovers (100), but with several critical modifications. These include the use of the reverse turn connecting strands β N5 and β N6, as opposed to the α -helical crossover, resulting in an antiparallel β N6, rather than the more common parallel sixth strand. In addition, Csa3 also lacks the Gly-X-Gly-X-X-(Gly/Ala) signature motif found in the α 1- β 1 loop of the dinucleotide-binding domain. Instead, Sso1445 and other Csa3 sequences contain a conserved Gly-(Phe/Ile) sequence in their α N1- β N1 loops (Figures 5-6 and 5-7).

The polypeptide then exits the N-terminal domain, and connects to the C-terminal domain via a 12 residue linker. The linker is present as two turns of α -helix, followed by 4 additional residues in an extended conformation. Following the linker, we find a C-terminal domain comprised of residues 145-212 that form a MarR-like, winged helix-turn-helix (wHTH) DNA binding domain (Figure 4-1A). Finally, electron density is lacking for the remaining 23 to 25 residues of chains B and A, respectively, and these residues are apparently disordered.

This MarR-like wHTH fold, which is commonly found in bacteria and archaea, is part of the larger superfamily of helix-turn-helix proteins that are frequently involved in DNA recognition (106). In the C-terminal domain of Csa3, helices $\alpha C1$, $\alpha C2$ and $\alpha C3$ form a right-handed three-helix bundle where helices $\alpha C2$ and $\alpha C3$ constitute the “helix-turn-helix” motif and helix $\alpha C3$ (residues 173-786) corresponds to the DNA-recognition helix. This tri-helical bundle is followed by a β -hairpin or “wing” comprised of strands $\beta C1$ and $\beta C2$ and a fourth α -helix ($\alpha C4$) that helps to differentiate Mar-R like proteins from other members of the wHTH family (106, 107). In chain B, residues 192-196 at tip of the wing are disordered. Consistent with nucleic acid recognition, the overall pI of Csa3 is 9.2, and the pI of the winged HTH domain alone is 9.6, with a basic patch on the putative DNA-binding surface (Figure 4-2).

Small-Angle X-ray Scattering Identifies the Csa3 Solution-State Dimer

The behavior of Csa3 on the size exclusion column suggested that Csa3 is present as a dimer in solution. To further investigate the structure of this putative solution-state dimer, our collaborators, Dr. Kenneth Frankel, Dr. Susan Tsutakawa and Dr. John Tainer undertook small-angle X-ray scattering (SAXS) studies using the SIBYLS beamline at the Advanced Light Source (108, 109). The resultant small angle scattering was consistent with a globular protein (Figure 4-4A) and the Guinier plots were linear, indicating that the samples were monodisperse with a radius of gyration of 28 Å (Figure 4-3).

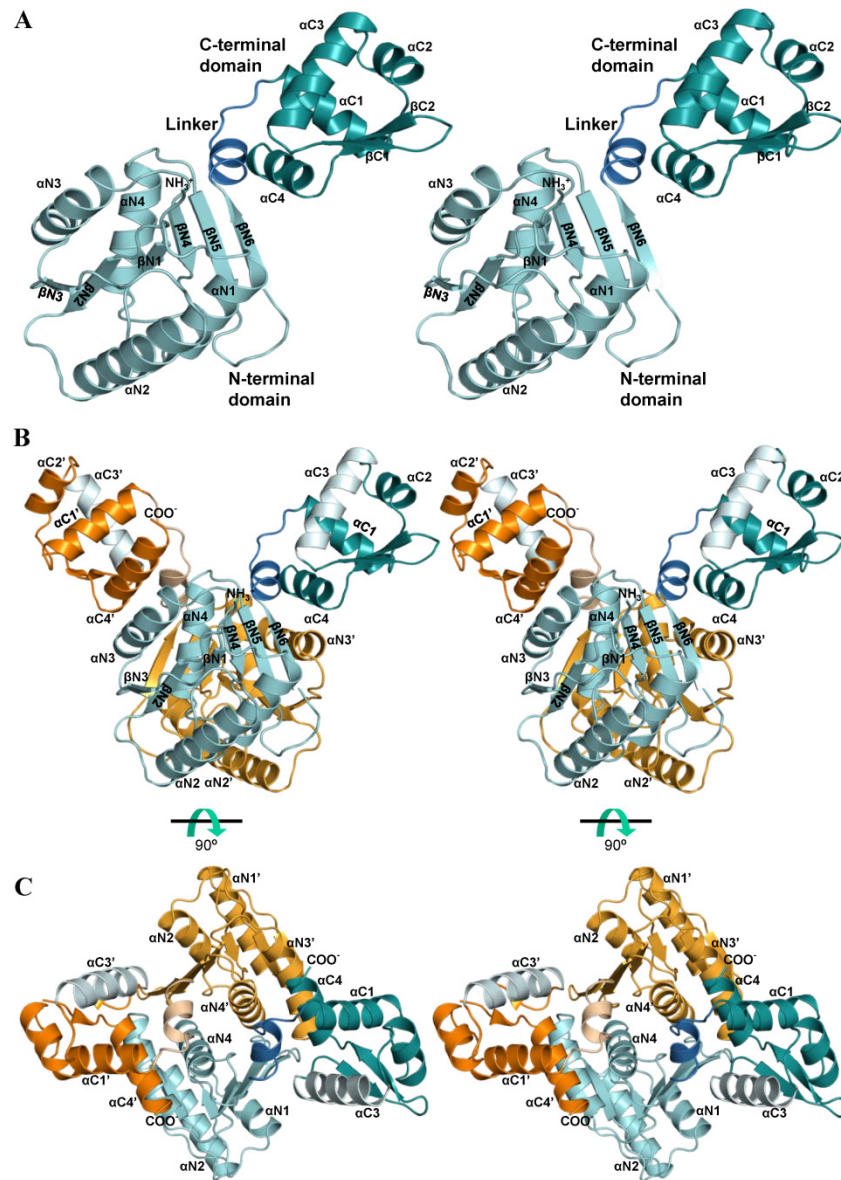


Figure 4-1: Csa3 is a dimer with two domains per subunit. (A) Stereo ribbon diagram of the Csa3 protomer. The N-terminal domain is colored cyan, the linker in blue and the C-terminal domain in deep teal. Secondary structural elements are labeled by domain and in ascending order from the N- to C-termini. (B) Stereo ribbon diagram of the Csa3 dimer. The N-terminal domain, linker and C-terminal domains are colored cyan, light blue and deep teal in chain A and light orange, tan and orange in chain B. The recognition helices in both chains are colored pale blue. (C) Stereo view of the Csa3 dimer looking down upon the winged helix-turn helix domains. Relative to panel B, the dimer has been rotated 90° about a horizontal axis in the plane of the page.

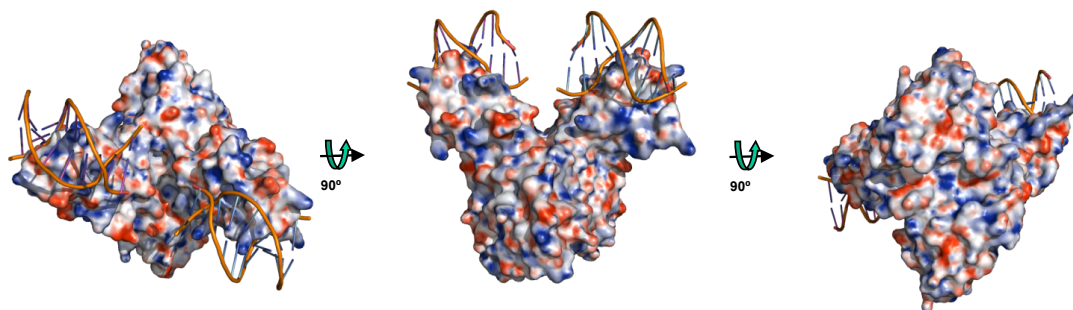


Figure 4-2: Surface electrostatics of Csa3. Surface electrostatics were calculated using APBS Tools with an ionic strength of 0.15 M and visualized with Pymol. The color ramp of the surface is from -25 Kt/e (red, acidic) to 25 Kt/e (blue, basic). A cartoon representation of DNA was included to show the hypothetical DNA-binding site.

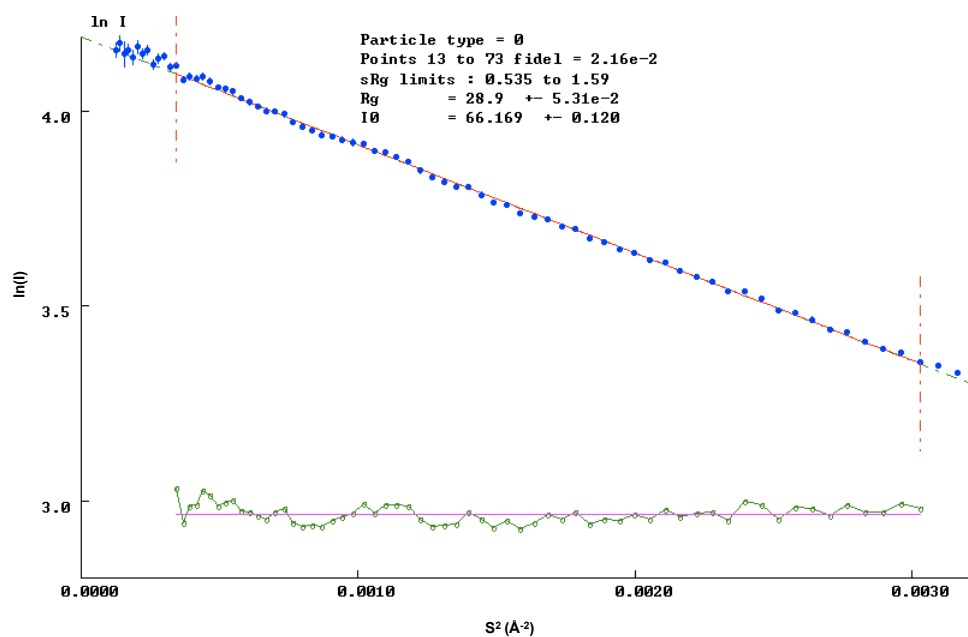


Figure 4-3: The Guinier Plot of Csa3 scattering indicates a monodisperse sample. The radius of gyration is 28.9 Å and the sample is monodisperse. The linear fit is shown in red. The green line depicts the deviation from the linear fit as a function of s^2 .

In order to identify the solution state dimer, the experimental scattering curve was compared to the hypothetical scattering curves of a Csa3 monomer and all potential dimers found in the crystal (Figure 4-4B). Only the dimer comprised of chains A and B from a single asymmetric unit gave a hypothetical scattering curve in agreement with the experimental scattering curve ($\chi= 3.223$) (Figure 4-4B). Comparison of the other potential dimers with the experimental scattering yielded χ values of 11.566, 12.221, 10.424 and 24.739. Small discrepancies between the curves are likely due to the presence of the disordered His₆-tag and C-termini in the crystal structure (97) or slight changes in the Csa3 conformation due to crystal packing. Visual examination of the Csa3 crystal structure and an analysis of crystal packing by the PISA server (110) also supported assignment of the A/B pair as the solution state dimer.

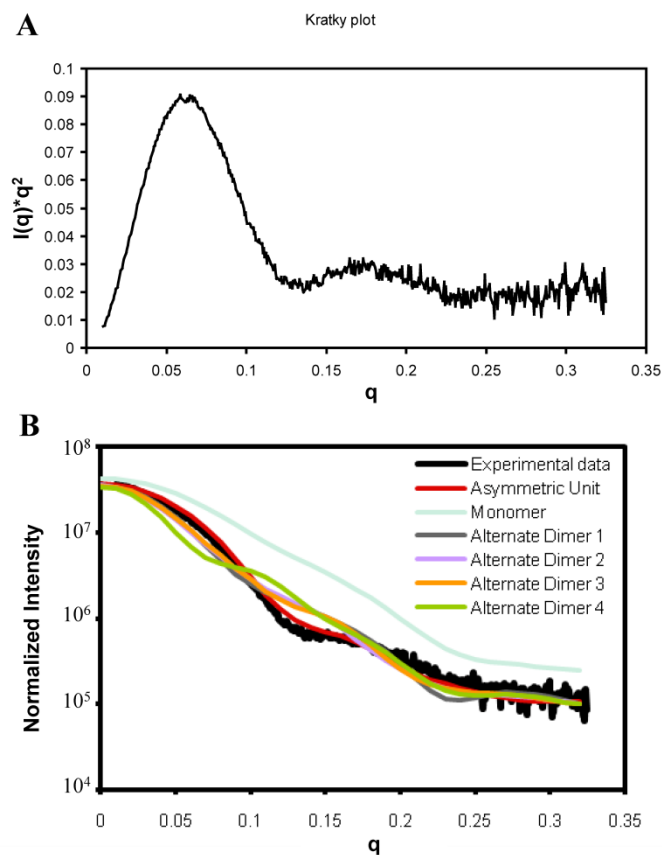


Figure 4-4: Validation of the biological unit by SAXS. (A) The parabolic nature of the Kratky plot at low values of q indicates that Csa3 is well folded. (B) The solution dimer corresponds to the crystallographic dimer present in a single asymmetric unit. The observed scattering curve of Csa3 (black) was compared with hypothetical scattering curves for the asymmetric unit (red), a potential monomer (light blue), and four alternative dimers (blue, violet, orange and green). Only the hypothetical scattering curve for the dimer corresponding to the crystallographic asymmetric unit agrees well with the observed scattering.

Structure of the Csa3 Dimer:

Dimer formation buries $\sim 2,200 \text{ \AA}^2$ of solvent accessible surface area per subunit, or 18.5%. The dimer interface includes 24 hydrophobic residues per subunit, complemented by 9 inter-subunit salt-bridges and 21 inter-subunit hydrogen bonds. One principal component is the interaction between the N-terminal domains (N_A - N_B), which

accounts for approximately half the interface. The remainder comes from interactions between the N- and C-terminal domains in symmetry related subunits. Thus, an additional one quarter of the buried surface area per subunit is found on the N-terminal domain, where it interacts with the C-terminal domain of a symmetry related chain (N_A-C_B), and in a mirror image, another quarter is on the C-terminal domain, where it interacts with the symmetry related N-terminal domain (C_A-N_B). When analyzed singly by the PISA server (110), both the N_A-N_B and the N_A-C_B interfaces give complexation significance scores of 1.000, implying that each interface plays an essential role in complex formation.

An interesting facet of the N_A-N_B interface is the role of helix $\alpha N4$. In the traditional dinucleotide-binding domain, strands $\beta N5$ and $\beta N6$ are covered by an additional helix that runs parallel to $\alpha N4$, and connects strands $\beta N5$ and $\beta N6$. Because of the reverse turn connecting $\beta N5$ and $\beta N6$ (discussed above), this covering helix is lost in the Csa3 subunit. However, dimer formation restores this structural element, placing $\alpha N4'$ of the symmetry related subunit in an equivalent position, where it runs parallel to helix $\alpha N4$, while covering $\beta N5$ and $\beta N6$ (Figure 4-5). Thus, one function of the $\beta N5$ - $\beta N6$ reverse turn is the construction of the Csa3 dimer interface.

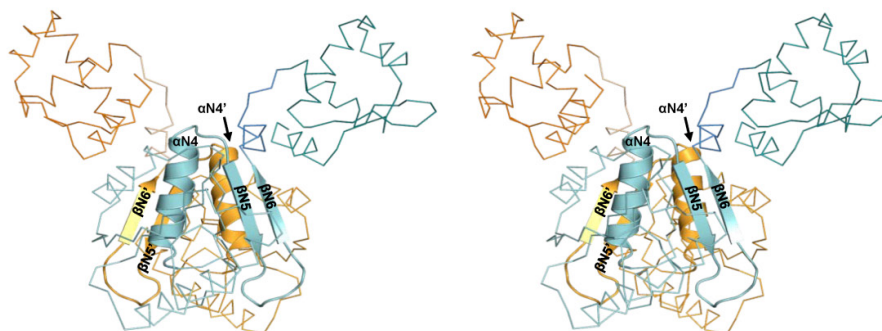


Figure 4-5: Csa3 dimer highlighting the interaction of helix N4 with strands N5' and N6' across the dimer interface. Helix N4 of the opposite chain fulfills the structural role of the $\beta 5$ - $\beta 6$ crossover in the canonical dinucleotide-binding domain.

The role of $\alpha C4$, the last helix in the wHTH domain is also noteworthy. In Mar-R like proteins this helix frequently serves to orient the wHTH domain with respect to other domains or subunits (30, 34, 106, 107, 111, 112). Accordingly, in Csa3 $\alpha C4$ is found at the domain interface where it contacts the linker helix in the same subunit and $\alpha N3'$ in the N-terminal domain of the symmetry related subunit (Figure 4-1). These contacts appear to anchor the position of the C-terminal wHTH domain with respect to the N-terminal domains of the Csa3 dimer, and thus, also fix the position of the wHTH domains with respect to each other, at least in the crystal.

Potential Csa3/DNA Interactions

A Dali search (84) identified the MexR repressor from the multidrug efflux operon of *Pseudomonas aeruginosa* as the nearest structural homologue to the Csa3 wHTH domain [PDB ID 1LNW, Z=12.5, 1.1 Å RMSD for 68 equivalent residues, 16% identity (113)]. However, the best match to a protein-DNA complex was OhrR, a transcription

factor from *Bacillus subtilis* [PDB ID 1Z9C, Z=10.8, 1.5 Å RMSD for 68 equivalent residues, 18% identity (114)]. Like Csa3, MexR and OhrR are dimeric 2-domain proteins whose structures consist of a dimerization domain, which is unrelated to the Csa3 N-terminal domain, and a MarR-family winged HTH DNA-binding domain. For OhrR, the dimerization domain includes an allosteric effector site where, under oxidative stress, Cys15 of OhrR is oxidized to Cys-sulphenic acid, resulting in derepression of the *ohrA* promoter (114).

Superposition of the OhrR/DNA complex on the wHTH domain of Csa3 shows that many features responsible for DNA recognition in OhrR are also present in Csa3. These include putative insertion of the recognition helix (α C3) into the major groove, with multiple opportunities for residues 172-184 to make base specific interactions or to interact with the ribose-phosphate backbone. Notably, of the residues positioned to make base-specific contacts, Thr175 is the best conserved among the Csa3 orthologues (Figure 4-7). Like most wHTH proteins, OhrR also uses the wing to interact with the ribose-phosphate backbone. Similarly, the Csa3 wing includes four basic residues, Lys192, Lys194, Arg196 and Lys197 that are likely to serve a similar function (Figure 4-6B). Relative to many wHTH proteins, however, the wing in OhrR is quite long, which also allows the tip of the wing to access the minor groove for base specific contacts. In contrast, the Csa3 wing, which is 5 residues shorter than the wing in OhrR, is more typical in length and appears too short to access the minor groove (Figures 4-6 and 4-7).

The similarity to OhrR suggests two additional sets of protein/DNA interactions. First, like OhrR, it appears that helix α C2 will contribute to DNA recognition by utilizing

the N-terminal end of the helix dipole to interact with phosphate in the DNA backbone. Second, OhrR also utilizes a “helix-helix” motif in DNA recognition (114), which in Csa3, corresponds to the linker helix and helix α C1. The superpositional docking suggests these helices may also play a minor role in DNA recognition, perhaps utilizing Arg141 in the connecting loop, and Arg145 at the N-terminus of helix C1 to facilitate interactions with the ribose-phosphate backbone (Figure 4-6).

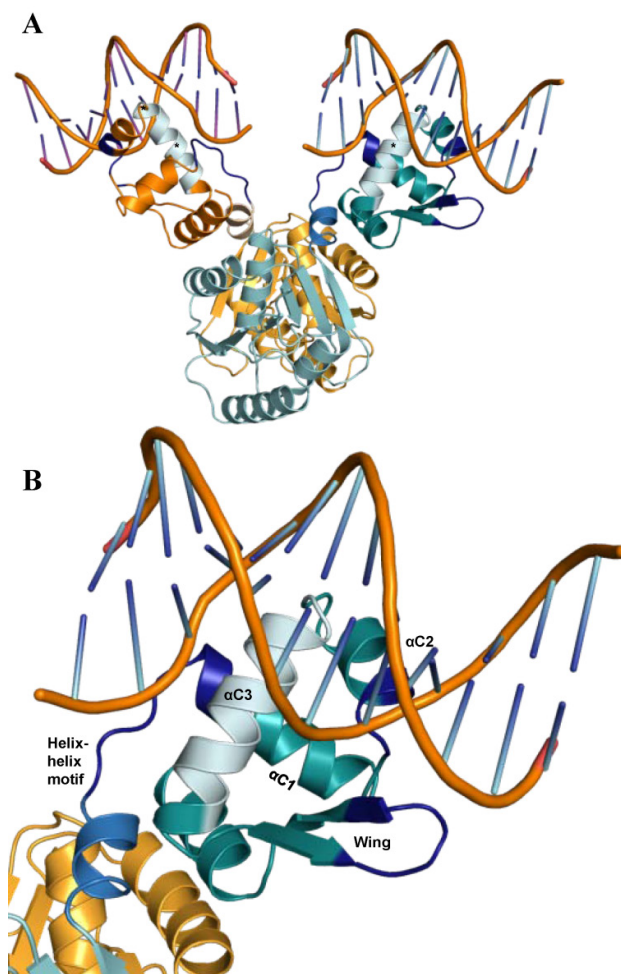


Figure 4-6: Superpositional docking of DNA to Csa3. (A) The Csa3 dimer is docked to the *ohrR* promoter. Csa3 is colored as in Fig. 5-1 except that the recognition helix is colored light teal and the additional putative DNA-interacting regions are colored blue. (B) Close-up view of the chain A wHTH domain docked to DNA. Structural features predicted to interact with the DNA are labeled.

The dimeric, two-fold symmetric structure of Csa3 suggests that the two wHTH domains will simultaneously bind to DNA, recognizing a palindromic or pseudo-palindromic sequence. However, while the OhrR DNA can be docked by superposition to a single chain of Csa3, it could not be simultaneously docked to both chains of Csa3 without a significant steric clash. Similarly, we were unable to satisfactorily dock generic B-form DNA to the Csa3 dimer. This suggests the crystallized conformation of Csa3 is unfavorable for DNA binding, or that Csa3 will recognize distorted B-form DNA. Indeed, the analogy to OhrR supports both possibilities. DNA binding by OhrR was observed to cause a 25° rotation of the winged HTH subunit and the OhrR DNA is bent by 10° and unwound by 1.4°.

We thus asked whether the small angle X-ray scattering data might indicate movement of the wHTH domains relative to each other within the Csa3 dimer. To this end, the small angle X-ray scattering data were re-examined with BILBO-MD, a program that generates a series of models by molecular dynamics and identifies models that agree with the experimental scattering data (115). However, no apparent domain movement was identified. While this does not rule out interdomain movement in DNA recognition, the SAXS data analysis suggests the anchoring N-C interface restricts conformational heterogeneity in Csa3, at least in the absence of DNA or some regulatory ligand.

