X-RAY CRYSTALLOGRAPHIC STUDIES OF THE PROTEINS FROM *SULFOLOBUS* SPINDLE-SHAPED VIRUSES (SSVs).

by

Smita Kesavankutty Menon

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

of

Doctor of Philosophy

in

Biochemistry

MONTANA STATE UNIVERSITY
Bozeman, Montana

July 2009

©COPYRIGHT

by

Smita Kesavankutty Menon

2009

All Rights Reserved

APPROVAL

of a dissertation submitted by

Smita Kesavankutty Menon

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 rules of the Library. I further agree that copying of this dissertation is allowable only for scholarly purposes, consistent with "fair use" as prescribed 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."

Smita Kesavankutty Menon

July 2009

ACKNOWLEDGEMENTS

I express my sincere gratitude to my teachers for their coveted guidance and help during all these years. I also convey my heartfelt thanks to my mentor Dr. Martin Lawrence for his constant encouragement and to all my colleagues, especially the past and present members of the Lawrence laboratory for all their help and co-operation. I would also like to thank my beloved husband and my family for their unconditional love and support during my time of need through all these years.

TABLE OF CONTENTS

TABLE OF CONTENTS – CONTINUED

TABLE OF CONTENTS – CONTINUED

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES – CONTINUED

Figure Page

ABSTRACT

Viruses populate virtually every ecosystem on the planet. Fuselloviridae are ubiquitous crenarchaeal viruses found in high-temperature acidic hot springs around the world. However, compared to eukaryotic and bacterial viruses, our knowledge of viruses infecting the archaea is limited. Fuselloviral genomes show little similarity to other organisms, generally precluding functional predictions. However, structural studies can reveal distant evolutionary relationships and provide functional insights that are not apparent from the primary amino acid sequence alone. Several such structural studies have already contributed to our understanding of the Sulfolobus Spindle-shaped viruses (*Fuselloviridae*). Here we report the structure of two proteins, SSV1 F112 and SSVRH D212. Biochemical, proteomic and structural studies of F112 reveal a monomeric intracellular protein that adopts a winged helix DNA binding fold. Continuing these efforts, a second structure was also determined where the overall fold and conservation of active site residues place D212 within the PD-(D/E)XK nuclease superfamily. Notably, the structure of F112 contains an intrachain disulfide bond, prompting analysis of cysteine usage in this and other hyperthermophilic viral genomes. The analysis supports a general abundance of disulfide bonds in the intracellular proteins of hyperthermophilic viruses and the evolutionary implications of such distribution are discussed. Here we review and describe our progress towards understanding these viruses at a molecular level.

CHAPTER 1

INTRODUCTION

Viruses represent the most abundant biological systems on earth [1] and are found to infect all cellular life forms. For these reasons, they significantly impact all systems of life. Recent studies have identified a key role for viruses not only in driving the evolution of their hosts but also in maintaining the ecology of the biosphere [1]. Due to the impact on their host and the ecosystem in which they exist, viruses from the bacterial and eukaryotic domains have proved to be essential paradigms for their more complex host organisms. Similarly, it is expected that the studies on viruses infecting archaea (the third domain of life) are most likely to provide a better insight into the mysterious archaea and their ecosystems. Although rapid advancement has been made in the field of archaeal viruses our understanding of these unique viruses at a molecular level is still in its infancy. The six chapters discussed in this thesis are an attempt to review and describe my work and our progress towards the understanding of the viruses which infect the hyperthermophilic archaeon, *Sulfolobus.* This chapter, in particular, provides background about one such family of viruses, the *fuselloviridae*, and describes in details about the genomes and proteomes of two of its members *Sulfolobus* Spindle-shaped Virus 1 (SSV1) and *Sulfolobus* Spindle-shaped Virus Ragged Hills (SSVRH).