A Putative Ligand-Binding Site in the N-Terminal Domain

As mentioned above, the Gly-X-Gly-X-X-Ala/Gly motif that is a hallmark of the dinucleotide binding domain is absent in Csa3. In addition, superpositional docking with

NAD(P)H or FADH₂ results in severe steric clashes within a single subunit, and even greater clashes when considering the Csa3 dimer. Thus, it seems Csa3 is unlikely to bind these dinucleotides.

To identify surface features that may be functionally important, we further examined the conserved residues in the N-terminal domain of Csa3. Because many members of COG0640 are not CRISPR associated, that is, they are not Csa3 proteins (58), the genome properties tool (88) was used to query the Comprehensive Microbial Resource Database for specific *csa3* genes, with the additional stipulation that they were encoded in genomes that also contained *cas1*, *cas2* and at least one *CRISPR* locus. All the Csa3 orthologues identified are within 9 kb of the nearest *CRISPR* locus and adjacent to, or part of, putative Cas operons. Ten sequences were retrieved and aligned using 3D-COFFEE (116). The multiple sequence alignments reveal two conserved sequence motifs in the N-terminal domain. The first motif is Thr-h-Gly-Phe-(Asn/Asp)-Glu-X₄-Arg, where h represents a hydrophobic residue, and is present in the β N1- α N1 loop. The second motif, Leu-X₂-Gly-h-Arg, is found within the β N4- α N4 loop (Figure 4-7).

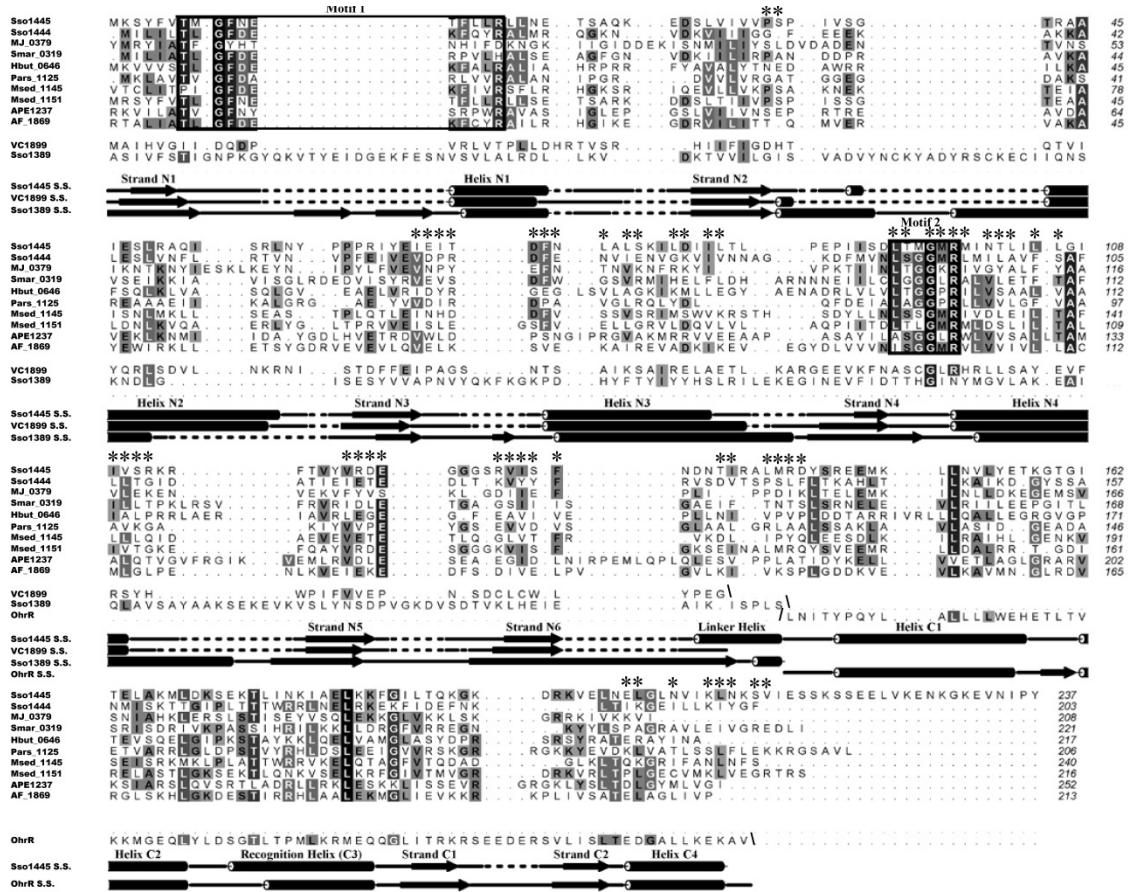


Figure 4-7: Two conserved sequence motifs are found in the Csa3 N-terminal domain. The genome properties tool (88) and CRISPR database (45) were used to identify *csa3* genes that are near clusters of *cas* genes and at least one *CRISPR* locus. The Csa3 sequences were aligned using 3D-COFFEE (116). The two sequence motifs which are conserved among Csa3 orthologs are outlined with boxes. Csa3 residues contributing to the dimer interface are marked with asterisks (*). Structure-based pairwise alignments of the VC1899 and Sso1389 (Csa3) N-terminal domains and the OhrR Whth domain with Sso1445 were calculated using Dali (84). Secondary structural elements of Sso1445 (Csa3), VC1899, Sso1389 (Csx1) and OhrR were identified using DSSP. The sequence motifs, which are conserved among Csa3 orthologs are not present in VC1899 or Csx1. Csx1 has a 16-residue insertion at the site corresponding to Motif 1 in Csa3. The figure was generated using ALINE (117).

In addition to these two larger motifs, Glu122 in the reverse turn connecting strands β N5 and β N6 is also strongly conserved. Interestingly, the residues in motifs 1 and 2, and

the β N5- β N6 loop harboring Glu122 are spatially close to each other on the surface of the N-terminal domain, where they form a single, prominent cleft that spans the dimer interface. Thus, in addition to contributing to the architecture of the subunit interface, the reverse turn connecting β N5 and β N6 also defines two symmetry related walls of this cleft. Two additional walls are formed by the 2 copies of the β N1- α N1 loop, while the β N4- α N4 loop forms the floor. This conserved cleft, which is distal to the winged HTH domain, is approximately 35 Å long (C_α of Glu67 chain A to C_α of Glu67 chain B), 9 Å deep (C_Z of Phe10 to main-chain N of Mse97), and varies between 11 (C_α of Gly132 chain A to C_β of Val 39 chain B) and 18 Å (C_β of Phe10 chain A to C_β of Phe10 chain B) in width. The side-chain of Phe10 (motif 1) extends out over the cleft, while the Arg98 side-chains (motif 2) lie on the floor of the cleft (Figure 4-8).

In all, this cleft has the hallmarks of a conserved ligand binding site (118). Because the cleft spans the 2-fold symmetric dimer interface, a potential ligand could mirror this symmetry. Alternatively, the cleft might accommodate an asymmetric ligand in either of two equivalent orientations. Interestingly, difference maps identified four-sigma difference electron density within the cleft, centered on the two-fold axis. This density most likely represents a component of the crystallization mother liquor or a series of several overlapping partial-occupancy water sites. We have tentatively modeled this density as a fragment of polyethylene glycol (PEG), a component of the crystal mother liquor, at 50% occupancy. Weaker electron density is present in other regions of the pocket, suggestive of additional ordered interactions with PEG, water or other mother liquor components, but has not been modeled.

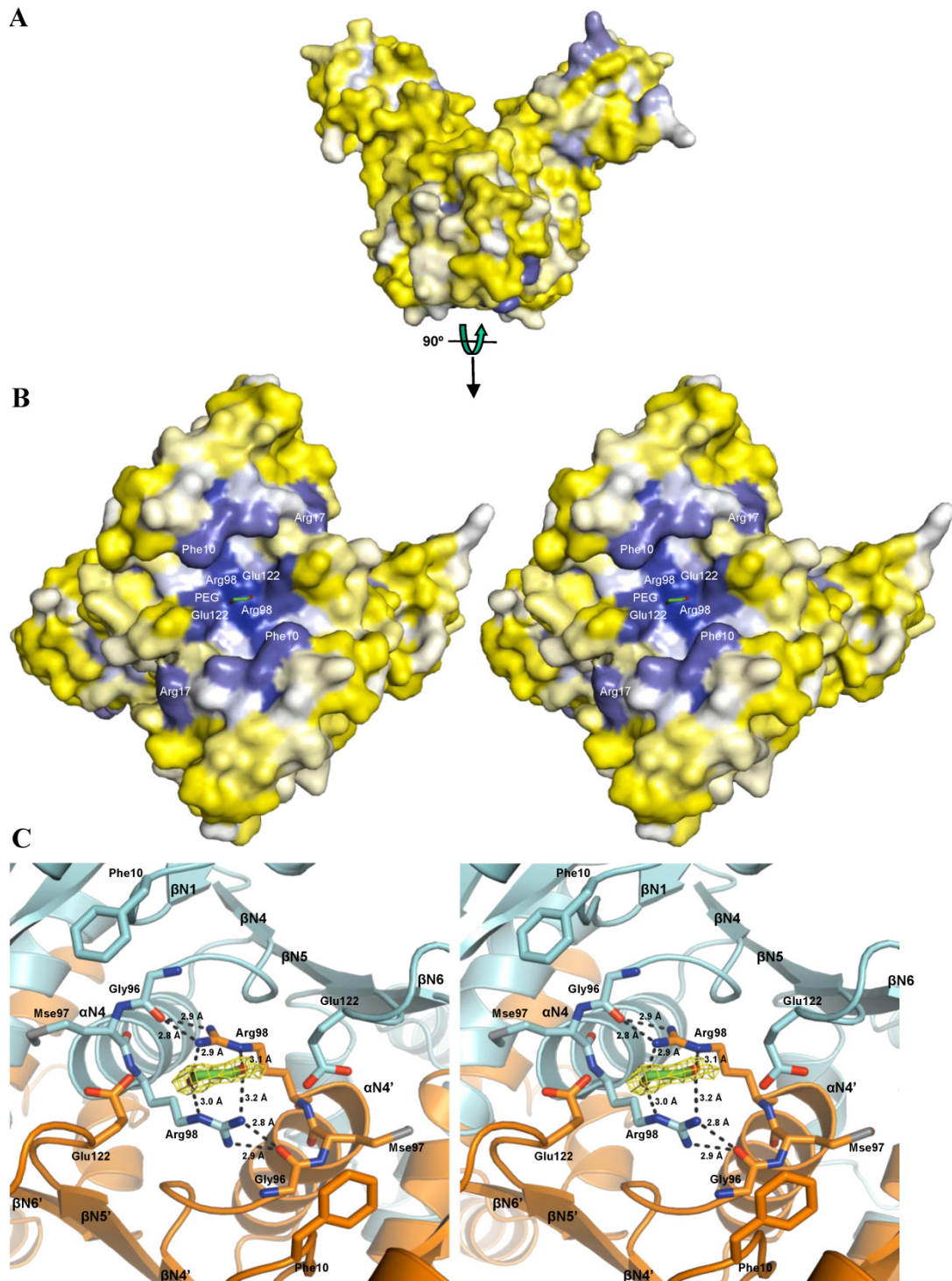


Figure 4-8: Putative ligand-binding pocket on the Csa3 N-terminal domain.

Figure 4-8 (Continued) (A) Surface conservation of Csa3. Residues are colored based on their relative conservation among the Csa3 orthologs using Consurf (119). Blue residues are more conserved and yellow residues are less conserved. (B) Stereo image of the conserved pocket on the N-terminal domain. The most highly conserved residues are labeled. Because the pocket spans the dimer interface, there are two copies of each residue. A four-atom length of ordered PEG is shown in stick rendering. The oxygen atoms are colored red and the carbon atoms are colored green. (C) Close-up view of the putative ligand-binding pocket. Chain A is colored cyan and chain B is colored orange. Phe10, Glu122, and the Gly-h-Arg motif (Gly96, selenomethionine (Mse) 97, and Arg98) are shown in stick rendering with nitrogen atoms in blue and oxygen atoms in red. The four atoms of ordered PEG are shown in stick rendering with the carbon atoms colored green and oxygen atoms red. The yellow mesh depicts the 2Fo-Fc omit map for the PEG contoured at 1σ . The black dotted lines depict inter-protomer hydrogen bonds between Arg98 and Gly96, and hydrogen bonds between Arg98 and the ordered PEG.

Structural Homology to the Csx1 Family of CRISPR-Associated Proteins

The N-terminal domain of Csa3 was used to query the protein databank using Dali (84). It was found to be structurally homologous to a similar domain in members of the COG1517-like and Csx1 families of CRISPR associated proteins. As anticipated by Makarova *et al.* (59), the N-terminal domain of Csa3 superimposes on the N-terminal domain of *Vibrio cholera* VC1899 with a 2.3 Å RMSD (PDB ID: 1XMX, Z = 14.2, 124 equivalent residues, 12% identity, Midwest Center for Structural Genomics, unpublished). The Csa3 N-terminal domain also superimposes on Sso1389 from *S. solfataricus* with a 3.9 Å RMSD (PDB ID: 2I71, Z = 8.8, 128 equivalent residues, 14% identity, Midwest Center for Structural Genomics, unpublished). The VC1899 and Csx1 N-terminal domains (residues 1-122 and 1-186 respectively) were realigned with the Csa3 N-terminal domain plus the linker (residues 1-144) using the SSM server yielding an rmsd of 2.05 over 118 residues for the Csa3-VC1899 alignment and an rmsd of 3.18 Å over 102 residues for the Csa3-Csx1 alignment (Figure 4-9).

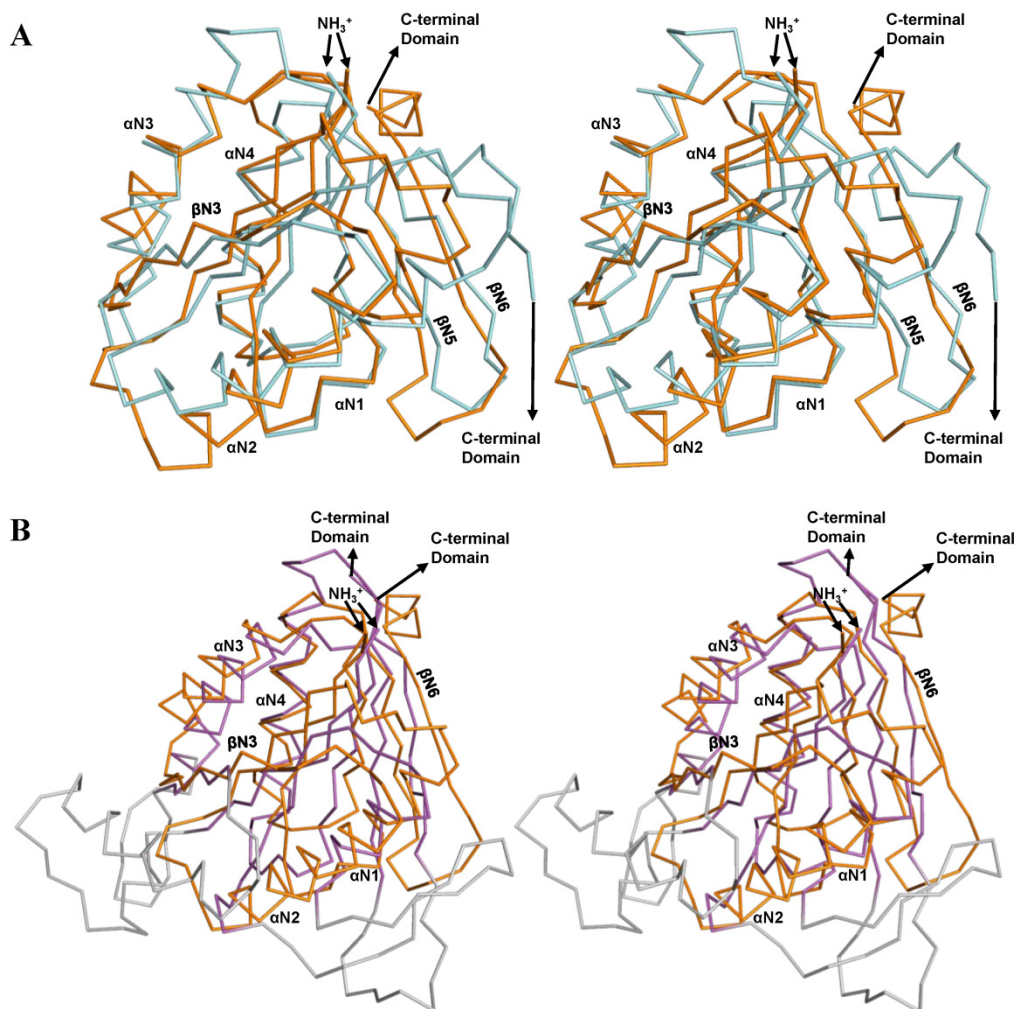


Figure 4-9: The N-terminal domain of Csa3 is structurally similar to the N-terminal domains of VC1899 and Csx1. (A) SSM Structural alignment of the Csa3 N-terminal domain (orange) with the VC1899 N-terminal domain (cyan). (B) SSM Structural alignment of the Csa3 N-terminal domain (orange) with the Csx1 N-terminal domain (violet and gray). In addition to the core di-nucleotide binding domain-fold, which it shares with Csa3, the Csx1 N-terminal domains has several additional structural features inserted into the loops (gray).

However, these are distant relationships, which are not currently detected by an NCBI conserved domain database (CDD) search (120), or an Hhpred (121) query of the COG database. Although Csa3, VC1899 and Csx1 share a common fold for their N-

terminal domains, they are dissimilar with regard to their C-terminal domains and overall domain architecture. Indeed, relative to Csa3 and Csx1, the C-terminal domains of VC1899 are found on the opposite face of the dimeric N-terminal domains (Figure 4-10). While specific functions for Csa3, VC1899 and Sso1389 remain to be determined, collectively, their structures suggest functions related to nucleic acid binding and/or nucleic acid metabolism.

As noted above, not all proteins containing the Csa3-like N-terminal domain are expected to be CRISPR associated proteins (58, 59), and this appears to be the case for VC1899. The O1 biovar El Tor strain N16961, from which VC1899 hails, does not harbor a CRISPR. In addition, for *Vibrio* genomes that do harbor CRISPRs, they are far from any VC1899 orthologs. For these reasons, the relevance of VC1899 to CRISPR/Cas biology is uncertain.

In contrast, Sso1389, which contains a C-terminal domain of unknown function, is a clear member of the Csx1 and COG1517 families of CRISPR associated proteins. Relative to the Csa3 N-terminal domain, the N-terminal domain of Sso1389 contains an extra helix inserted between β N2 and α N2 (Csa3 numbering), and two additional β -strands running along each edge of the central beta sheet, yielding a 10 stranded mixed β -sheet. Two of these strands, connected by a reverse turn, are inserted into the β N1- α N1 loop and run along β N3. The other two strands run along β N6, and are provided by an extension from the C-terminal domain of unknown function (Figure 4-10).

Visual examination of the Sso1389 crystal structure reveals a crystallographic dimer that is mediated, in large part, by the N-terminal domain of Csx1. The interdomain

interactions are strikingly similar to those in the Csa3 N-terminal domain, although there are fewer hydrophobic residues at the interfaces, particularly on helix α N4. This suggests Sso1389 might also form a Csa3-like dimer in solution, though the association may not be as tight as that seen for Csa3. This conclusion is supported by the PISA server, which gives a complexation significance score of 0.639, suggesting this interface could play an important role in the formation of a putative Sso1389 dimer. VC1899 also appears to form a dimer with an interface similar to Csa3 and Csx1 (Figure 4-10). Thus, the dimer interface observed in Csa3 and Csx1 appears to be a common feature of the Csa3 and COG1517 N-terminal domains.

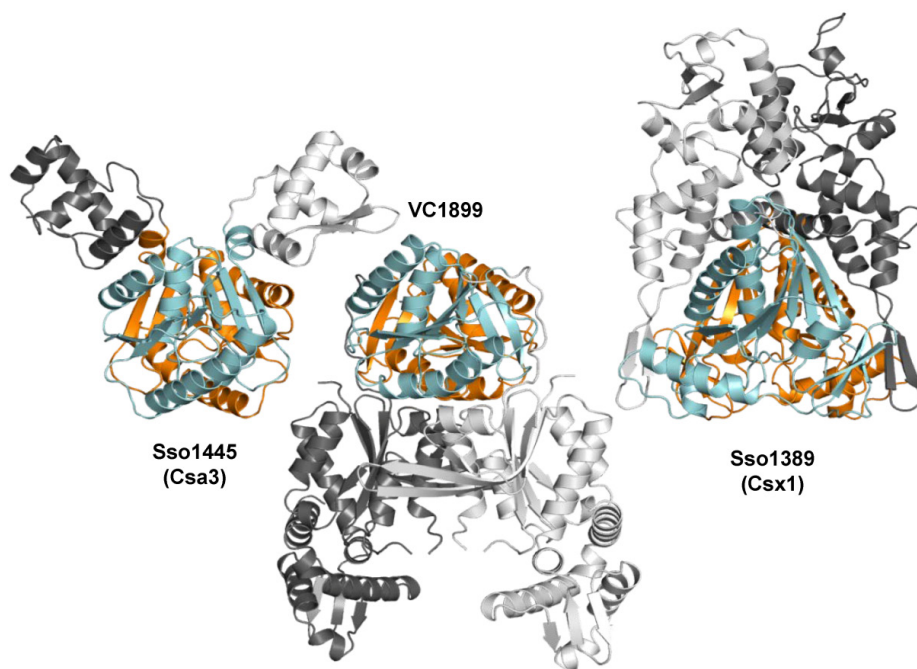


Figure 4-10: Csa3, VC1899 and Csx1 share similar N-terminal domains but differ in overall domain architecture. Csa3, VC1899, and Sso1389 (Csx1) are shown in equivalent orientations based on structural superposition of their N-terminal domains (cyan and orange). The structurally dissimilar C-terminal domains are shown in shades of grey. Note that the C-terminal domains in VC1899 are found on the opposite side of the Csa3-like N-terminal domain.

While the conserved sequence motifs in the N-terminal domain of Csa3 are not found in Sso1389/Csx1, this putative Csx1 dimer does contain a large cleft in the N-terminal domain that spans the dimer interface. In addition, a sequence alignment of CRISPR associated Csx1 proteins identifies three conserved sequence motifs in the Csx1 family that line this cleft (Supplemental Figures S5 and S6), suggesting that this may be a functional ligand binding site in Csx1 as well. However, the lack of sequence similarity between Csx1 and Csa3 families suggests that if these are functional ligand binding sites, the identity of the ligand may differ. Finally, a fourth motif is found in the Csx1 C-terminal domain of unknown function, suggesting a second possible ligand binding site in this protein family (Figures 5-11 and 5-12).

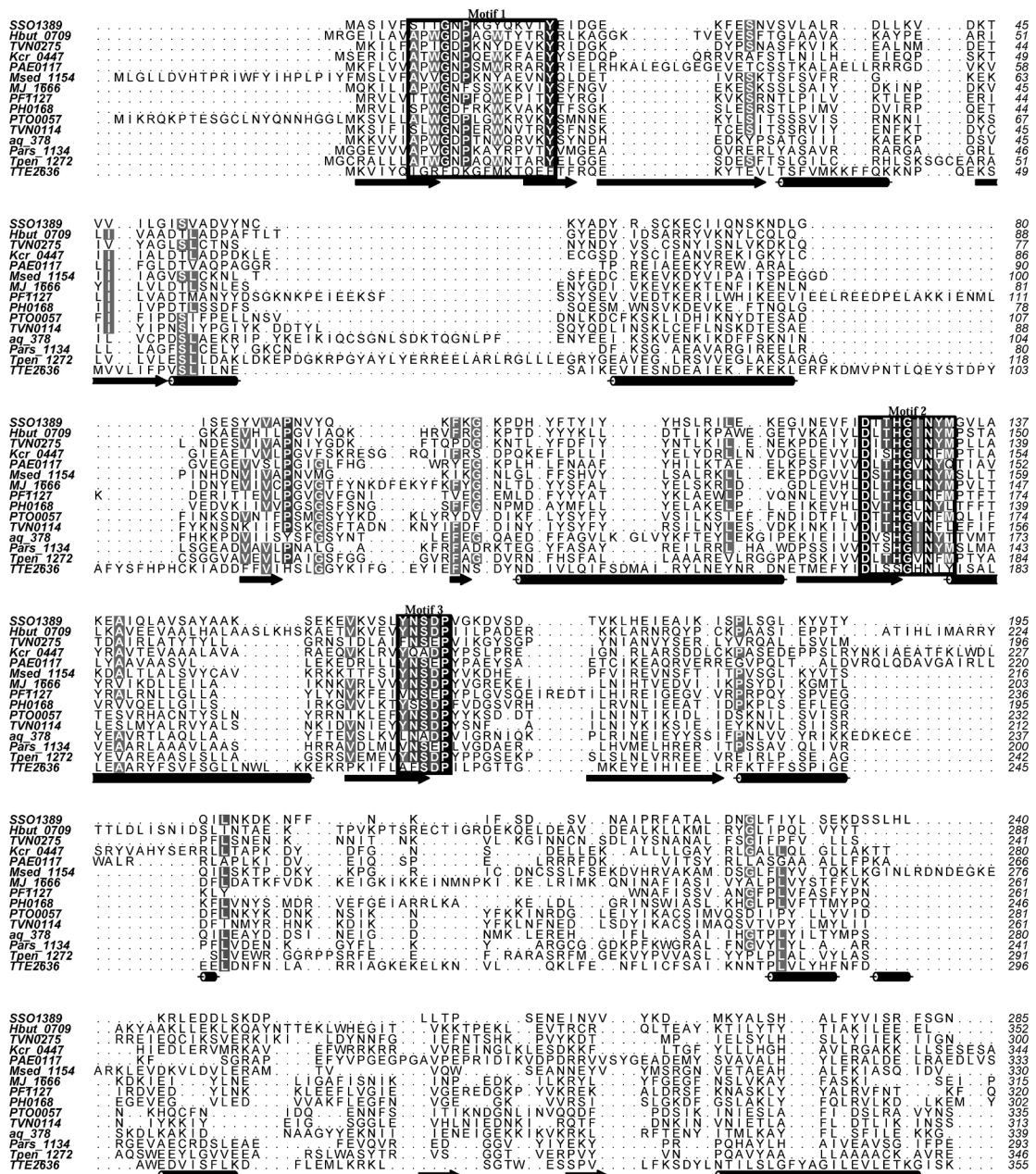


Figure 4-11: There are four sequence motifs that are conserved among Csx1 orthologues.

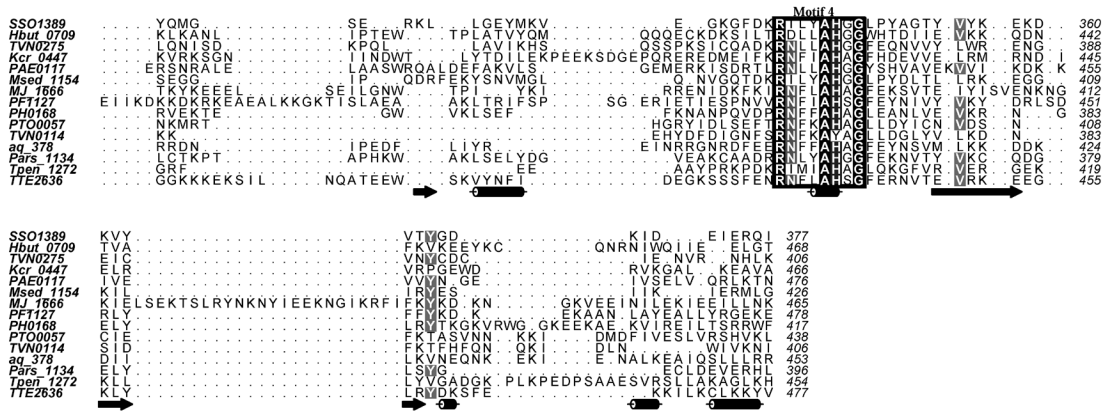


Figure 4-11 (Continued): There are four sequence motifs that are conserved among Csxl orthologues. All genomes in the comprehensive microbial resource databank were searched for Csxl with the TIGR01897 HMM. All identified orthologues are from genomes that include at least one CRISPR. 15 retrieved sequences were retrieved and aligned using 3D-Coffee followed by manual adjustment around the gaps. Secondary structural elements were calculated using DSSP from the Sso1389 structure (2I71). Motifs 1 (Ser/Ala-h₂-Gly-Asn/Asp-Pro-X₇-Tyr), 2 (Asp-X-Thr-His-Gly-h-Asn-Tyr/Phe-h) and 3 (Tyr-Asn-Ser-Asp/Glu-Pro) are found in a cleft on the N-terminal domain that spans the dimer interface. Motif 4 (Arg-X₃-Ala-His-Gly/Ala-Gly) is found in a cleft on the C-terminal domain that also spans the dimer interface and is distal to the winged HTH domain. The figure was generated using ALINE.

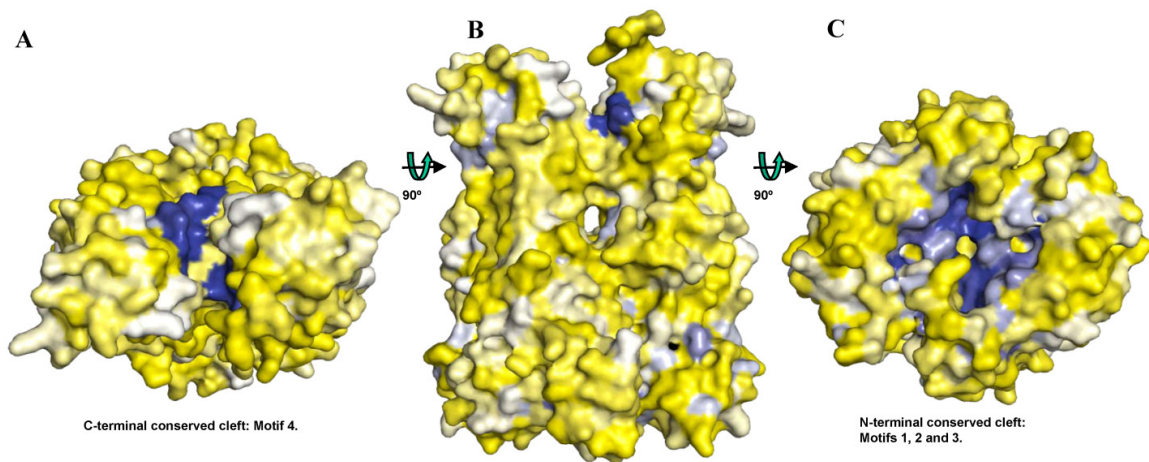


Figure 4-12: Csxl has conserved 2-fold symmetric clefts that span the dimer interface on the N- and C-terminal domains. Consurf image was generated using the multiple sequence alignment shown in figure 4-11. (A) Cleft on the C-terminal domain containing motif 4. (B) Csxl shown in the orientation equivalent to figure 4-10. (C) Conserved cleft on the N-terminal domain containing motifs 1, 2, and 3.

Discussion and Concluding Remarks

While the molecular mechanisms of CRISPR/Cas systems are an area of active research, little is known regarding the transcriptional regulation of CRISPR-Cas in most organisms. Present insight comes from *E. coli* K12, where transcription of CRISPR-I and the Cas operon encoding Cse1-4, Cas5e and Cas1-2 is suppressed by heat-stable nucleoid-structuring (H-NS) protein (72) and derepressed by the transcriptional activator LeuO (73). In addition, some elements of the *T. thermophilus* CRIPR-Cas system appear to be controlled by the Camp receptor protein (122). However, it's clear that other regulatory factors remain to be identified in *T. thermophilus* (122), and perhaps in *E. coli* as well.

The first structure of a Csa3 protein, determined using X-ray crystallography and solution state small-angle X-ray scattering, suggests that Csa3 is a transcriptional regulator with a novel binding site for an allosteric effector in the conserved cleft of the N-terminal domain. The presence of this conserved ligand-binding site in the N-terminal domain of Csa3 suggests the existence of a small-molecule regulator of archaeal CRISPR-Cas systems. In this light, the identity of a controlling allosteric effector for Csa3 is of significant interest. The putative Csa3 ligand-binding site is two-fold symmetric and includes a strictly conserved positively charged residue (Arg98), and a strictly conserved solvent-accessible aromatic sidechain (Phe10). While we cannot rule out the possibility that conserved cleft binds one or two copies of an asymmetric small-molecule regulator, the structure of Csa3 suggests the ligand may be a two-fold symmetric molecule with negatively charged and hydrophobic or aromatic moieties. Known small-molecule

regulators that exhibit these characteristics include a number of dinucleoside polyphosphates (123, 124).

Alternatively, Csa3 might function as part of a larger complex during adaptation or interference, perhaps in the recognition of foreign DNA, or the generation and integration of new spacers. In either case, the conserved cleft in the N-terminal domain might serve to recruit Csa3 into a larger complex. We thus attempted to identify proteins that interact with Csa3, both by incubating immobilized Csa3 with *Sulfolobus solfataricus* lysate and by purifying tandem-affinity tagged Csa3 expressed in *Sulfolobus solfataricus*. Neither approach was successful in identifying partners for Csa3 (data not shown). While the negative result may be due to the presence of the affinity tag, a loose interaction, or low levels of the partner protein, it is also consistent with our first hypothesis, that the conserved cleft represents a small-molecule ligand-binding site.

Also of interest are the identities of any transcriptional targets that might be controlled by Csa3. One possibility, of course, is that Csa3 might regulate expression of the neighboring *CRISPRs* (*CRISPR C* and *D*). However, using the recombinant apo-protein, we were unable to demonstrate recognition of the putative promoters using an electromobility shift assay (data not shown). Whether DNA recognition requires the presence of an unidentified regulatory ligand, or Csa3 recognizes DNA in some other context, remains to be determined. Efforts to identify conserved (pseudo-)palindromic motifs by *in silico* analysis of the *csa3* genomic neighborhood in *S. Solfataricus*, or in the immediate neighborhoods surrounding *csa3* orthologs, were unsuccessful.

Importantly, the structural findings in regards to Csa3 can be extended to Csx1, a second CRISPR-associated protein, and the non-CRISPR-associated protein, VC1899. Both are proteins of unknown function solved as part of structural genomics efforts. While Csa3 (COG0640), Csx1 (COG1517) and VC1899 (predicted to be similar to COG1517 (59)) have different overall structures, they share a common N-terminal domain. The fold of this domain is similar to that of a dinucleotide binding domain with the critical modification that strand 6 of the 6-stranded β -sheet (Csa3 numbering) runs antiparallel to strands 1-5. In all three structures this antiparallel sixth strand negates the need for one helical crossover and facilitates formation of a common dimer interface. The presence of a similar dimer interface in all three crystal structures, and in solution for Csa3, suggests that dimerization is one function of this unique fold. Accurate determination of the biological unit from crystal structures can sometimes be a non-trivial issue. Thus, the biological unit may be incorrectly assigned for VC1899 (monomer) and ambiguously assigned for Csx1 (two possible dimers) in their respective protein databank entries (PDB IDs: 1XMX and 2I71 www.rcsb.org). This underscores the inherent synergy between X-ray crystallography, which provides an accurate atomic model of Csa3, and small-angle X-ray scattering, which identifies the oligomeric state of Csa3 in solution. Together, as evidenced by recent results on such combined methods(125), the crystallographic and SAXS data provide significant new insight into the prokaryotic adaptive immune system, as well as an accurate structural foundation for further investigation of the roles of Csa3 and Csx1 in CRISPR-Cas.

ISOLATION AND CHARACTERIZATION OF AN ARCHAEAL CASCADE

Introduction

CASCADE is a large protein-RNA machine that is at the heart of most CRISPR-directed DNA interference systems. CASCADE binds processed crRNA, recognizes complementary protospacers in dsDNA, which presumably requires a partial unwinding of the double helix, forms R-loops and RNA-DNA hybrids upon recognizing a protospacer and presumably recruits Cas3 to degrade the invading DNA (41, 69). The major protein of *E. coli* CASCADE, CasC, belongs to the COG1857-like superfamily of CRISPR-associated proteins (59). COG1857, along with COG1688 is the most phylogenetically widespread of the CASCADE components and are found in six of the eight CRISPR-Cas subtypes (41, 59, 69). This chapter describes the isolation and purification of *S. solfataricus* CASCADE, the first for an archaeon (which we thus named aCASCADE). The chapter also describes the structure of the major protein component of aCASCADE, Csa2 and the unique architecture of aCASCADE. This chapter also proposes a preliminary structural model for CASCADE. The results presented in this chapter are currently in preparation for submission to the EMBO journal.

ResultsIsolation of a CASCADE-like Protein-RNA Complex from *S. solfataricus*

The *S. solfataricus* genome encodes orthologs for two components of *E. coli* CASCADE, Csa2 (COG1857) and Cas5a (COG1688) (59). To determine whether Csa2

and Cas5a participate in a CASCADE-like complex in *S. solfataricus*, N-terminal tandem-affinity (StrepII and His₈) tagged Csa2 (Sso1442) was expressed in *S. solfataricus* strain PH1-16 and affinity purified using Strep-tactin resin followed by Ni-NTA resin. SDS-PAGE analysis of the purified material revealed several bands, which were identified by in-gel tryptic digest and LC-MS/MS (Figure 5-1A, Table 5-1). The major band migrated at 38 kDa and was identified as Csa2 (Sso1442), while a fainter 28 kDa band was identified as Cas5a (Sso1441). Two bands between the major Csa2 and Cas5a bands were also identified as Csa2 (Sso1442) and likely represent proteolyzed or endogenous, untagged Csa2. Two additional bands were identified as PccB and AccC, which are components of an unrelated biotinylated enzyme (126) complex that is recognized by streptactin. To confirm the relative abundance of Csa2 and Cas5a in the complex, a superose-6 column was used to separate the complex from any excess, uncomplexed Csa2 arising from overexpression. We observed a broad elution peak that was consistent with a molecular weight of approximately 350-500 kDa, and a minor amount of faster migrating material extending up to the void volume (Figure 5-2). Similar to *E. coli* CASCADE (41), Csa2(COG1857) was significantly more abundant in the complex than Cas5a(COG1688), even after fractionation on the superose-6 column (Figure 5-1A). We propose the term “aCASCADE” for this Aperm subtype archaeal complex.

Because *E. coli* CASCADE binds processed crRNA and DNA (69), we examined the aCASCADE complex for co-purifying nucleic acid using basic phenol-chloroform extraction followed by RNaseA or DNase I digestion. TBE-urea-PAGE with Sybr-gold staining revealed RNA species of 60-70 nt, plus low amounts of higher molecular weight

RNA that included faint bands at two and three times the molecular weight of the major species. The aCASCADE complex also co-purified with a smaller amount of high molecular weight (≥ 300 nt) DNA (Figure 5-1B). It is not clear whether this represents a specific crRNA-DNA interaction, or like *E. coli* CASCADE, the complex copurifies with nonspecifically bound DNA (41, 69). The RNA was found to co-purify with the aCASCADE complex through the streptactin, Ni-NTA and size-exclusion chromatography (Figure 5-1C). Further, when subjected to ribonuclease protection assays the RNA in the complex showed no visible degradation, even after 24 hours, indicating that the RNA is tightly bound and protected along its entire length (Figure 5-1D).

To confirm that the RNA was CRISPR-derived, the 60-70 nt band was gel-extracted, cloned, sequenced and compared with the *S. solfataricus* P2 genome and two available *S. solfataricus* P1 CRISPR sequences (47). Fifteen of 16 sequenced clones were clearly CRISPR-derived with fragments of direct repeat sequence on the 5' and 3' ends that were separated by variable spacer sequence (Figure 5-1E). All three CRISPR repeat sequences found in *S. solfataricus* P2 were represented among the clones, indicating the aCASCADE complex binds each type of crRNA. Clone 7 contained the shortest spacer, 38 bases, while the clone 6 spacer is the longest at 44 bases. Twelve clones contained spacers present in strain P2 CRISPRs and thus could be assigned to individual CRISPR loci (B, C, D and F) (Figure 5-1E). Three additional clones had spacers which were not present in the sequenced *S. solfataricus* P1 or P2 CRISPRs and based on the direct repeat sequence could belong to either CRISPR A or B. The longest clones represented a complete repeat-spacer unit with 8 nt of repeat sequence at the 5' end and 16-17 nt of

repeat at the 3' end, reminiscent of CRISPR transcript processing in *P. furiosus* by Cas6 (60). Some clones had shortened 5'-ends that could represent cloning artifacts. Several clones also had single nucleotide mismatches with the P2 CRISPR repeat sequences, potentially due to differences between the PH1-16 and P2/P1 strains or the use of an error-prone polymerase to amplify the cDNA.

To identify additional proteins that bind weakly or are present in lower abundance, purification of the aCASCADE complex was limited to the streptactin resin and analyzed by in-solution tryptic digest and LC-MS/MS in three independent experiments (Table 5-2). In addition to the expected Csa2 (Sso1442) and Cas5a (Sso1441), we also identified the crRNA processing endonuclease Cas6 (Sso1437) and Csa5 (Sso1443). Csa5 is a 150 residue protein of unknown function encoded immediately upstream from Csa2 in many archaeal genomes and may thus represent an archaeal-specific CASCADE subunit. We also identified two Csa2 paralogues (Sso1399 and Sso1997), two Cas5a paralogs (Sso1400 and Sso1998), the Csa5 paralog Sso1996 and the Cas6 paralog Sso2004. After size-exclusion chromatography, Csa2 (Sso1442) and Cas5a (Sso1441) copurified as expected, but the other proteins were not detectable by mass spectrometry, suggesting that they are more weakly associated in aCASCADE. Alternatively, the presence of multiple paralogs might be explained through bridging of aCASCADE paralogs by incompletely processed crRNA present during early stages of the purification (Figure 5-1B).

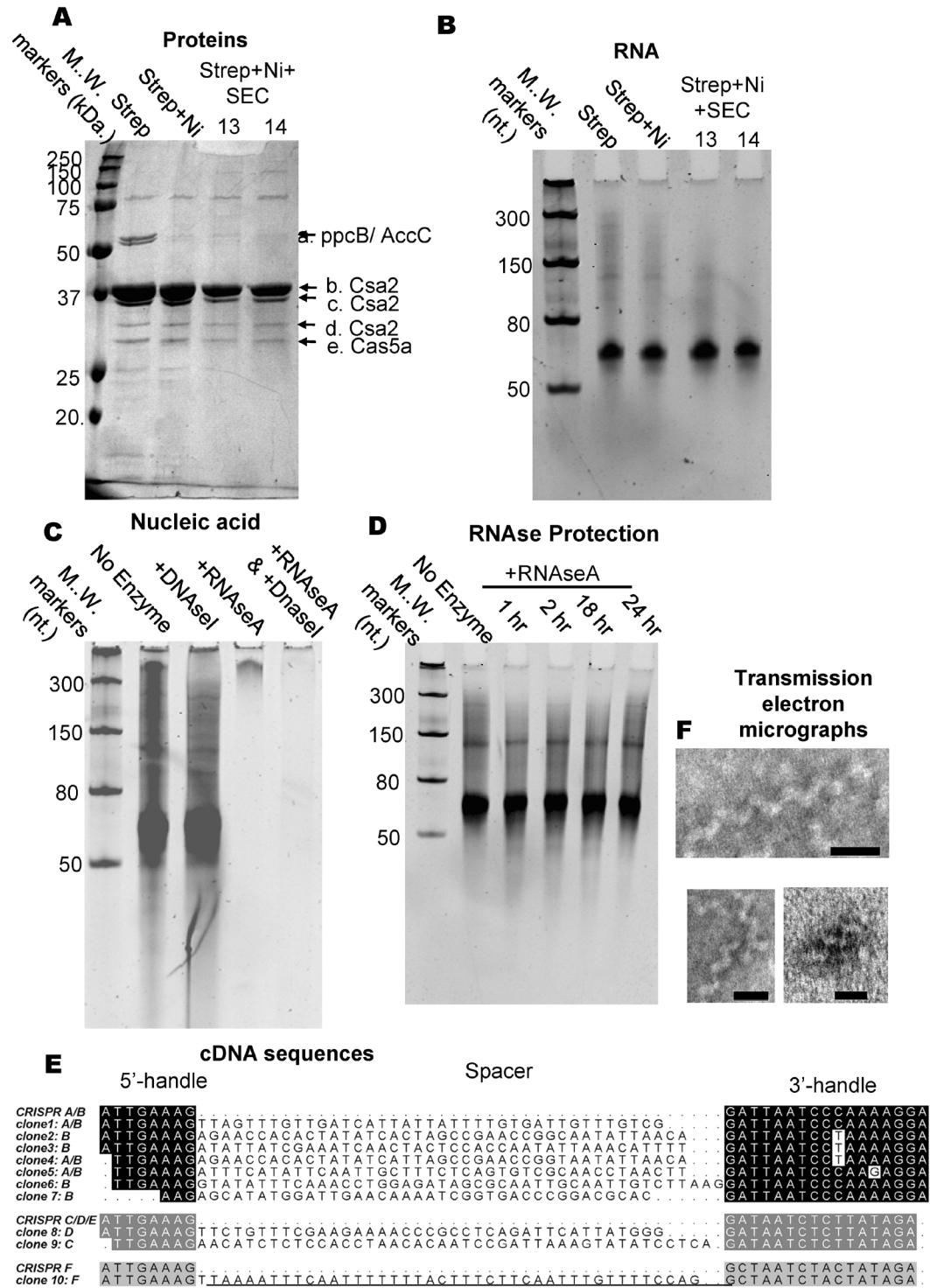


Figure 5-1: Isolation and characterization of aCASCADe from *S. solfataricus*.