Archaea

Based on the 16S ribosomal RNA (rRNA) sequence analysis Carl Woese and George Fox demonstrated that archaeal cells were quite different as compared to the bacterial cells and hence proposed their classification as a separate domain of life [2]. The archaea are further divided into four phyla (Figure 1-1), the Crenarchaeota, the Euryarchaeota, the Korarchaeota, and the Nanoarchaeota [2-4].

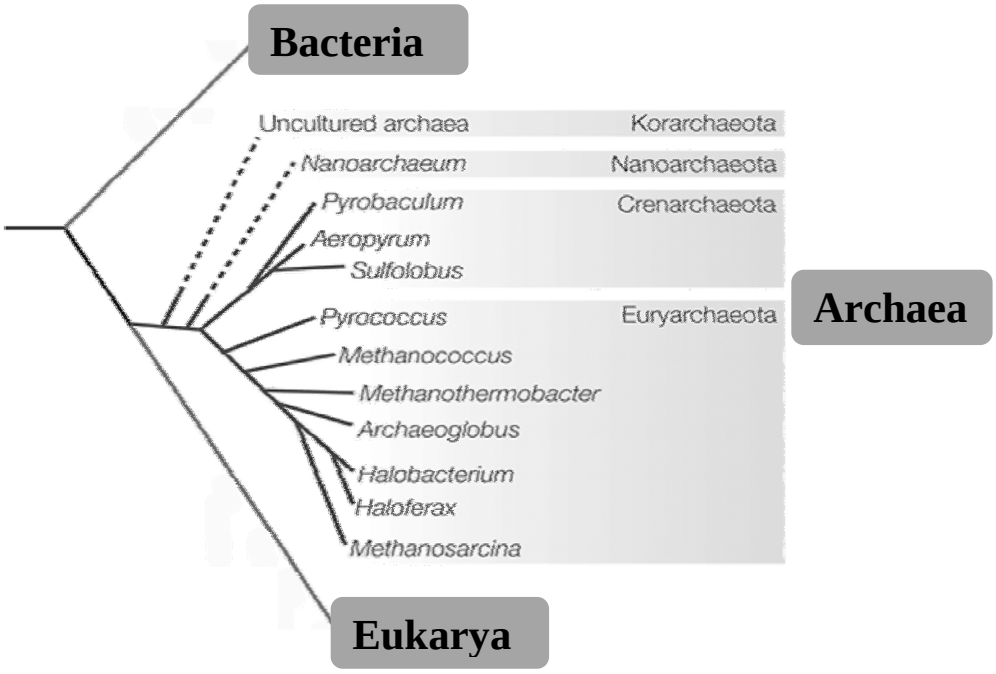


Figure 1-1. The Universal Tree of Life.
The universal tree is organized into three domains based upon rRNA sequence. The archaea are divided into four kingdoms the Crenarchaeota, the Euryarchaeota, the Nanoarchaeota, and the Korarchaeota. Figure has been modified from [5].

The Euryarchaeota [2] are primarily made up of anaerobic methanogens (methane producing organisms), aerobic halophiles (organisms that survive extreme salt concentrations) and thermophiles. The Crenarchaeota [2] were originally considered to be

either thermophilic or psychrophilic extremophiles. However, more recently they have also been identified as symbionts of temperate water sponge in mesophilic marine environments [6]. The Korarchaeota [3] are a small group of isolates from high temperature hydrothermal pools from Yellowstone National park with features distantly related to other archaea. While the fourth phylum the Nanoarchaeota [4] are tiny symbiotic microbes originally isolated from the hydrothermal vents in the deep ocean near Iceland, that require the presence of its host organism *Ignicoccus* for its survival. In addition, a newly discovered group of archaea from other members of the archaeal domain have been identified in the Iron Mountain mines of California and are named ARMAN for Archaeal Richmond Mine Acidophilic Nano organisms [7].