Figure 5-1 (Continued): (A) Colloidal-Coomassie stained SDS-PAGE gel showing the copurification of Csa2 and Cas5a through StrepTactin (lane 2), Ni-NTA (lane 3) and size-exclusion chromatography (lanes 4 and 5) steps. (B) Sybr-Gold stained UREA-PAGE gel showing nucleic acids extracted from the proteins using basic Phenol-Chloroform after the StrepTactin step and subjected to no enzymatic treatment (lane 2), DNaseI (lane 3), RNaseA (lane 4) or both enzymes (Lane 5). Lanes 2 and 3 are overloaded to demonstrate the relative amounts of DNA and RNA in the complex. (C) Sybr-Gold stained UREA-PAGE gel showing that RNA co-purifies with aCASCADE through all three purification steps. (D) RNase-protection assays. There is no degradation of the bound RNA is observed after 24 hours of incubation with RNaseA. (E) The co-purifying RNA is CRISPR derived and comes from all three *S. solfataricus* CRISPR types. Alignment of non-redundant cDNA sequences associated with aCASCADE. The labels indicate which CRISPR the clone is derived from. Some clones could be from either of two or three CRISPRs. One spacer targets the SIRV genome and is underlined. (F) aCASCADE purified from *S. solfataricus* forms unusual helical filaments. Several images of the helical filaments observed by transmission electron microscopy. The black bars are 20 nm.

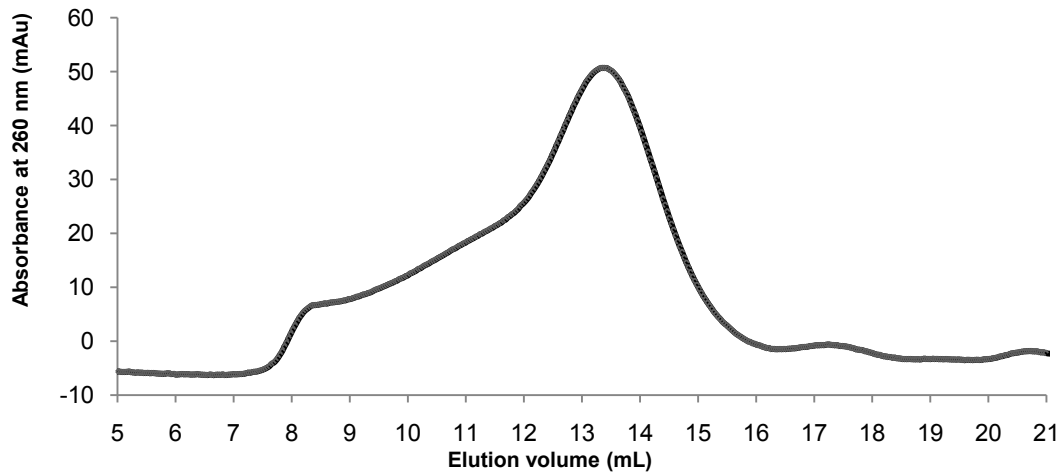


Figure 5-2: Behavior of the CASCADE-like complex purified from *S. solfataricus* on a superose-6 column. The major elution peak is centered at 13.6 mL, and corresponds to a molecular weight of ~350-500 kDa, and there is an additional “tail” of higher molecular weight material.

Table 5-1: Identification of high-abundance proteins identified using in-gel trypsin digestion followed by LC-MS/MS				
Band (Fig 5-1A)	Protein	Mascot Score	Unique peptides	Sequence Coverage
a.	Sso1399	106	2	7%
b.	AccC	289	8	16%
	PpcB	274	5	12%
c.	Sso1442	329	8	31%
d.	Sso1442	492	10	34%
e.	Sso1442	85	2	6%
f.	Sso1441	414	5	31%

Table 5-2: Proteins Co-purifying with Sso1442(Csa2) as identified by solution trypsin digestion followed by LC-MS/MS							
Protein	Family	Expt A		Expt B		Expt C (EtBr wash)	
		score	coverage	score	coverage	score	coverage
Sso1442	Csa2	24937	78%	33795	80%	32149	82%
Sso1441	Cas5	2040	53%	1770	57%	1726	63%
Sso1399	Csa2	999	66%	1189	40%	912	50%
Sso2466	Biotin carboxylase	593	33%	67	6%	377	59%
Sso1400	Cas5	396	41%	622	22%	327	22%
Sso1443	Csa5	130	25%	372	65%	377	59%
Sso0962	Alba	95	53%	173	72%	173	46%
Sso1437	Cas6	92	17%	132	17%	61	18%
Sso1401	Hypothetical	-	-	153	10%	84	8%

Recombinant *S. solfataricus* Cas6 Generates Fragments Identical to Those in aCASCADE

Cas6 from *P. furiosus* has been shown to generate processed crRNA species (60, 127) consistent with those identified in the *S. solfataricus* aCASCADE complex, but the equivalent activity has not been demonstrated in *S. solfataricus*. The annotated Cas6 orthologs in *S. solfataricus* (Sso1381, Sso1406, Sso1437 and Sso2004) share negligible sequence similarity with *P. furiosus* Cas6 and even structure based threading using the

Phyre server (128) does not find a match with *P. furiosus* Cas6. Accordingly, it was important to ascertain whether the putative *S. solfataricus* Cas6 orthologs cleaved crRNA *in vitro*. Our collaborators, Melina Kerou and Dr. Malcolm F. White cloned the *sso2004* gene and expressed it in *E. coli*, allowing purification of the recombinant protein. Recombinant *S. solfataricus* Cas6 cleaved an *in vitro* transcript comprising the first two repeat-spacer units of the *S. solfataricus* P2 CRISPR A locus, yielding a pattern consistent with cleavage at a single position within the repeat at the same position cleaved by *P. furiosus* Cas6 (Figure 5-3, A & C). This was confirmed by the cleavage pattern generated from an RNA oligonucleotide comprising a single 25 nt repeat sequence with a 15U 5' extension, which was cut at a single site (5'-AGGA/AUUG) (Figure 5-3, B & D), yielding the 8 nt 5' tag ("psi-tag") identified previously. Thus, the *S. solfataricus* Cas6 protein that purifies with the Csa2-Cas5a complex is capable of generating the crRNA products found in this complex. This is reminiscent of *E. coli* CASCADE where the crRNA cleaving subunit CasE (cse3) is a constituent of the complex (41).

Investigation of the Core CASCADE Subunits

For structural studies and activity assays, recombinant Csa2 (Sso1442) was expressed in *E. coli*, both alone and with Cas5a (Sso1441). For His-tagged Csa2 alone, size exclusion chromatography was unable to distinguish between monomer and dimer. Analytical ultracentrifugation (AUC) revealed a monomer-dimer equilibrium with a dissociation constant of 4.5 μ M, indicating that in the absence of other CASCADE components, recombinant Csa2 is predominantly monomeric at physiologically relevant concentrations. In contrast, Csa2 and Cas5a co-expressed in *E. coli*, formed a stable

complex that could be purified to homogeneity. The Coomassie staining suggested an excess of Csa2 over Cas5a (Figure 5-4A), and AUC showed a complex mixture of heterodimer and larger complexes, suggesting a dynamic mixture of subunits in the complex that were presumed to reflect different numbers of bound Csa2 subunits.

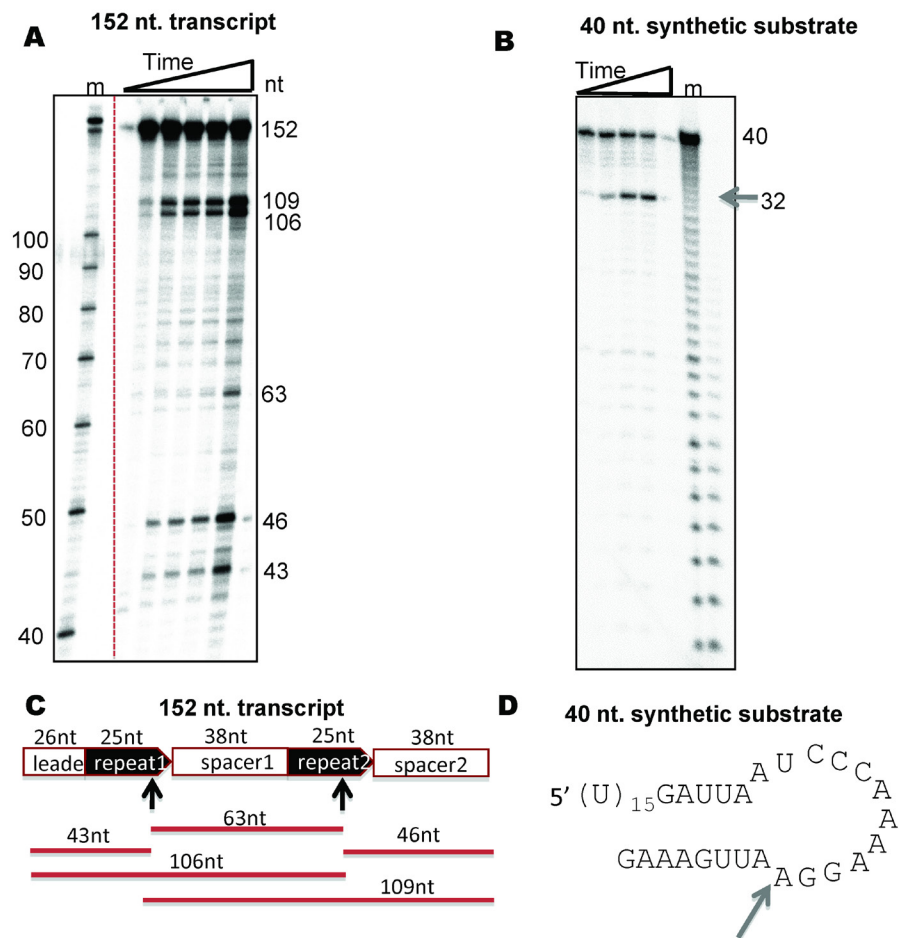


Figure 5-3: *S. solfataricus* Cas6 generates the crRNA (A) A two repeat-spacer unit CRISPR transcript was cleaved by Cas6 at a single site in each repeat, yielding fragments of 109 and 43 nt for cleavage at repeat 1, 106 and 46 nt for cleavage at repeat 2 and the 63 nt mature crRNA for cleavage at both repeats. (B) A synthetic RNA corresponding to a single CRISPR repeat with a 15U 5' extension is cleaved by Cas6 at a single site, generating an 8 nt repeat-derived 5' extension ("psi-tag"). (C) Schematic illustrating the 2-repeat transcript and the expected cleavage products. (D) Schematic illustrating the synthetic substrate. [Figure by Melina Kerou and Dr. Malcolm F. White]

The Csa2-Cas5a Complex Binds crRNA and Forms Ternary Complexes with Target DNA

E. coli CASCADE utilizes bound crRNA to target viral DNA, forming a ternary complex that is thought to result in cleavage of the DNA target by other CAS proteins, most likely Cas3 (41, 69). To determine whether Csa2-Cas5a had similar functionality, we carried out electrophoretic mobility shift assays (EMSA) to visualize crRNA and DNA binding using the sequence of CRISPR locus A, spacer 1. We first tested the ability of Csa2 and the Csa2-Cas5a complex to bind radiolabelled crRNA (Fig. 5-4B). Both bound the crRNA with roughly similar affinities, suggesting that Csa2 is the major RNA binding subunit of the complex. In the absence of crRNA-A1, the Csa2-Cas5a complex showed very little binding to a labeled target DNA species (Target-A1f) (Table 5-3, Fig. 5-4C, lanes 1-5). However, when the protein complex was preincubated with crRNA-A1, which can base pair with the central region of the tA1+n oligonucleotide, an RNA-DNA heteroduplex was formed that was gel-shifted efficiently by the Csa2-Cas5a complex (lanes 6-10). The Csa2-Cas5a-crRNA complex did not bind the reverse complementary DNA strand (Target-A1r), which cannot form a heteroduplex with crRNA-A1 (lanes 11-15). These data demonstrate that the Csa2-Cas5a complex has a crRNA-dependent DNA binding activity that is consistent with its presumed function in CRISPR-mediated antiviral defense, analogous to *E. coli* CASCADE. Recent data in *S. solfataricus* suggests that target DNAs are only cleaved if they include a “Protospacer Adjacent Motif” (PAM) sequence, typically CCN, at the 5' end of the protospacer (11, 48). This may be a mechanism to allow discrimination between foreign DNA and the chromosomal CRISPR loci, which lack PAM sequences and are therefore not targeted by the CRISPR system

(62). The target oligonucleotide A1 used for gel-shifting included a PAM sequence, however binding to an alternative oligonucleotide lacking a PAM sequence gave similar results (data not shown). This suggests that there is no discrimination based on PAM presence or absence in the minimal recombinant system tested here.

Table 5-3: Oligonucleotides used in this study (CRISPR repeat derived sequences are in bold and the PAM is underlined.)	
Oligonucleotide name	sequence
crRNA-A1	5'- AUUGAAAGGAACUAGCUUAUAGUUUAGAAGAAAACAAACAAUAAUGAU UAAUCCCAAAA
U15 CRISPR repeat	5'-UUUUUUUUUUUUUUU GAUUAUCCCAAAAGGAAUUGAAAG
Target-A1f	5'- TAATACGACTCACTATAGGGTATTATTTGTTTGTTCCTTCTAACTATAAGC TAGTTC TGG AGAGAAGGTG
Target-A1r	5'- CACCTTCTCT CC AGAACTAGCTTATAGTTTAGAAGAAAACAAACAAATAATA CCCTATAGTGAGTCGTATTA
crRNA transcript	5'- GGGAUAGGAAGUAUAAAAACACAACAGAU UAAUCCCAAAAGGAAUUGAA AGGAACUAGCUUAUAGUUUAGAAGAAAACAAACAAUAAUGAUUAAUCC CAAAAGGAAUUGAAAGAUUUUCAGCUGAAAAUUUGAAAUCUGUAGAUUU GGAUG

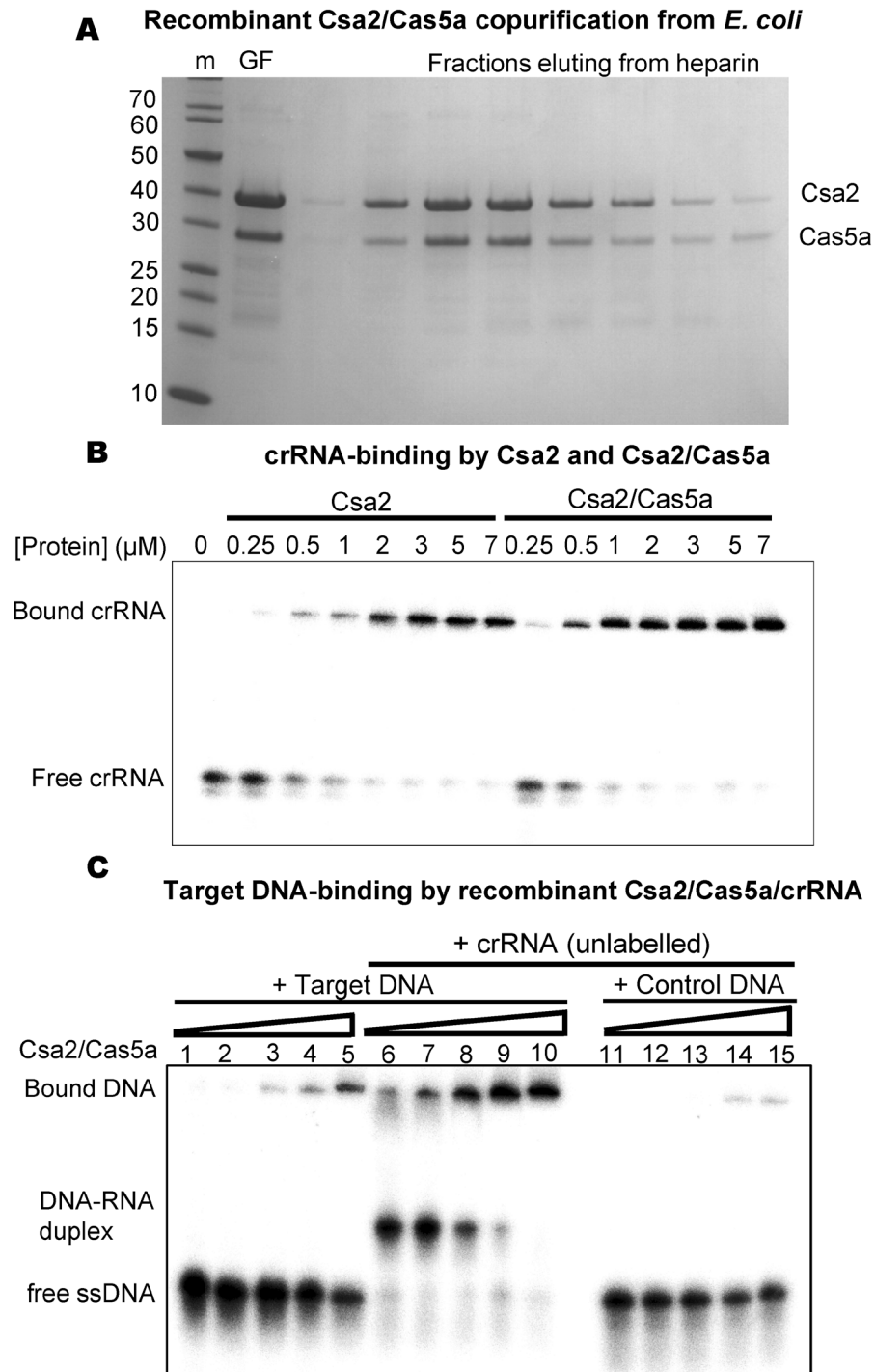


Figure 5-4: crRNA and target-DNA binding by a Csa2/Cas5a complex.

Figure 5-4 (Continued): (A) Recombinant Csa2 and Cas5a form a complex when co-expressed in *E. coli*. Coomassie-stained SDS-PAGE gel showing material purified using Ni-NTA affinity chromatography and gel-filtration chromatography (Lane 2) and further purified using a heparin columns (Lanes 3-10). (B) Electromobility shift assays (EMSA) showing the binding of Csa2 alone (lanes 2-8) and the Csa2/Cas5a complex (lanes 9-15). Lane 1 shows crRNA alone. Protein concentrations for both Csa2 and Csa2-Cas5a complex were 0.25, 0.5, 1, 2, 3, 5 and 7 μ M and the crRNA concentration was 100 nM. (C) Target-DNA binding by the recombinant complex. Increasing amounts of the Csa2-Cas5a complex was incubated with radiolabelled target ssDNA in the absence of crRNA (Lanes 1-5), in the presence of crRNA (Lanes 6-10) and in the presence of crRNA and non-target DNA (Lanes 11-15). [Figure by Melina Kerou and Dr. Malcolm F. White]

Structural Studies of A-CASCADE

In order to investigate the overall structure of the aCASCADE complex purified from *S. solfataricus*, the purified material was visualized using negative stain transmission electron microscopy (TEM). Following elution from the streptactin column, and at all subsequent stages of the purification including the peak fraction from the size exclusion column, we observed protein filaments with a width of approximately 6 nm. Interestingly, the filaments are present as extended right-handed helices of variable length, with an average helical width of 11.5 nm and a pitch of 14 nm (Fig. 5-1F). These particles are clearly larger than that suggested by size exclusion chromatography, probably due to their non-spherical nature.

Importantly, the observed right handed helical assemblies indicate that Csa2, the major protein in aCASCADE, is capable of forming oligomers with open, as opposed to closed symmetry. While these assemblies are significantly larger than those seen in *E. coli* CASCADE, we note that this unusual open oligomeric assembly is consistent with the relative abundance of the COG1857 protein in both *S. solfataricus* (Csa2) and *E. coli* (CasC) CASCADE. In addition, the observation that recombinant Csa2 expressed in *E.*

coli is predominantly monomeric in the absence of Cas5a and crRNA, suggests that Cas5a and crRNA may be responsible for nucleation and/or stabilization of this helical assembly. For these, and other reasons discussed in more detail below, we believe that many features of this extended complex are likely to be relevant to endogenous aCASCADE.

The Structure of Csa2

To better understand the role of the COG1857 family of proteins in CASCADE in general, and the role of Csa2 in aCASCADE in particular, the structure of recombinant Csa2 was solved using X-ray crystallography. Csa2 crystallized in space group $P2_12_12_1$ with four copies per asymmetric unit. However, the protein-protein contacts in the Csa2 crystal appear unlikely to recapitulate protein-protein interactions in aCASCADE. While there are substantial contacts at the A/B and C /D interfaces the surfaces are discontinuous and exhibit improper closed symmetry, inconsistent with the apparent open-symmetry of aCASCADE.

The structure of the Csa2 protomer reveals a 3-domain, crescent shaped structure that is 65 Å in length, tip to tip (Figure 5-5A). The central domain is comprised of a five-stranded anti-parallel β -sheet ($\beta 6$, $\beta 7$, $\beta 1$, $\beta 8$ and $\beta 9$), flanked by four α -helices. The first four strands of the central β -sheet along with helices $\alpha 2$ and $\alpha 8$ display the $\beta\alpha\beta\beta\alpha\beta$ topology of the RNA-recognition motif (RRM). We thus refer to these structural elements as the RRM-like sub-domain (purple, Figure 5-5A). In Csa2, the RRM is elaborated upon by a C-terminal addition comprised of residues 256-320 that begins with an extended 13-residue connection leading into helix $\alpha 9$. This is followed by a short connection to $\beta 9$ that adds as a fifth antiparallel strand to the β -sheet of the RRM, and $\alpha 10$, which sits

“underneath” the β -sheet (Figure 5-5A). Interestingly, Csa2 lacks the conserved sequence motifs in strands $\beta 1$ and $\beta 7$ (Figure 5-6B) that recognize single-stranded nucleic acid in strands $\beta 1$ and $\beta 3$ of the canonical RRM (104), although a solvent exposed aromatic residue is conserved C-terminal to $\beta 6$ (Tyr141). Consistent with the lack of a canonical ssRNA-binding motif, the extended $\beta 8$ - $\alpha 9$ loop and helix $\alpha 9$ lie on “top” of the antiparallel β -sheet, partially occluding the RNA binding face of the typical RRM fold. Thus, while the central domain of Csa2 may have evolved from the RRM, it is likely that RNA recognition by Csa2 will be different from other RRM containing proteins currently in the Protein Data Bank.

Two additional bipartite domains, one at each tip of the crescent shaped protein, are formed by four insertions into the RRM fold (Figure 5-5A, Figure 5-6). The “1-3” domain is found “above” the RRM domain as it is pictured in Figure 5-5A, and is formed from insertions 1 and 3. The first insertion (residues 27-46) includes $\alpha 1$, followed by a disordered loop that connects to the $\beta 2$ - $\beta 3$ hairpin, which extends to the upper tip of the “crescent”. The 1-3 domain is then completed by insertion 3 (residues 145-180), which contributes helix $\alpha 7$. However, this helix is ordered only in chains A and C, where it sits against the face of the $\beta 2$ - $\beta 3$ hairpin, while at least 7 residues in the $\alpha 7$ - $\beta 7$ loop are disordered in all chains.

The 2-4 domain, found “below” the RRM domain, is likewise composed of insertions two and four. The second insertion, residues 68-136, contributes a mixed α/β structure consisting of $\alpha 3$ - $\alpha 6$, which lie along one face of the short antiparallel $\beta 4$ - $\beta 5$ hairpin, while the opposite face of the $\beta 4$ - $\beta 5$ hairpin remains solvent exposed on the

concave face of the crescent. The fourth insertion, residues 192-216, contains another extended connection followed by the N-terminal half of $\alpha 8$. The $\alpha 8$ helix is kinked at residue 216 where it leaves the 2-4 domain and forms the second helix of the $\beta\alpha\beta\beta\alpha\beta$ RRM-like fold.

To identify functionally important residues, including potential sites for RNA recognition, or interactions with other aCASCADE subunits, we examined the locations of conserved surface features. Makarova *et. al* identified 3 conserved sequence motifs in CRISPR-associated COG1857-like proteins, specifically: 1) s-h-Asn, 2) Arg, and 3) (Phe/Pro/His/Gly)-Gly, where s and h indicate small and hydrophobic residues, respectively (59). In Csa2 these correspond to: 1) Ser14-Leu15-Asn16, 2) Glu58 and 3) Gly121-Gly122. All six residues are solvent exposed and found on the concave surface of the Csa2 crescent (Figure 5-5B). Among these, Asn16 is poorly ordered (Chain A) or disordered (Chains B, C and D).

Because few residues are generally conserved among all COG1857-like CRISPR-associated proteins, we also examined the location of surface residues conserved just among the Csa2 orthologues (Figures 5-5B and 5-7). Most of the strictly-conserved, solvent-exposed residues are found in two closely spaced clusters, which are coincident with the COG1857-like motifs discussed above, and are thus also found on the concave surface of the crescent. The first cluster is on the surface of the 1-3 domain near Asn16, while the second cluster lies at the interface between the RRM-like sub-domain and the 2-4 domain, near Glu58 and Gly121-Gly122. We thus refer to these clusters as the asparagine (Asn) and glycine (Gly) clusters, respectively (Figures 5-5B). These surfaces

are hydrophilic and somewhat basic (Figures. 5-5C). While either cluster might indicate a surface involved in subunit interactions, the identities of several conserved residues (His160, Arg162, His55 and Asn16) are more suggestive of nucleic-acid rather than protein recognition. In agreement with this, mutation of His160 to an alanine resulted in a significant reduction in the binding affinity of Csa2 for crRNA (Figures. 5-5E).

These conserved surface features are adjacent to three disordered loops, two of which contain additional residues that are strictly conserved among Csa2 orthologues; Gly22 and Asn23 are found in the α 1- β 2 loop, and Arg240 is present in the α 8- β 8 loop (Figures 5-5B). Further evidence for flexibility is also seen in the orientation of the β 2- β 3 hairpin of the 1-3 domain, which is shifted by 10.4 Å in chains B and D, relative to that in chains A and C (Figures 5-8). The conformational change is accompanied by the loss of additional ordered density in the α 7- β 7 loop in chains B and D, including strictly conserved His160 and Arg162 (Figures 5-8). The presence of conserved, flexible loops in Csa2 and the lowered affinity of the His160Ala variant for crRNA (above, Figure 5-5E) suggests these flexible loops may be involved in the recognition of crRNA or in crRNA directed DNA recognition. In addition, we also note the presence of a hydrophobic patch on the back side of the β 2- β 3 hairpin, at the interface of between the 1-3 domain and the RRM-like subdomain, a likely candidate for subunit interactions within the larger aCASCADE.

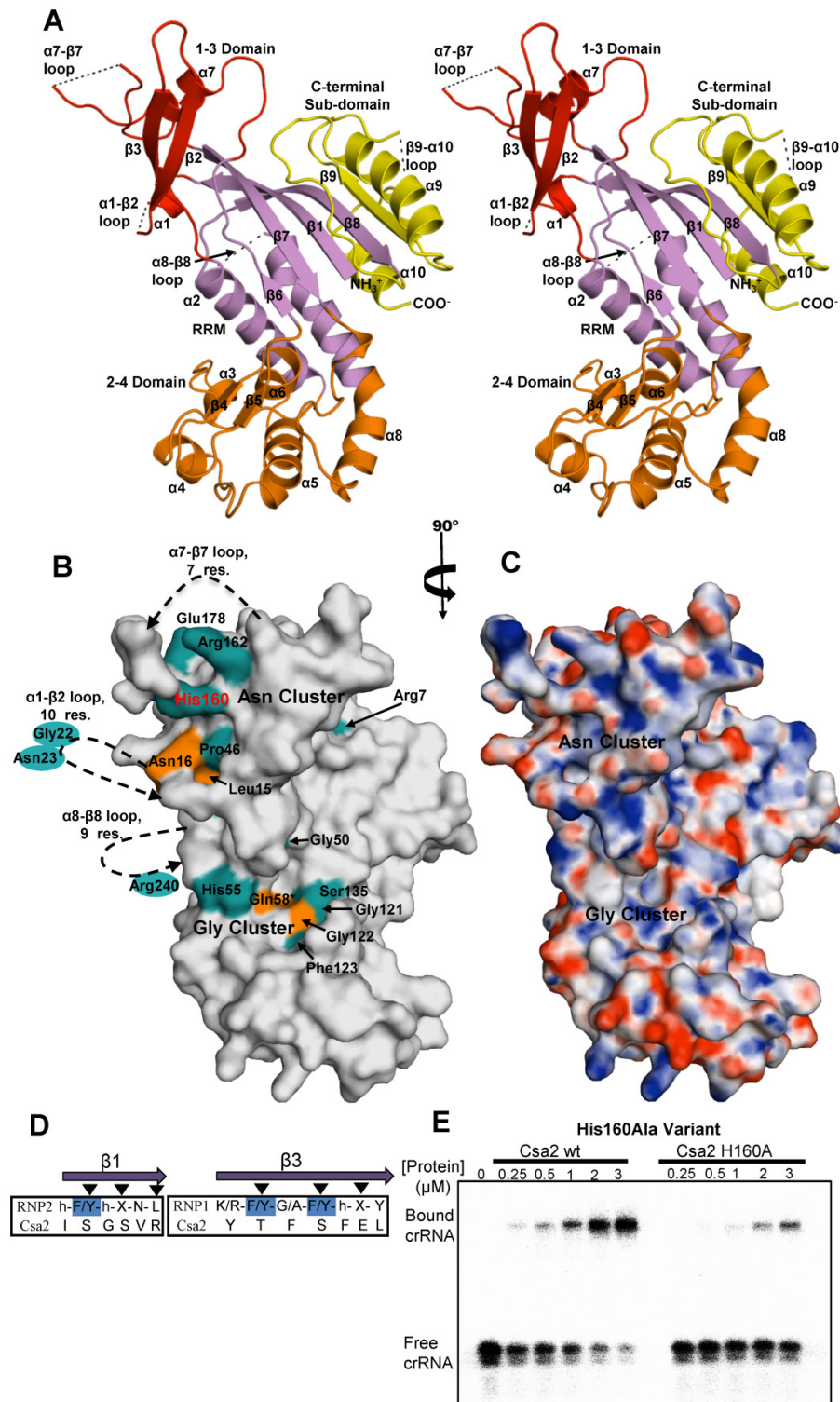


Figure 5-5: The structure of Csa2 reveals a novel domain architecture containing an RNA-recognition motif (chain A shown).

Figure 5-5 (Continued): (A) Stereo ribbon diagram of the Csa2 monomer. The RNA-recognition motif is colored violet, the 1-3 domain is red, the 2-4 domain is orange and the C-terminal subdomain is colored yellow. (B) Conserved residues cluster on the edge of the RRM and the 1-3 domain. Surface representation of Csa2 rotated 90° about the vertical axis relative to A. Sequence motifs which are conserved among the COG1857-like superfamily are colored orange and residues which are strictly conserved among Csa2 proteins are cyan. The approximate locations of the three disordered loops are indicated by dotted lines and the strictly conserved residues that are located in the disordered loops are indicated by cyan ovals. The two conserved residue clusters are indicated. (C) Electrostatic surface map of Csa2 calculated using APBS tools and shown in the same orientation as panel B. The color ramp of the surface is from -20 kT/e (red, acidic) to 20 kT/e (blue, basic). (D) Csa2 $\beta 1$ and $\beta 7$ lack the conserved ssRNA-binding motif found on $\beta 1$ and $\beta 3$ of the canonical RRM. These two canonical RNA-binding motifs are aligned with the Csa2 residues in the structurally equivalent positions. The solvent-exposed side-chains are indicated with black triangles. The aromatic residues which normally make base-specific contacts are highlighted in blue (104). (E) EMSA the reduced crRNA-binding activity of the His160Ala Csa2 variant.

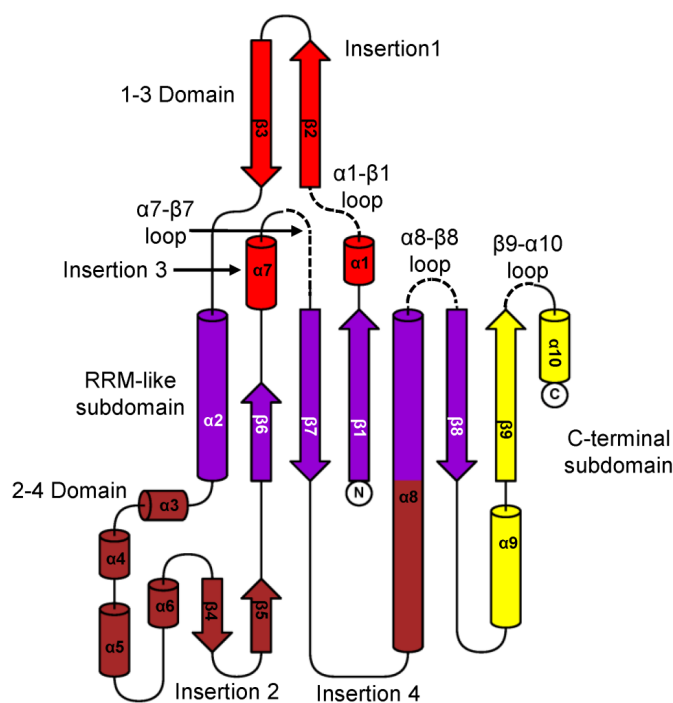


Figure 5-6: Connectivity of the Csa2 structural features. The RRM-like subdomain is colored violet, the 1-3 domain is red, the 2-4 domain is brown and the C-terminal subdomain is yellow. The disordered loops are depicted with dotted lines. Secondary structures are labelled as in Figure 5-5. The insertions into the core RRM-like fold are labelled and numbered sequentially.

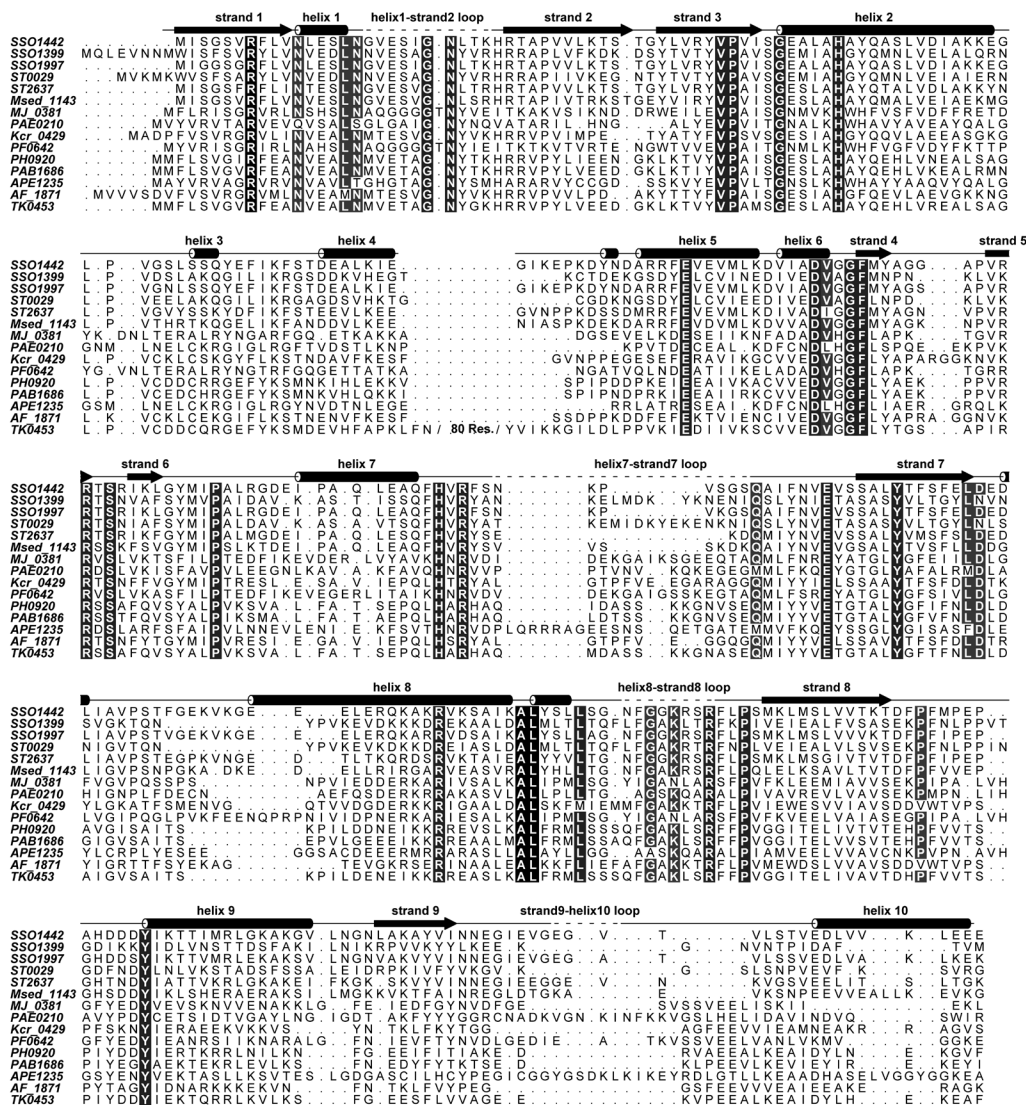


Figure 5-7: Multiple sequence alignment of Csa2 orthologs. Csa2 orthologs encoded in genomes with CRISPRs, Cas1-6 and Csa proteins were aligned using T-COFFEE. The secondary structural elements were assigned with DSSP using chain A of the Sso1442 structure and are the names correspond to figure 4. The disordered residues in chain A are depicted with dotted lines. TK0453 has an extra 80-residue insertion between helices four and five which is not shown in the alignment. Identified Csa2 orthologs are from *Sulfolobus solfataricus* (SSO), *Sulfolobus tokodaii* (ST), *Metallosphaera sedula* (Msed), *Methanocaldococcus jannaschii* (MJ), *Pyrobaculum aerophilum* (PAE), *Candidatus Korarchaeum cryptofilum* (Kcr), *Pyrococcus furiosus* (PF), *Pyrococcus horikoshii* (PH), *Pyrococcus abyssi* (PAB), *Aeropyrum pernix* (APE), *Archaeoglobus fulgidus* (AF), and *Thermococcus kodakarensis* (TK).

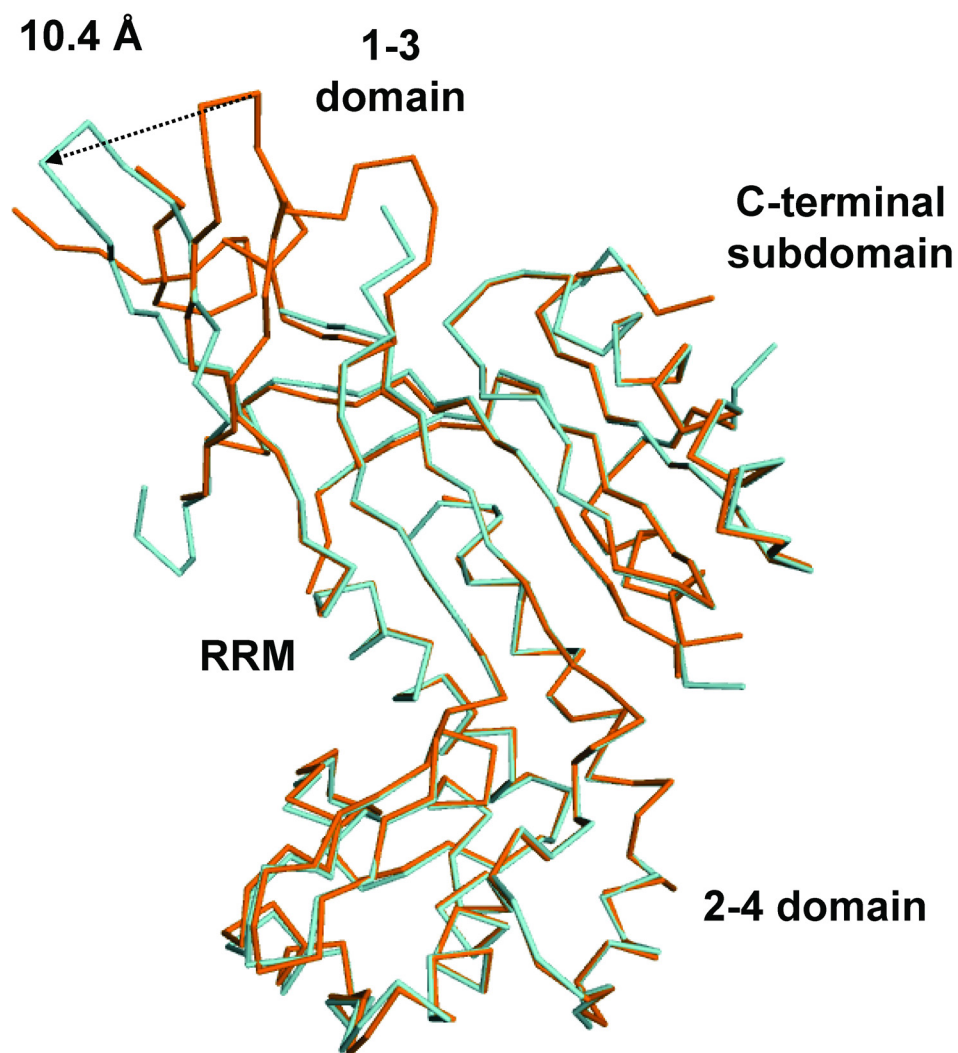


Figure 5-8: Two conformations of the 1-3 domain are present in the asymmetric unit. Csa2 Chain B (cyan) was SSM superposed onto chain A (orange). In chain B the end of the β -hairpin is shifted by 10.4 Å. This conformational change requires an additional 18-23 residues to become disordered.

DALI and SSM searches (84, 86) identified *Pyrococcus furiosus* Cas6 (60, 129) as the closest structural homolog to Csa2, but the similarity is limited to the RRM-like subdomain (Cas6 PDB ID 3I4H, 2.9 Å RMSD over 87 residues for chain A). Cas6, a

member of the RAMP superfamily, displays tandem ferredoxin- or RRM-like domains, with the N-terminal domain of Cas6 showing the greatest similarity to Csa2 (Figure 5-9). Cas6 is a metal-independent ribonuclease, in which His46, Tyr41 and Lys52 forming a putative catalytic triad (60, 127, 129). The SSM superposition places these residues in the vicinity of the Gly cluster on Csa2. While, the Gly cluster lacks a recognizable catalytic triad and is thus unlikely to represent a nuclease active site, the structural alignment with the Cas6 active site further suggests the Gly cluster may function in nucleic acid recognition. In contrast to the RRM domain, Dali queries of the protein databank with the 1-3 and 2-4 domains alone did not yield statistically significant matches, these domains thus appear to be unique to the Csa2 structure.

We next asked how the structure of the Csa2 protomer might relate to the helical assemblies observed by TEM, and began by considering the relative scale of these two structures. Interestingly, the width of the helical assemblies (6 nm) observed by TEM is approximately equal to the tip-to-tip diameter of the crescent shaped Csa2 protomer (65 Å). The right handed helical assembly can thus be crudely modeled by placing multiple copies of the Csa2 protomer in a right-handed helical arrangement such that the long axis of the Csa2 protomer runs perpendicular to the protein filament, and by requiring the pitch and width of the model helix to be consistent with the TEM images. In addition, in our models, we also chose to require the conserved crescent shaped face of Csa2, including His160 which appears to be involved in RNA recognition, to remain solvent exposed. We emphasize that there are additional ways these helices might be modeled, and that there is no reason to believe that the Csa2/Csa2 interface employed in our models corresponds to

that in the real complex. However, the exercise is valuable in that it suggests six to twelve protomers per turn of helix.

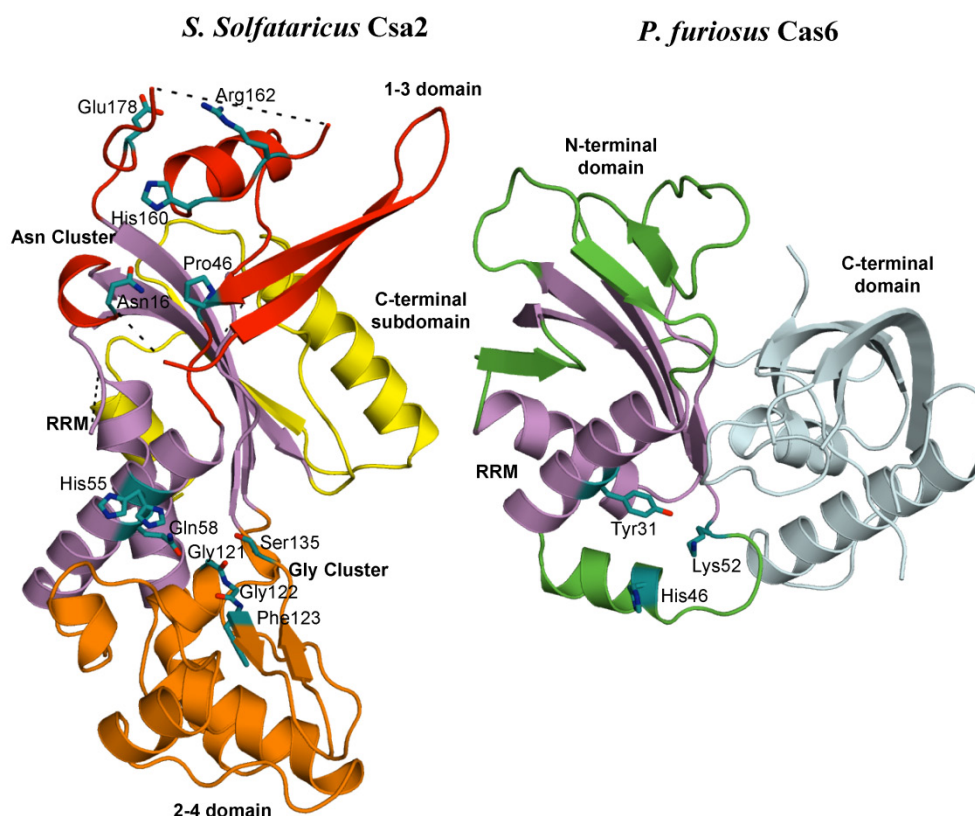


Figure 5-9: The structural similarity between Csa2 and Cas6 is limited to the RNA-recognition motif. Csa2 and Cas6 are shown in equivalent orientations based on the SSM Structural alignment. The RRM-like subdomains are colored violet in both structures. The Csa2 1-3 domain is colored red, the 2-4 domain is colored orange and the C-terminal subdomain is colored yellow. The portions of the Cas6 N-terminal domain that are not part of the RRM are green and the C-terminal domain is light cyan. The two conserved clusters on Csa2 are shown with “sticks” and colored dark cyan. The putative active site residues on Cas6 are also shown with “sticks” and colored dark cyan.

Discussion

Here we report the isolation and characterization of a complex that bears many of the hallmarks expected of an Archaeal CASCADE. Similar to *E. coli* CASCADE (41, 69), aCASCADE includes a CasC/COG1857-like protein (Csa2), a CasD/Cas5e ortholog (Cas5a), and this complex copurifies from *S. solfataricus* with processed crRNA. Further, the recombinant Csa2/Cas5a complex produced in *E. coli* specifically binds crRNA, and this complex, in turn, recognizes single stranded “target” DNA *in vitro*. Finally, the complex from *Sulfolobus* also copurifies with the more weakly interacting or lower abundance components, Csa5 and Cas6. In *Sulfolobus* aCASCADE, Cas6 appears to serve a function analogous to that of *E. coli* CasE/Cse3. Thus, there are clear orthologs in aCASCADE for each component of *E. coli* CASCADE except CasA and CasB, components that appear to be limited to the Ecoli CRISPR/Cas subtype. Importantly, the presence of core CASCADE components in *S. solfataricus* aCASCADE (Csa2/Cas5a), as well as Cas6, suggests that structural and functional studies of aCASCADE are relevant not only to the Aperi subtype CASCADE, but are also generally relevant to their orthologs in other CRISPR/Cas subtypes, especially the Tneap, Hmari and Mtube subtypes that also contain Cas6 (41, 58, 59).

S. solfataricus Cas6

Although annotated as a Cas6 ortholog, the crenarchaeal Cas6 proteins are highly diverged from the well characterized euryarchaeal Cas6 protein (60, 127). The two protein families share little sequence similarity beyond the glycine rich region that is the hallmark

of these proteins, and the proposed Tyr-His-Lys catalytic triad (60, 127) of *Pyrococcus* Cas6 does not appear to be conserved. It is therefore significant to see that *S. solfataricus* Cas6 functions like its *Pyrococcus* counterpart *in vitro*, processing the pre-CRISPR RNA transcript to generate crRNA, and that like *Pyrococcus* Cas6, Csy4 and CasE(Cse3) (41, 60, 62), *S. solfataricus* Cas6 generates crRNA with an 8 base 5' handle that contains the conserved GAAA(C/G) motif identified by Kunin et al. (130). Our data also suggest that Cas6 shows only moderate affinity for the aCASCADE complex. Like many organisms utilizing Cas6, *S. solfataricus* contains both the CMR and CASCADE systems. Indeed, it appears that Cas6 is associated with all CRISPR subtypes (Tneap, Hmari, Aperi, Mtube) predicted to contain unstructured CRISPR repeats and both the CMR and CASCADE systems (58, 130). The potentially loose association of Cas6 in *S. solfataricus* aCASCADE may allow it to function in the initial processing of the CRISPR transcript for both CRISPR systems.

Structural Models for A-CASCADE

The over expression of Csa2 in *S. solfataricus* results in the production of extended right-handed helical assemblies of variable length. The ability of the extended Csa2 assembly to bind crRNA and to protect the crRNA from RNase digestion suggests that the crRNA is bound by protein along its entire length. That the assembly copurifies with Cas5a, Csa5 and Cas6, in addition to the crRNA, also suggests that many aspects of the assembly are physiologically relevant. However, the extended helical filaments observed in preparations from *S. solfataricus* are longer than needed to accommodate a single crRNA; one turn of the Csa2 helix should be more than sufficient to accomplish

this. Thus, if these helices are physiologically relevant, they are likely to harbor multiple crRNAs and could potentially be used in succession to screen potential target DNA for a match to the collection of bound crRNA.

However, the open symmetry of the Csa2 assembly coupled with high concentrations of Csa2 from over-expression in *S. solfataricus* and substoichiometric amounts of endogenous Cas5a, Csa5 or Cas6, might allow the assembly to grow to physiologically irrelevant lengths, particularly in the presence of stabilizing crRNA. Thus, we propose a shorter model, in which native aCASCADE includes a single crRNA and a limited number of Csa2 subunits, resulting in an arch shaped structure corresponding to less than one turn of helix. Indeed, this second model is consistent with the most recent model for *E. coli* CASCADE, which binds a single crRNA and is observed as a smaller arch shaped particle with no discernible symmetry (69); the arch-shaped backbone of *E. coli* CASCADE is nicely explained by the presence of a half turn or more of CasC helix (Figure 5-10).

In either model, the major function of Csa2 appears to be the construction of an extended assembly that functions to support the crRNA spacer sequence along its entire length, protecting it from RNase digestion. At the same time, the bases must be exposed and available for interaction with target DNA, such that aCASCADE effectively templates or presents the spacer sequence for DNA recognition. This suggests that the conserved surface features on Csa2 will tightly interact with RNA in a sequence independent manner. They might also serve to stabilize a hybrid RNA/target-DNA complex, and

perhaps to even destabilize dsDNA, allowing DNA within the cell to be surveyed for complementarity to the CRISPR spacer.

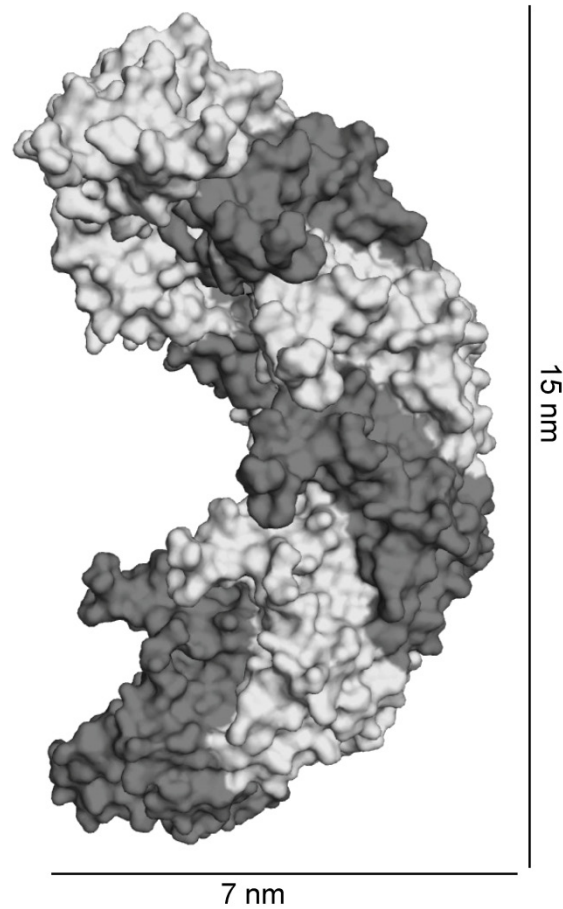


Figure 5-10: A hypothetical Csa2 hexamer approximates the core of *E. coli* CASCADE. The size and shape of a modeled Csa2 hexamer is similar to that of the hexameric casC core of *E. coli* CASCADE (69).

How is crRNA specifically recruited to aCASCADE? We envision a cooperative CASCADE assembly process that is dependent on the presence of crRNA, as we do not see extended helices of recombinant Csa2 or Csa2/Cas5a in the absence of crRNA. In

addition, while the Csa2 backbone of aCASCADE is expected to bind the variable CRISPR spacer in a sequence-independent manner, the complex clearly distinguishes crRNA from other cellular RNAs, most likely through sequence-specific interactions with the 5' and 3' handles. For these reasons, it is attractive to consider roles for the additional aCASCADE components (Cas5a, Csa5, Cas6) in specific crRNA recognition. Such interactions might also serve to nucleate growth of the Csa2 helix (crRNA induced oligomerization), and/or terminate growth of the helix, governing the length of the Csa2 backbone. We thus propose the model for aCASCADE presented in Figure 5-11. The structural core of CASCADE is modeled as a partial turn of Csa2 helix, with crRNA running along the length of the Csa2 assembly, and Cas5a, Csa5 and/or Cas6 at the 5'- and 3'-ends, where they may serve to initiate and/or terminate growth of the complex.

Once assembled, CASCADE must probe the DNA within the cell for sequences complimentary to the bound crRNA spacers. While we cannot definitively predict which surface- CASCADE utilizes for this process, we note that the concave surface formed by a partial turn of the Csa2 helix is large enough to accommodate or wrap around dsDNA. Further, several of our TEM-based models indicate that the partial Csa2 helix can be docked to dsDNA in a coaxial arrangement, that is, with the Csa2 helical axis coincident with that of the DNA double helix. In such an arrangement, the Csa2 protomers are positioned along the dsDNA, each equidistant from the DNA, potentially facilitating DNA surveillance. Thus, tentatively place crRNA in the aCASCADE-model such that it runs along the concave surface of the Csa2 arch where it would be available to probe dsDNA. However, while larger complexes often assemble with symmetry axes aligned, we must

emphasize that there is, as yet, no direct experimental evidence to support the position of the crRNA and DNA within the proposed model.

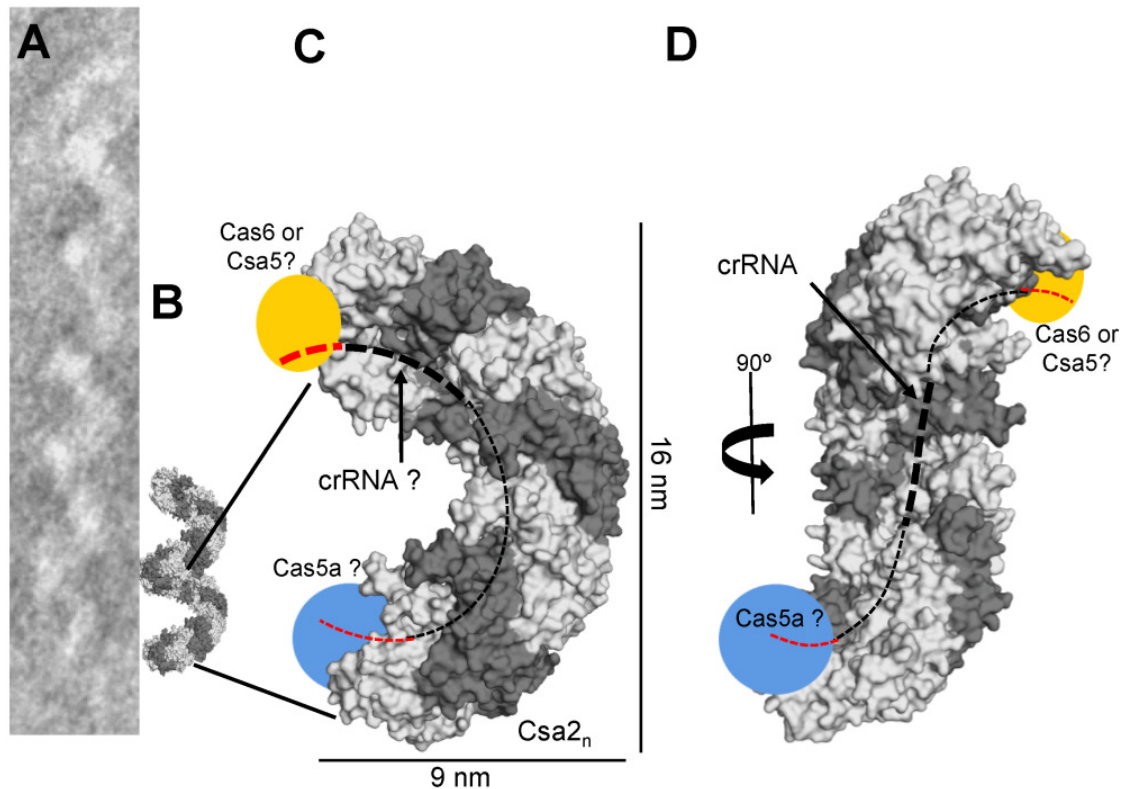


Figure 5-11: Preliminary structural models for aCASCADe. (A) TEM image of helical assembly. (B) Model of a Csa2 helical assembly with 10 Csa2 protomers per turn of helix. The Csa2 subunits are alternately colored dark and light grey. The Gly clusters are colored cyan and the Asn clusters are colored orange. (C) Proposed model for aCASCADe. The structural core of the model is formed from several copies of Csa2 in a helix-like assembly (colored as in B). At one end of the assembly is a single copy of Cas5a. The other end may have a protein, which caps the growth of the Csa2 assembly. The crRNA is tentatively placed along the inner surface of the Csa2 helical assembly and is indicated by a dotted line with the repeat portions shown in red and spacer shown in black. (D) The model in C is rotated 90° about the vertical axis to show the conserved residues.

Upon reflection, additional advantages of the unusual open symmetry of the Csa2 oligomer become apparent, particularly in light of the variable spacer sequences seen both within and between different organisms utilizing CRISPR-Cas. *E. coli* K12 shows a

limited range of spacer lengths, generally 32-33 bases (Figure 5-12). Assuming the spacer sequence is largely bound by the 6 CasC subunits in the backbone of *E. coli* CASCADE, each CasC subunit would accommodate about 5-6 bases of the crRNA spacer. Spacer lengths are generally longer in Archaea, however, and *S. solfataricus* spacer lengths vary between 34 and 44 bases, with a 39 base spacer the most common (Figure 5-12). While there are other ways the helix parameters might be adjusted, the additional length of these archaeal spacers might be accommodated by extending the Csa2 helix from 6 to 7, or even 8 subunits, depending on the size of the particular spacer. Similarly, the possibility that crRNA length may define the number of CasC subunits has also been considered for *E. coli* CASCADE (69). Thus, a role for the crRNA spacer in governing the number of COG1857-like subunits in the CASCADE “backbone” may be a general feature of CASCADE architecture, allowing crRNA of variable lengths to be accommodated, both within and across species. The architecture of CASCADE might also impose constraints on the size of a functional CRISPR spacer. Specifically, the majority (greater than 99.6%) of spacers in the CRISPR database (45) are 50 nt. or shorter (Figure 5-12). This may, in part, reflect the need for growth of the Csa2 oligomer, and COG1857-like oligomers in general, to terminate before completing a full turn of helix, allowing the inner surface of the helical Csa2 backbone to remain accessible to DNA.

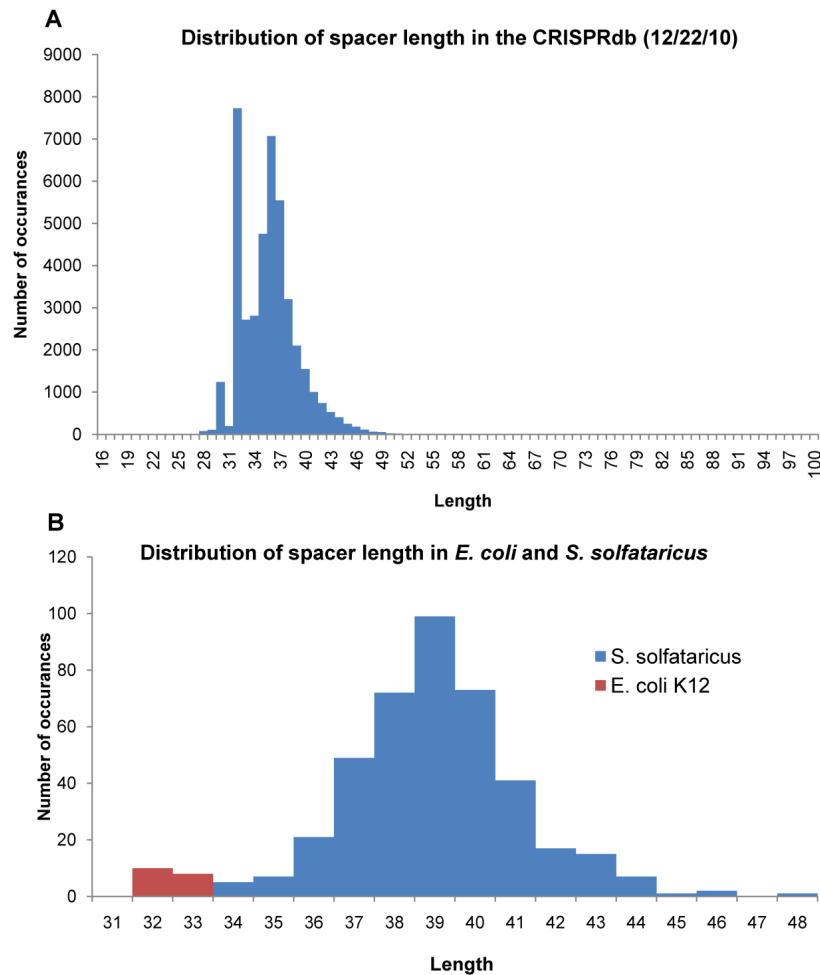


Figure 5-12: Distribution of spacer sizes among CRISPRs. Histogram showing **(A)** the distribution of lengths for all spacers in the CRISPR database (45) on 12/22/10 and **(B)** the spacer lengths in *S. solfataricus* and *E. coli*. The open symmetry of CASCADE may allow it to accommodate crRNA of variable size.

CRISPR-Mediated Viral Defense in the Archaea

The data presented here for *S. solfataricus*, coupled with previous work on the CMR complex in *P. furiosus* (63), allows the construction of an emerging model for the CRISPR system in many Archaea. CRISPR transcripts are processed by Cas6 to produce crRNA and incorporated into aCASCADE, which has a stable core composed of Csa2-

Cas5, with the former present at a higher copy number like its *E. coli* homolog CasC/cse4. Additional subunits may include Csa5, which appears unique to aCASCADe, and Cas6. This complex can form a ternary complex with target DNA that is presumably cleaved by the Cas3 helicase-HD nuclease protein(s), again in line with the situation in *E. coli*. Alternatively, the crRNA can be further processed by an unknown nuclease to remove the 3' repeat-derived RNA, generating smaller RNAs (psiRNAs) that are loaded into the CMR complex and used to target viral RNA. The two systems, CASCADe and CMR, may work in parallel in a two-pronged approach to maximize the utility of the CRISPR system (Figure 5-13).

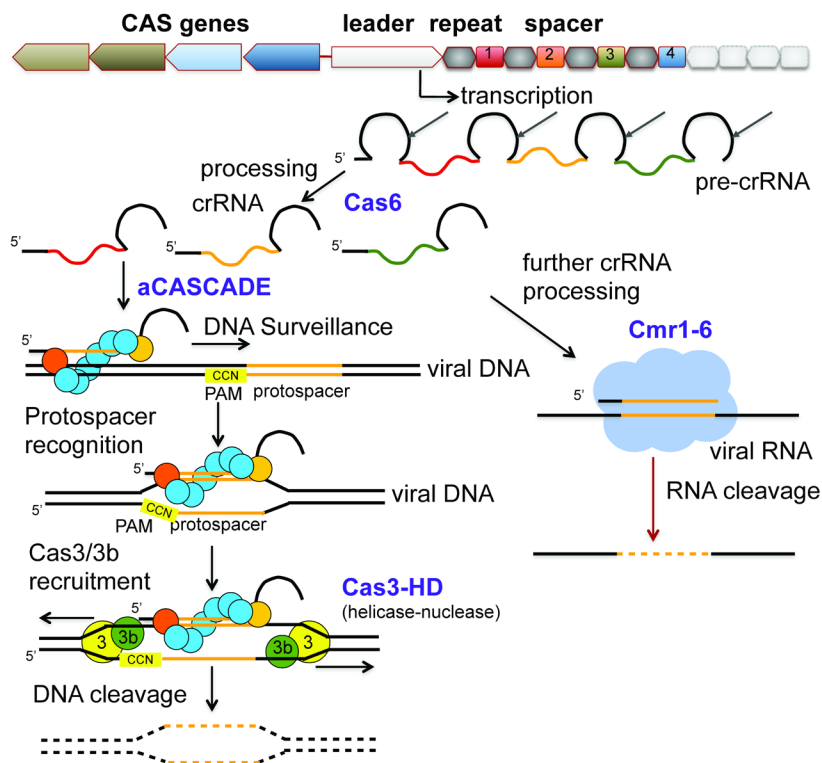


Figure 5-13: The two arms of CRISPR-mediated viral defense in *S. solfataricus* and other archaea. CASCADe and Cas3 target invading DNA and Cmr1-6 target invading RNA.

CONCLUDING REMARKS AND FUTURE DIRECTIONS

Virus-host interactions are one of the most important drivers of biological diversity and evolution. Despite the importance of viruses, there is a lack of extensive research into virus-host interactions in the domain Archaea, one of the three domains of life. Initial studies of archaeal viruses revealed tremendous molecular and morphological diversity, particularly among those infecting the crenarchaea (20). The work presented here advances detailed molecular insight into a model crenarchaeal virus-host system. Additionally, the work on the CRISPR-Cas system provides new insight into CRISPR-Cas systems in general.

This work initially sought to determine functions for proteins encoded by the crenarchaeal virus STIV, specifically A81. While the structure does not yield a clear prediction of function it does, give us a basis from which to form hypotheses and design further experiments. It is possible that A81 encircles single-stranded nucleic acid. Band-shift assays of A81 using various ss-nucleic acid substrates are clearly warranted by this prediction. Development of a functional assay for A81 would then allow the hypothesis that A81 encircles single-stranded nucleic acid to be tested using site-directed mutagenesis to disrupt the putative nucleic acid binding sites and oligomerization interfaces. A potential issue with any nucleic-acid-binding experiment is that the A81 octameric ring would presumably have to assemble and disassemble on a single-stranded nucleic acid and this may make A81 unable to bind substrate in the absence of other factors.

Identification of binding partners for A81 would also provide insight into its specific molecular function and the role of that functionality. Along these lines, polyclonal

antibodies against A81 have been produced. These would allow immunoprecipitation of A81 from infected cells and co-precipitating proteins to be identified by trypsin-digestion followed by LC-MS/MS. Alternatively, tagged A81 could be expressed in *S. solfataricus* and co-purifying binding partners could be identified. In this case, it would be desirable to utilize virally-infected cells to provide any potential virus-encoded binding partners and to increase the probability that any host binding partners would be expressed. Along these lines, development of a uracil-auxotrophic version of the STIV-susceptible P2-2-12 strain would be desirable. Future co-purification experiments should also be tried with a chemical crosslinking step to help identify weaker binding partners (131).

The structure of Csa3 suggests that the CRISPR-Cas system or of its some components are regulated by an unknown small-molecule regulator. The clear next steps in research on Csa3 are identifying the genes that are regulated by Csa3 and the molecule that binds to the conserved pocket on the N-terminal domain. We are able to express recombinant tagged Csa3 in *S. solfataricus*. Thus, the first attempt to identify the DNA sequence Csa3 binds should be using ChIp-on-Chip or ChIp-Seq. Alternatively, the sequence could be identified using SELEX. However, if the regulator must be bound for Csa3 to bind DNA, SELEX would be unlikely to work. Once a DNA-binding site is identified, the effect of Csa3 on transcription, specifically whether Csa3 is an activator or repressor, could be determined by overexpressing Csa3 in *S. solfataricus* and monitoring the transcription of the downstream genes by RT-qPCR by using a reporter-gene system.

The structure of Csa3 suggests a novel small-molecule regulator. Therefore, the identity of that regulator is also of great interest. The regulator could be identified by

incubating Csa3 with a small-molecule fraction of *S. solfataricus* lysate, separating the unbound molecules, and identifying the bound molecules by LC-MS. Navid Movahed in Dr. Brian Bothner's research group has initiated preliminary experiments to do this and identified several species, which apparently bind the Csa3. None of the masses correspond to metabolites in existing databases, making identification more difficult. Additionally, these trials were done with the mass spec. instrument in positive ionization mode, however the strictly conserved Arg residue in the conserved pocket suggests that the binding partner is negatively charged and the experiment should be repeated in negative ionization mode. If an unknown species is consistently identified as binding to Csa3 and enough of it can be isolated, then the structure of the binding partner could be solved using NMR. If the molecule turns out to be commercially available, the binding to Csa3 could be confirmed using isothermal titration calorimetry (ITC). The binding of the ligand to the conserved pocket could be then confirmed by solving the ligand-bound structure or, failing at that, by site-directed mutagenesis.

If a DNA-binding site and a small-molecule regulator are identified, the effect of Csa3 with and without the regulator could be assayed using the *Sulfolobus in vitro* transcription system that has recently been developed (132). If this assay could be developed it would provide a productive system to explore the effects of various Csa3 and promoter mutations on transcription from Csa3-regulated promoters.

The isolation and characterization of the *S. solfataricus* CASCADE reported herein, the first CASCADE isolated from an archaeon provides many potential directions for future research. Because the core of *S. solfataricus* CASCADE appears to consist of

only a COG1857 protein and a Cas5 protein, the most conserved CASCADE components, *S. solfataricus* CASCADE provides a model that will be applicable to many CRISPR-Cas subtypes. Many questions about *S. solfataricus* CASCADE and CASCADE in general remain unanswered. These include, what is the true size of native *S. solfataricus* CASCADE? Given that the majority of a crRNA is variable in sequence, how does CASCADE specifically recognize crRNA and not all cellular RNA? How does CASCADE scan DNA, a process which presumably requires a partial unwinding of the helix to probe for complementarity to the crRNA spacer? Upon recognizing a protospacer, by what mechanism does CASCADE unwind DNA, independent of ATP or divalent cations and recruit Cas3? If the COG1857 backbone does have an open symmetry, how is its length regulated? How, mechanistically does CASCADE recognize the lack of a protospacer adjacent motif adjacent to the CRISPR-locus in the chromosome (133) and avoid targeting its own genome?

Another outstanding question is the size of the native *S. solfataricus* CASCADE. One method which may answer is to fractionate cleared lysate from *S. solfataricus* on a size-exclusion column and probe the fractions for Csa2 using Western blotting. It is possible, however, that the complex could run anomalously, so the native complex may need to be isolated using an antibody affinity column with anti-Csa2 antibodies or by developing a multi-step chromatography method and following the Csa2 by Western blotting. The size of the isolated native complex could then be determined by analytical ultracentrifugation or possibly non-covalent mass spec. Although, the variable sequence

of the crRNA may introduce too much mass heterogeneity for non-covalent mass spec analysis.

A high-resolution structure of a CASCADE would provide significant insight into the remaining questions. In this area, the isolation of a truly native *S. solfataricus* CASCADE would be of high value. *E. coli* CASCADE has so far proven to not be amenable to crystallographic studies (69). It is possible that the *S. solfataricus* CASCADE will be more suitable. Furthermore, because most of the protein in *S. solfataricus* CASCADE is Csa2 the crystal structure of Csa2 provides a convenient molecular-replacement model for phase determination. There are several impediments that must be overcome before *S. solfataricus* CASCADE is suitable for crystallization trials. The complexes that have been isolated as part of these studies are clearly heterogeneous and are thus, unsuitable for crystallographic studies. A native, homogenous, complex must be isolated for crystallographic studies. It is possible that a purification protocol could be developed for endogenous, non-overexpressed aCASCADE by fractionating *S. solfataricus* P2 lysate using several steps and following CASCADE by western blotting. This approach is likely to be tedious and would require very large amount of cellular material. A more promising approach would be to reconstitute a native-like CASCADE from protein components expressed in *E. coli* and synthetic RNA. This would approach would have the added advantage over the truly native complex of having every complex bound to the same RNA sequence. Along these lines, recombinant Csa2, Cas5a, Csa5 and Cas6 have been expressed and purified by either us or our collaborators. A reconstituted complex would also be useful for the development of functional assays to explore the

mechanisms of the many activities of CASCADE including RNA-binding, DNA-unwinding, DNA-probing, target-DNA recognition, DNA-unwinding and Cas3 recruitment.

The understanding of CRISPR-Cas has grown dramatically in only the past few years. However, in spite of all the new research, we still have almost no understanding of the adaptation phase of CRISPR-Cas. Key questions include, how does the system initially recognize viral DNA, before there is a spacer in the CRISPR? Short conserved motifs called protospacer adjacent motifs (PAMs) have been determined to be important for this process (134). PAMs are only 2-3 nucleotides long and thus, in a best-case scenario, would occur by chance once every 64 nucleotides ($4^3=64$) in random DNA sequence. Thus, the PAM requirement cannot account for the fact that the chromosome is not targeted by CRISPR-Cas for new spacer generation. And how is the DNA processed into spacers and incorporated into CRISPRs? The key proteins predicted to be involved in these processes are Cas1 Cas2 and the RecB-like nuclease (Cas4 or Csa1) but the reported activities of Cas1 and Cas2 alone cannot account for the incorporation of new spacers.

As of yet, there is no published structural or functional data on the predicted CRISPR-associated RecB-like nuclease (Cas4 or Csa1). Thus, the Csa1 protein, Sso1391, expressed as part of these studies provides an attractive avenue of future research. Activity assays should be conducted with Csa1 to determine if it actually possesses RecB-like activity and additional efforts should be made to crystallize Csa1. Csa1 purified from *E. coli* has a brown color suggesting a bound co-factor. This co-factor should also be identified by ICP-MS if it is a metal or LC-MS if it is an organic molecule. It is possible

that incomplete occupancy by this co-factor is introducing heterogeneity and interfering with crystallization.

Because the activities of Cas1 and Cas2 alone cannot account for the incorporation of new spacers, it is likely that they interact with other proteins, possibly even “non-CRISPR-associated” proteins. Biology has many examples of proteins which normally have one function being recruited for something different. One possibility is that Cas1 and Cas2 recruit the cell’s general DNA-repair machinery for the integration of new spacers. Indeed, many CRISPR-Cas systems lack the RecB-like nuclease (58, 59), the simplest explanation for this is that these systems instead utilize their RecB. In fact, the Cas1 from *E. coli*, which lacks a CRISPR-associated RecB-like nuclease, was very recently shown to interact with the RecB, RecC and RuvB normally involved in DNA-repair (64). Research into the adaptation phase of CRISPR-Cas has been hampered by the inability of researchers focused on *E. coli* and *P. aeruginosa* to develop a reliable biological protocol to assay spacer-incorporation phenotypes (64), which leads to the inability to investigate the importance of given genes for spacer incorporation. Thus, additional model systems, including *S. solfataricus*, should be investigated. As part of these studies we have developed protocols and techniques for the identification of binding partners for *S. solfataricus* proteins. Thus, it should be determined whether this interaction with general DNA-repair machinery extends to the ApeII subtype by expressing tagged Cas1, Cas2, Cas4 and Csa1 in *S. solfataricus* and look for co-purifying binding partners. In any case, CRISPR-Cas remains an exciting area of research with several key discoveries yet to be made.

PRESENTATION OF STRUCTURAL DATA USING THE 3-D PDF FORMAT

Recent advances in the Adobe Acrobat software have allowed the inclusion of interactive, 3-dimensional (3D) content directly in publications (135-137). Readers can interact with the 3D content by rotating, panning zooming and toggling layers on and off. Preset views can be designated, allowing the author to direct the reader to specific aspects of the model. The content can also include multiple embedded layers which can be toggled on and off to display various aspects of the model or data. The paper based on chapter 4 of this thesis (105) includes an interactive 3-dimensional figure. This short chapter describes how such figures can be generated. It is meant to be useful as a how-to manual for other researchers. Generation of 3-D PDF content requires Pymol and Adobe 9 Pro Extended. To view 3D interactive content readers must have the free Adobe Reader version 9 or later. The 3D-PDF shown in figure 7-1 can be generated by downloading the coordinates of Csa2 from the protein databank (PDB ID: 3PS0), deleting chains B, C, D and the waters, and following the directions in this chapter.

Generating Input Files with Pymol

A Pymol script must be written which loads the coordinate file(s) to be presented and assigned “selections” to each aspect of the model which will be displayed as a separate layer or would be colored separately in the final figure (For the Csa2 example see Appendix 1). The colors assigned to various parts of the model in Pymol will not be preserved when converted to the 3-D pdf format with the exception of atom colors in Pymol’s “sticks” rendering. Therefore each portion of the model that will be colored

differently or displayed separately must be saved as a separate object. Once loaded the model and selections are loaded, the user must use the dropdown menu to hide everything. The user then “shows” each domain or set of residues individually using the dropdown menu. The user then saves each image in the VRML2 format (file extension: “.wrl”). In the Csa2 example, the user must use the dropdown menu to hide everything and then show cartoon for the selection “1-3_domain” and show dashes for the distance selections “d1” and “d2” which are set up to draw dotted lines across the two gaps in the 1-3 domain. The user then saves the resulting image in the VRML2 format and repeats the process with the 2-4 domain, the RRM-like subdomain and the C-terminal subdomain. To generate the surface rendering colored by conservation, the user must first show the “surface” of the entire chain then hide conserved residues using the dropdown menu. The non-conserved surface is then saved as a VRLM2 file. Finally the surface of the conserved residues only are “shown” and saved in the VRLM2 file.

Setting up 3D Content with Adobe 3D Reviewer

The interactive content must then be setup using Adobe 3D Reviewer, a component of the Adobe 9 Pro Extended package. Adobe 3D Reviewer must be launched all the “.wrl” file generated in Pymol must opened simultaneously using “cntrl” to select multiple files. Each portion of the model is to be re-colored as desired by right-clicking on it, either in the model or in the list of layers at the left and selecting “properties”. The best way to select colors is, in the color menu to choose “additional options.” The best way to reproduce the Pymol colors is to enter the RGB values that Pymol defines each

color as (http://www.pymolwiki.org/index.php/Color_Values). In the Csa2 (Fig. 7-1) example, the RRM is colored violet, the C-terminal subdomain is yellow, the 1-3 domain is red and the 2-4 domain is orange. In the “surface” layers the non-conserved features are gray, the residues conserved among COG1857 are orange and the residues conserved among Csa2 are cyan. The resulting model is then exported in the Universal 3D (.u3d) format.

Incorporating 3D Content into Documents with Adobe 9 Pro Extended

The Universal 3D file be either incorporated into an existing or a new pdf file. Incorporating the model into a new file is slightly simpler but the model can only appear on its own page without other text or objects in the final document. The model can be incorporated into a new file by launching Adobe Acrobat 9 Pro Extended then clicking “File >> Create PDF >> From File” and selecting the .u3d file exported from Adobe 3-D Reviewer. Alternatively, 3D content can be incorporated into an existing file. This method is useful if the author wants it to appear as if it were a figure in a paper. This second method was used to generate figure 7-1. In this case, the rest of the document must first be written using an appropriate software such as Word. An appropriately sized area must be left blank where 3D content will be placed. The document must then be converted to PDF and opened using Adobe 9 Pro Extended. The 3-D PDF can be incorporated into the document by selecting Tools >> Multimedia >> 3-D Tool. The author then selects a rectangular canvas where the 3-D content will appear and is prompted to select a file; choose the .u3d file exported from Adobe 3-D Reviewer.

Once the 3D content is imported into Adobe Acrobat 9 Pro Extended, the “poster”, background, and views can be set up. A “poster” is an image that will appear in place of the 3D PDF before the reader activates the content by clicking on it. The poster can simply be the default view of the 3D model or it can be another image. To use another image the user must right-click on the 3D content, select “properties”, select “create poster from file” under the “launch settings” tab and select an image file in .pdf or .jpg format. The image file must have the same dimensions as the 3D-PDF canvas or it will be stretched to fill the canvas and distorted. The background color of the 3D content can also be changed using the “properties” window under the “3D” tab.

The author can then designate pre-set views that the reader will have the option of choosing from. When the 3D content is activated by clicking on it a tool-bar will appear above the content. Immediately to the right of a drop-down menu labeled “views” there is an icon which toggles the “Model-tree” on and off. The model tree can, in turn, be used to toggle various layers of the model on and off. The model can be rotated by left-clicking and dragging, zoomed by right-clicking and dragging or panned by holding Ctrl, left-clicking and dragging. Pre-set views for the reader to choose from can be generated by moving the model to the desired position and toggling the desired portions on using the Model tree then selecting “Manage Views” from the dropdown “Views” menu. The “Manage Views” window can be used to create new views, delete views, change the order in which the views will appear in the list presented to the reader and designate one view to be the “Default view” which the reader will first see upon activating the content. For the Csa2 example, the “default view was generated by changing the background color to

white, toggling on all four “cartoon” layers and toggling off all four “surface” layers.

Another view, showing the conserved surface features was generated in which the cartoon layers were toggled “off” and the surface layers were toggled “on.” Finally a third view was generated in which the surface rendering was rotated 90° to show the most conserved region. This procedure was used to generate figure 7-1.

Discussion

The 3D-PDF format has great potential to be a useful tool for the presentation of not just protein atomic structures but any data that can be nicely displayed as a 3-dimensional interactive model including cryo-EM maps, electron tomography models and fluorescence microscopy Z-stacks. While Adobe Acrobat 9 Pro Extended has the ability to do direct “screen captures” of 3-D content from programs like Pymol, it was usually unable to correctly render some features, particularly “surfaces.” The renderings are also too low in resolution to be of publication quality. With direct “screen capture” the user also loses the ability to assign multiple layers to one figure. Therefore I recommend the procedure outlined in this chapter.

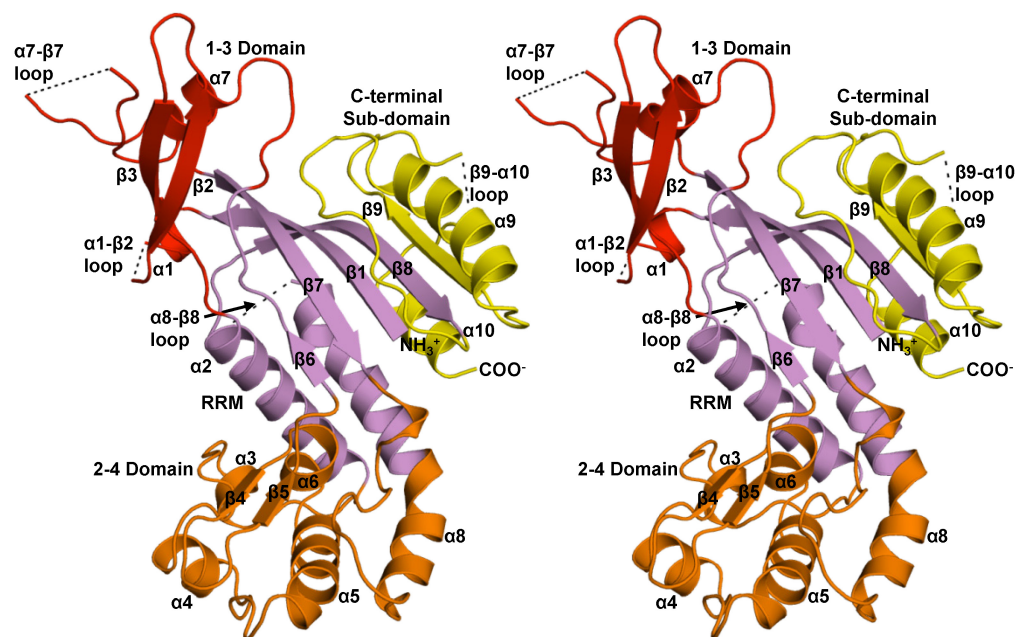


Figure 7-1: (Embedded 3D content, see below for instructions). (Poster) The structure of Csa2 reveals a novel domain architecture containing an RNA-recognition motif (chain A shown). Stereo ribbon diagram of the Csa2 monomer. The RNA-recognition motif is colored violet, the 1-3 domain is red, the 2-4 domain is orange and the C-terminal subdomain is colored yellow. (Embedded interactive content requires the free Adobe Reader software, version 9 or later and can be activated by clicking on any part of the figure. The model can be manipulated interactively using the mouse. Options for selecting, rotating, panning and zooming are available in the toolbar or contextual menu. Parts of the model can be individually accessed and toggled on or off using the model tree. Preset views can be accessed using the dropdown “views” menu and include a “cartoon” rendering, a “surface” rendering with conserved features highlighted, and a cartoon rendering with conserved residues depicted with “sticks.” To end 3D viewing, right-click on the model and select “disable content”; for MAC users, Ctrl+click.

REFERENCES CITED

1. Wilhelm, S. W. & Suttle, C. A. (1999) *Bioscience* **49**, 781-788.
2. Whittaker, R. H. (1959) *Q Rev Biol* **34**, 210-26.
3. Woese, C. R., Kandler, O. & Wheelis, M. L. (1990) *Proc Natl Acad Sci U S A* **87**, 4576-9.
4. Barns, S. M., Delwiche, C. F., Palmer, J. D. & Pace, N. R. (1996) *Proc Natl Acad Sci U S A* **93**, 9188-93.
5. Huber, H., Hohn, M. J., Rachel, R., Fuchs, T., Wimmer, V. C. & Stetter, K. O. (2002) *Nature* **417**, 63-7.
6. Albers, S. V. & Driessen, A. J. (2008) *Archaea (Vancouver, B.C.)* **2**, 145-9.
7. Beloglazova, N., Brown, G., Zimmerman, M. D., Proudfoot, M., Makarova, K. S., Kudritska, M., Kochinyan, S., Wang, S., Chruszcz, M., Minor, W., Koonin, E. V., Edwards, A. M., Savchenko, A. & Yakunin, A. F. (2008) *The Journal of biological chemistry*.
8. Berkner, S. & Lipps, G. (2008) *Archives of microbiology* **190**, 217-30.
9. Han, D. & Krauss, G. (2009) *FEBS Lett* **583**, 771-6.
10. Happonen, L. J., Redder, P., Peng, X., Reigstad, L. J., Prangishvili, D. & Butcher, S. J. (2010) *J Virol* **84**, 4747-54.
11. Lillestol, R. K., Shah, S. A., Brugger, K., Redder, P., Phan, H., Christiansen, J. & Garrett, R. A. (2009) *Mol Microbiol* **72**, 259-72.
12. Maaty, W. S., Ortmann, A. C., Dlakic, M., Schulstad, K., Hilmer, J. K., Liepold, L., Weidenheft, B., Khayat, R., Douglas, T., Young, M. J. & Bothner, B. (2006) *J Virol* **80**, 7625-35.

13. Martusewitsch, E., Sensen, C. W. & Schleper, C. (2000) *Journal of bacteriology* **182**, 2574-81.
14. Ortmann, A. C., Brumfield, S. K., Walther, J., McInnerney, K., Brouns, S. J., van de Werken, H. J., Bothner, B., Douglas, T., van de Oost, J. & Young, M. J. (2008) *Journal of virology* **82**, 4874-83.
15. Ortmann, A. C., Wiedenheft, B., Douglas, T. & Young, M. (2006) *Nat Rev Microbiol* **4**, 520-8.
16. Peng, X., Brugger, K., Shen, B., Chen, L., She, Q. & Garrett, R. A. (2003) *J Bacteriol* **185**, 2410-7.
17. She, Q., Singh, R. K., Confalonieri, F., Zivanovic, Y., Allard, G., Awayez, M. J., Chan-Weiher, C. C., Clausen, I. G., Curtis, B. A., De Moors, A., Erauso, G., Fletcher, C., Gordon, P. M., Heikamp-de Jong, I., Jeffries, A. C., Kozera, C. J., Medina, N., Peng, X., Thi-Ngoc, H. P., Redder, P., Schenk, M. E., Theriault, C., Tolstrup, N., Charlebois, R. L., Doolittle, W. F., Duguet, M., Gaasterland, T., Garrett, R. A., Ragan, M. A., Sensen, C. W. & Van der Oost, J. (2001) *Proc Natl Acad Sci U S A* **98**, 7835-40.
18. She, Q., Zhang, C., Deng, L., Peng, N., Chen, Z. & Liang, Y. X. (2009) *Biochemical Society transactions* **37**, 92-6.
19. Wiedenheft, B., Mosolf, J., Willits, D., Yeager, M., Dryden, K. A., Young, M. & Douglas, T. (2005) *Proc Natl Acad Sci U S A* **102**, 10551-6.
20. Lawrence, C. M., Menon, S., Eilers, B. J., Bothner, B., Khayat, R., Douglas, T. & Young, M. J. (2009) *J Biol Chem* **284**, 12599-603.
21. Suttle, C. A. (2007) *Nat Rev Microbiol* **5**, 801-12.
22. Forterre, P. (2006) *Virus Res* **117**, 5-16.
23. Forterre, P. (2006) *Proc Natl Acad Sci U S A* **103**, 3669-74.

24. Albers, S. V., Jonuscheit, M., Dinkelaker, S., Ulrich, T., Kletzin, A., Tampe, R., Driessen, A. J. & Schleper, C. (2006) *Appl Environ Microbiol* **72**, 102-11.
25. Aucelli, T., Contursi, P., Girfoglio, M., Rossi, M. & Cannio, R. (2006) *Nucleic acids research* **34**, e114.
26. Berkner, S., Grogan, D., Albers, S. V. & Lipps, G. (2007) *Nucleic acids research* **35**.
27. Zillig, W., Prangishvili, D., Schleper, C., Elferink, M., Holz, I., Albers, S., Janekovic, D. & Gotz, D. (1996) *FEMS Microbiol Rev* **18**, 225-36.
28. Prangishvili, D., Garrett, R. A. & Koonin, E. V. (2006) *Virus Res* **117**, 52-67.
29. Kraft, P., Kummel, D., Oeckinghaus, A., Gauss, G. H., Wiedenheft, B., Young, M. & Lawrence, C. M. (2004) *J Virol* **78**, 7438-42.
30. Kraft, P., Oeckinghaus, A., Kummel, D., Gauss, G. H., Gilmore, J., Wiedenheft, B., Young, M. & Lawrence, C. M. (2004) *J Virol* **78**, 11544-50.
31. Menon, S. K., Eilers, B. J., Young, M. J. & Lawrence, C. M. (2010) *J Virol* **84**, 5890-7.
32. Menon, S. K., Maaty, W. S., Corn, G. J., Kwok, S. C., Eilers, B. J., Kraft, P., Gillitzer, E., Young, M. J., Bothner, B. & Lawrence, C. M. (2008) *Virology* **376**, 270-8.
33. Schlenker, C., Menon, S., Lawrence, C. M. & Copie, V. (2009) *Biomol NMR Assign* **3**, 219-22.
34. Larson, E. T., Eilers, B., Menon, S., Reiter, D., Ortmann, A., Young, M. J. & Lawrence, C. M. (2007) *Virology* **368**, 249-61.
35. Larson, E. T., Eilers, B. J., Reiter, D., Ortmann, A. C., Young, M. J. & Lawrence, C. M. (2007) *Virology* **363**, 387-96.

36. Larson, E. T., Reiter, D., Young, M. & Lawrence, C. M. (2006) *J Virol* **80**, 7636-44.
37. Rice, G., Tang, L., Stedman, K., Roberto, F., Spuhler, J., Gillitzer, E., Johnson, J. E., Douglas, T. & Young, M. (2004) *Proc Natl Acad Sci U S A* **101**, 7716-20.
38. Fulton, J., Bothner, B., Lawrence, M., Johnson, J. E., Douglas, T. & Young, M. (2009) *Biochem Soc Trans* **37**, 114-7.
39. Khayat, R., Tang, L., Larson, E. T., Lawrence, C. M., Young, M. & Johnson, J. E. (2005) *Proc Natl Acad Sci U S A* **102**, 18944-9.
40. Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D. A. & Horvath, P. (2007) *Science (New York, N.Y.)* **315**, 1709-12.
41. Brouns, S. J., Jore, M. M., Lundgren, M., Westra, E. R., Slijkhuis, R. J., Snijders, A. P., Dickman, M. J., Makarova, K. S., Koonin, E. V. & van der Oost, J. (2008) *Science (New York, N.Y.)* **321**, 960.
42. Horvath, P. & Barrangou, R. (2010) *Science* **327**, 167-70.
43. Sorek, R., Kunin, V. & Hugenholtz, P. (2008) *Nature reviews. Microbiology* **6**, 181-6.
44. van der Oost, J., Jore, M. M., Westra, E. R., Lundgren, M. & Brouns, S. J. (2009) *Trends Biochem Sci* **34**, 401-7.
45. Grissa, I., Vergnaud, G. & Pourcel, C. (2007) *BMC Bioinformatics* **8**, 172.
46. Grissa, I., Vergnaud, G. & Pourcel, C. (2007) *Nucleic acids research* **35**, W52-7.
47. Lillestol, R. K., Redder, P., Garrett, R. A. & Brugger, K. (2006) *Archaea* **2**, 59-72.
48. Gudbergdottir, S., Deng, L., Chen, Z., Jensen, J. V., Jensen, L. R., She, Q. & Garrett, R. A. (2011) *Mol Microbiol* **79**, 35-49.

49. Karginov, F. V. & Hannon, G. J. (2010) *Mol Cell* **37**, 7-19.
50. Banfield, J. F. & Young, M. (2009) *Science* **326**, 1198-9.
51. Andersson, A. F. & Banfield, J. F. (2008) *Science* **320**, 1047-50.
52. Mojica, F. J., Diez-Villasenor, C., Garcia-Martinez, J. & Soria, E. (2005) *J Mol Evol* **60**, 174-82.
53. Bolotin, A., Quinquis, B., Sorokin, A. & Ehrlich, S. D. (2005) *Microbiology* **151**, 2551-61.
54. Pourcel, C., Salvignol, G. & Vergnaud, G. (2005) *Microbiology* **151**, 653-63.
55. Shah, S. A., Hansen, N. R. & Garrett, R. A. (2009) *Biochemical Society transactions* **37**, 23-8.
56. Tyson, G. W. & Banfield, J. F. (2008) *Environmental microbiology* **10**, 200.
57. Jansen, R., Embden, J. D., Gastra, W. & Schouls, L. M. (2002) *Mol Microbiol* **43**, 1565-75.
58. Haft, D., Selengut, J., Mongodin, E. & Nelson, K. (2005) *Plos Computational Biology* **1**, 474-483.
59. Makarova, K. S., Grishin, N. V., Shabalina, S. A., Wolf, Y. I. & Koonin, E. V. (2006) *Biology direct* **1**, 7.
60. Carte, J., Wang, R., Li, H., Terns, R. M. & Terns, M. P. (2008) *Genes & development* **22**, 3489-96.
61. Wiedenheft, B., Zhou, K., Jinek, M., Coyle, S. M., Ma, W. & Doudna, J. A. (2009) *Structure* **17**, 904-12.

62. Marraffini, L. A. & Sontheimer, E. J. (2008) *Science (New York, N.Y.)* **322**, 1843-5.
63. Hale, C. R., Zhao, P., Olson, S., Duff, M. O., Graveley, B. R., Wells, L., Terns, R. M. & Terns, M. P. (2009) *Cell* **139**, 945-56.
64. Babu, M., Beloglazova, N., Flick, R., Graham, C., Skarina, T., Nocek, B., Gagarinova, A., Pogoutse, O., Brown, G., Binkowski, A., Phanse, S., Joachimiak, A., Koonin, E. V., Savchenko, A., Emili, A., Greenblatt, J., Edwards, A. M. & Yakunin, A. F. (2011) *Mol Microbiol* **79**, 484-502.
65. Han, D., Lehmann, K. & Krauss, G. (2009) *FEBS Lett* **583**, 1928-32.
66. Ebihara, A., Yao, M., Masui, R., Tanaka, I., Yokoyama, S. & Kuramitsu, S. (2006) *Protein Sci* **15**, 1494-9.
67. Haurwitz, R. E., Jinek, M., Wiedenheft, B., Zhou, K. & Doudna, J. A. (2010) *Science* **329**, 1355-8.
68. Garneau, J. E., Dupuis, M. E., Villion, M., Romero, D. A., Barrangou, R., Boyaval, P., Fremaux, C., Horvath, P., Magadan, A. H. & Moineau, S. (2010) *Nature* **468**, 67-71.
69. Jore, M. M., Lundgren, M., van Dujn, E., Bultema, J. B., Westra, E. R., Waghmare, S. P., Wiedenheft, B., Pul, U., Wurm, R., Wagner, R., Beijer, M. R., Berndregt, A., Zhou, K., Snijders, A. P. L., Dickman, M. J., Doudna, J. A., Boekema, E. J., Heck, A. J. R., van der Oost, J. & Brouns, S. J. (2011) *Nature Structural and Molecular Biology* **In Press**.
70. Hale, C., Kleppe, K., Terns, R. M. & Terns, M. P. (2008) *Rna* **14**, 2572-9.
71. Shinkai, A., Kira, S., Nakagawa, N., Kashihara, A., Kuramitsu, S. & Yokoyama, S. (2007) *Journal of bacteriology* **189**, 3891.
72. Pul, U., Wurm, R., Arslan, Z., Geissen, R., Hofmann, N. & Wagner, R. (2010) *Mol Microbiol* **75**, 1495-512.

73. Westra, E. R., Pul, U., Heidrich, N., Jore, M. M., Lundgren, M., Stratmann, T., Wurm, R., Raine, A., Mescher, M., Van Heereveld, L., Mastop, M., Wagner, E. G., Schnetz, K., Van Der Oost, J., Wagner, R. & Brouns, S. J. (2010) *Mol Microbiol* **77**, 1380-93.
74. Tang, T. H., Polacek, N., Zywicki, M., Huber, H., Brugger, K., Garrett, R., Bachellerie, J. P. & Huttenhofer, A. (2005) *Mol Microbiol* **55**, 469-81.
75. Thoma, R., Obmolova, G., Lang, D. A., Schwander, M., Jenö, P., Sterner, R. & Wilmanns, M. (1999) *FEBS Lett* **454**, 1-6.
76. Otwinowski, Z. & Minor, W. (1997) in *Macromolecular Crystallography, Part A*, eds. Carter, C. & Sweet, R. (Academic Press, Vol. 276, pp. 307-326.
77. Terwilliger, T. C. & Berendzen, J. (1999) *Acta Crystallogr D Biol Crystallogr* **55**, 849-61.
78. Terwilliger, T. C. (2000) *Acta Crystallogr D Biol Crystallogr* **56**, 965-72.
79. Emsley, P. & Cowtan, K. (2004) *Acta Crystallogr D Biol Crystallogr* **60**, 2126-32.
80. Bailey, S. (1994) *Acta Cryst.* **D50**, 760-763.
81. Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997) *Acta Cryst.* **D53**, 240-255.
82. Painter, J. & Merritt, E. A. (2006) *Acta Crystallogr D Biol Crystallogr* **62**, 439-50.
83. Davis, I. W., Leaver-Fay, A., Chen, V. B., Block, J. N., Kapral, G. J., Wang, X., Murray, L. W., Arendall, W. B., 3rd, Snoeyink, J., Richardson, J. S. & Richardson, D. C. (2007) *Nucleic Acids Res* **35**, W375-83.
84. Holm, L. & Sander, C. (1993) *J Mol Biol* **233**, 123-38.
85. Gibrat, J. F., Madej, T. & Bryant, S. H. (1996) *Curr Opin Struct Biol* **6**, 377-85.

86. Krissinel, E. & Henrick, K. (2004) *Acta crystallographica. Section D, Biological crystallography* **60**, 2256-68.
87. DeLano, W. L. (2002).
88. Haft, D. H., Selengut, J. D., Brinkac, L. M., Zafar, N. & White, O. (2005) *Bioinformatics* **21**, 293-306.
89. Haft, D. H., Selengut, J. D. & White, O. (2003) *Nucleic acids research* **31**, 371-3.
90. Studier, F. W. (2005) *Protein expression and purification* **41**, 207.
91. Gasteiger, E., Hoogland, C., Gattiker, A., Duvaud, S., Wilkins, M. R., Appel, R. D. & Bairoch, A. (2005) in *The Proteomics Handbook*, ed. Walker, J. M. (Humana Press, pp. 571-607.
92. Brock, T. D., Brock, K. M., Belly, R. T. & Weiss, R. L. (1972) *Arch Mikrobiol* **84**, 54-68.
93. Candiano, G., Bruschi, M., Musante, L., Santucci, L., Ghiggeri, G. M., Carnemolla, B., Orecchia, P., Zardi, L. & Righetti, P. G. (2004) *Electrophoresis* **25**, 1327-33.
94. Maaty, W. S., Wiedenheft, B., Tarlykov, P., Schaff, N., Heinemann, J., Robison-Cox, J., Valenzuela, J., Dougherty, A., Blum, P., Lawrence, C. M., Douglas, T., Young, M. J. & Bothner, B. (2009) *PLoS One* **4**, e6964.
95. Bradford, M. M. (1976) *Anal Biochem* **72**, 248-54.
96. Svergun, D. I., Petoukhov, M. V. & Koch, M. H. (2001) *Biophys J* **80**, 2946-53.
97. Hura, G. L., Menon, A. L., Hammel, M., Rambo, R. P., Poole, F. L., 2nd, Tsutakawa, S. E., Jenney, F. E., Jr., Classen, S., Frankel, K. A., Hopkins, R. C., Yang, S. J., Scott, J. W., Dillard, B. D., Adams, M. W. & Tainer, J. A. (2009) *Nat Methods* **6**, 606-12.

98. Painter, J. & Merritt, E. A. (2006) *Journal of Applied Crystallography* **39**, 109-111.
99. Lovell, S. C., Davis, I. W., Arendall, W. B., 3rd, de Bakker, P. I., Word, J. M., Prisant, M. G., Richardson, J. S. & Richardson, D. C. (2003) *Proteins* **50**, 437-50.
100. Tong, L. A. & Rossmann, M. G. (1990) *Acta Crystallogr A* **46 (Pt 10)**, 783-92.
101. Tempel, W., Liu, Z. J., Schubot, F. D., Shah, A., Weinberg, M. V., Jenney, F. E., Jr., Arendall, W. B., 3rd, Adams, M. W., Richardson, J. S., Richardson, D. C., Rose, J. P. & Wang, B. C. (2004) *Proteins* **57**, 878-82.
102. Matias, P. M., Gorynia, S., Donner, P. & Carrondo, M. A. (2006) *J Biol Chem* **281**, 38918-29.
103. Krishna, T. S., Kong, X. P., Gary, S., Burgers, P. M. & Kuriyan, J. (1994) *Cell* **79**, 1233-43.
104. Maris, C., Dominguez, C. & Allain, F. H. (2005) *Febs J* **272**, 2118-31.
105. Lintner, N. G., Frankel, K. A., Tsutakawa, S. E., Alsbury, D. L., Copie, V., Young, M. J., Tainer, J. A. & Lawrence, C. M. (2011) *J Mol Biol* **405**, 939-55.
106. Aravind, L., Anantharaman, V., Balaji, S., Babu, M. M. & Iyer, L. M. (2005) *FEMS Microbiol Rev* **29**, 231-62.
107. Alekshun, M. N., Levy, S. B., Mealy, T. R., Seaton, B. A. & Head, J. F. (2001) *Nat Struct Biol* **8**, 710-4.
108. Putnam, C. D., Hammel, M., Hura, G. L. & Tainer, J. A. (2007) *Q Rev Biophys* **40**, 191-285.
109. Tsutakawa, S. E., Hura, G. L., Frankel, K. A., Cooper, P. K. & Tainer, J. A. (2007) *J Struct Biol* **158**, 214-23.

110. Krissinel, E. & Henrick, K. (2007) *J Mol Biol* **372**, 774-97.
111. Chen, C. S., White, A., Love, J., Murphy, J. R. & Ringe, D. (2000) *Biochemistry* **39**, 10397-407.
112. Wu, R. Y., Zhang, R. G., Zagnitko, O., Dementieva, I., Maltzev, N., Watson, J. D., Laskowski, R., Gornicki, P. & Joachimiak, A. (2003) *J Biol Chem* **278**, 20240-4.
113. Lim, D., Poole, K. & Strynadka, N. C. (2002) *J Biol Chem* **277**, 29253-9.
114. Hong, M., Fuangthong, M., Helmann, J. D. & Brennan, R. G. (2005) *Mol Cell* **20**, 131-41.
115. Pelikan, M., Hura, G. L. & Hammel, M. (2009) *Gen Physiol Biophys* **28**, 174-89.
116. O'Sullivan, O., Suhre, K., Abergel, C., Higgins, D. G. & Notredame, C. (2004) *J Mol Biol* **340**, 385-95.
117. Bond, C. S. & Schuttelkopf, A. W. (2009) *Acta Crystallogr D Biol Crystallogr* **65**, 510-2.
118. Liang, J., Edelsbrunner, H. & Woodward, C. (1998) *Protein Sci* **7**, 1884-97.
119. Landau, M., Mayrose, I., Rosenberg, Y., Glaser, F., Martz, E., Pupko, T. & Ben-Tal, N. (2005) *Nucleic Acids Res* **33**, W299-302.
120. Marchler-Bauer, A., Anderson, J. B., Chitsaz, F., Derbyshire, M. K., DeWeese-Scott, C., Fong, J. H., Geer, L. Y., Geer, R. C., Gonzales, N. R., Gwadz, M., He, S., Hurwitz, D. I., Jackson, J. D., Ke, Z., Lanczycki, C. J., Liebert, C. A., Liu, C., Lu, F., Lu, S., Marchler, G. H., Mullokandov, M., Song, J. S., Tasneem, A., Thanki, N., Yamashita, R. A., Zhang, D., Zhang, N. & Bryant, S. H. (2009) *Nucleic Acids Res* **37**, D205-10.
121. Soding, J. (2005) *Bioinformatics* **21**, 951-60.

122. Agari, Y., Sakamoto, K., Tamakoshi, M., Oshima, T., Kuramitsu, S. & Shinkai, A. (2010) *J Mol Biol* **395**, 270-81.
123. Christen, M., Kulasekara, H. D., Christen, B., Kulasekara, B. R., Hoffman, L. R. & Miller, S. I. (2010) *Science* **328**, 1295-7.
124. Monds, R. D., Newell, P. D., Wagner, J. C., Schwartzman, J. A., Lu, W., Rabinowitz, J. D. & O'Toole, G. A. (2010) *J Bacteriol* **192**, 3011-23.
125. Rambo, R. P. & Tainer, J. A. (2010) *Curr Opin Struct Biol* **20**, 128-37.
126. Hugler, M., Krieger, R. S., Jahn, M. & Fuchs, G. (2003) *Eur J Biochem* **270**, 736-44.
127. Carte, J., Pfister, N. T., Compton, M. M., Terns, R. M. & Terns, M. P. (2010) *Rna* **16**, 2181-8.
128. Kelley, L. A. & Sternberg, M. J. (2009) *Nat Protoc* **4**, 363-71.
129. Wang, R., Preamplume, G., Terns, M. P., Terns, R. M. & Li, H. (2011) *Structure* **19**, 257-64.
130. Kunin, V., Sorek, R. & Hugenholtz, P. (2007) *Genome biology* **8**, R61.
131. Sinz, A. (2010) *Anal Bioanal Chem* **397**, 3433-40.
132. Kessler, A., Sezonov, G., Guijarro, J. I., Desnoues, N., Rose, T., Delepierre, M., Bell, S. D. & Prangishvili, D. (2006) *Nucleic acids research* **34**, 4837-45.
133. Marraffini, L. A. & Sontheimer, E. J. (2010) *Nature* **463**, 568-71.
134. Mojica, F. J., Díez-Villaseñor, C., García-Martínez, J. & Almendros, C. (2009) *Microbiology (Reading, England)* **155**, 733-40.

135. Anonymous (2009) *J Biol Chem* **284**, 21101.
136. Kumar, P., Ziegler, A., Grahn, A., Hee, C. S. & Ziegler, A. (2010) *Trends Biochem Sci* **35**, 419-22.
137. Kumar, P., Ziegler, A., Ziegler, J., Uchanska-Ziegler, B. & Ziegler, A. (2008) *Trends Biochem Sci* **33**, 408-12.

APPENDIX A

PYMOL SCRIPT FOR GENERATION OF 3-D PDFs

What follows is an annotated Pymol script, which can be used to reproduce figure 7-1 3-D PDF's. It is formatted such that it will run as is if copied and pasted into a Pymol script.

```
# pymol script
#
# The following set of commands loads the coordinate files and names the
# resulting models. To run this example script, download the Csa2
# coordinates from the PDB (3PS0)
#####
load [filepath]\3PS0_chainA_nowaters.pdb, csa2
#####

# The following set of commands are settings, which I have found by
# experience, makes the figures appear the way I like them.

hide everything
set orthoscopic = 1\# 1 = perspective off; 0 = perspective on

set seq_view_overlay, 1
set cartoon_flat_sheets = 1
set cartoon_smooth_loops = 0
set cartoon_fancy_helices = 1
set cartoon_dumbbell_length, 1 # default width = 1.6
set depth_cue = 0
set opaque_background, off
set ray_shadows, 0
set specular, off
set dash_radius, 0.1
set dash_gap, 1
set dash_length, 0.4
set dash_round_ends, 0
space cmyk
bg white

#####

# The following set of commands assign "selections" to each portion of
# the model that I want to have in a different layer. This is the most
# important portion of the script for making 3-D PDF's

# "Selections" of the four domains/subdomains are assigned. Unlike in a
# standard pymol script the domains must overlap by one residue.

select 1-3_domain, chain A and resi 12-49+145-180
select c-term_subdomain, chain A and resi 256-320
select 2-4_domain, chain A and resi 68-136+191-216
select rrm, chain A and resi 136-145+49-69+216-246+180-191+ \
1-12+244-256
color grey90, Csa2
```

Selections of conserved residues are assigned.

```
select cons, csa2 and chain A and resi \
7+22+23+46+50+55+108+119+122+123+133+135+144+160+162+178+184+ \
218+225+240+243+269
select cons_1857, csa2 and chain A and resi 14+15+16+58+121+122
```

Atoms colors are assigned for the "sticks" layer.

```
color cyan, cons and name C*
color red, cons and name O*
color blue, cons and name N*
color orange, cons_1857 and name C*
color red, cons_1857 and name O*
color blue, cons_1857 and name N*
```

**# The distance command is used to draw dotted lines connecting
the gaps in the model.**

```
distance d1 = (chain A and resi 16 and name CA), (chain A and
resi 27 and name CA)
distance d2 = (chain A and resi 164 and name CA), (chain A and
resi 173 and name CA)
distance d3 = (chain A and resi 232 and name CA), (chain A and
resi 241 and name CA)
distance d4 = (chain A and resi 302 and name CA), (chain A and
resi 306 and name CA)
show, d1
show, d2
show, d3
show, d4
color black, d1
color black, d2
color black, d3
color black, d4
set label_size, 0
```

**# The following command places the model in the a "standard" orientation
I have chosen to use.**

```
set_view (\
  0.271650136, -0.611722827, -0.742962718,\
  0.840249121, 0.527151883, -0.126810476,\
  0.469230354, -0.589829087, 0.657207549,\
  8.000000000, 0.000000000, -231.613937378,\
  109.840484619, 72.236984253, 91.072860718,\
  187.938888550, 275.288970947, -20.000000000 )
#####
viewport 1000, 1000
```