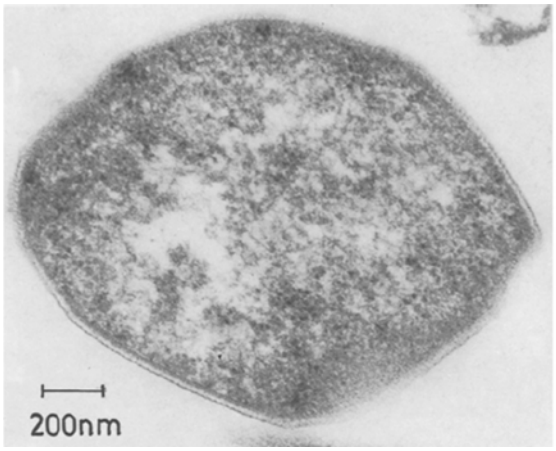


Figure 1-2. The Crenarchaeon *Sulfolobus.*
Electron micrographs of thin section of *Sulfolobus acidocaldarius* DSM639 [8]. *Sulfolobales* are aerobic sulfur oxidizing crenarchaeal organism. They serve as important model organism for the study of archaea.

Sulfolobales are one of the most widely studied crenarchaeal genera and are considered a model organism for the study of the archaea. *Sulfolobales* are aerobic crenarchaea ubiquitously found in high temperature (70-87^{o}C) acidic (pH 1.5 to 5.5) environments. They are flagellar regular or sometimes irregularly shaped coccoid cells

capable of growth on elemental sulfur or sulfur containing compounds. *Sulfolobus,* like most archaeal cells, possess a cell wall made up of surface-layer (S-layer) proteins. Complete genomes have been sequenced and deposited for three of the 50 known *Sulfolobus* species namely, *S.acidocaldarius* DSM 639 [9], *S.solfataricus* P2 [10], and *S.tokodaii* str. 7 [11]. Because of the ease in isolation and culturing of these cells in the laboratory they serve as a model for research on archaeal DNA replication, chromosomal integration, transcription and translation, RNA processing, and the archaeal cell cycle [12]. Despite significant progress, much remains to be learned about the crenarchaeal kingdom. One reason for this is the lack of availability of suitable genetic tools to manipulate *Sulfolobus* species, and other crenarchaeota. Currently, vectors based on *Sulfolobus* Spindle-shaped Virus 1, which can replicate both in *Sulfolobus solfataricus* and *Escherichia coli* are one of the few available tools for studying molecular biology in the archaea [5, 13, 14]. Along with genetic knockouts and microarray analysis, the archaeal viruses might also be an important tool in the field of characterization of the archaea particularly the *Sulfolobales*.

Archaeal Viruses

Archaeal viruses have primarily been isolated from the Euryarchaeota and the Crenarchaeota. The Euryarchaeal viruses infect extreme halophiles or methanogens or moderately thermophilic euryarchaea. These viruses frequently exhibit the commonly found head-and-tail bacteriophage like morphology and have been classified into two families the *Myoviridae* and the *Siphoviridae* [15]. These viruses enclose ~30-230 kilo base (kb) pair circular or linear double-stranded (ds) DNA genomes with a high G+C

content similar to their host [16]. More recently, a 7kb circular single-stranded (ss) DNA genome from a new haloarchaeal virus Halorubrum pleomorphic virus 1 (HRPV-1) has been isolated recently and is the only known example of an archaeal virus exhibiting a ssDNA genome [17]. In addition to their phage-like virion structures, the Euryarchaeal viruses also showcase several unique morphologies like the spindle shaped halovirus His1 [16]. These viruses to some extent resemble the bacteriophage in their genome content and possess a very mosaic genome which undergoes extensive genetic exchange [15]. No archaeal virus with an RNA genome has been isolated to date, possibly a bias introduced by the techniques used for virus isolation [16] or due to the greater instability of RNAs in the extreme environments where these viruses exist [18].

The Crenarchaeal viruses on the other hand have been isolated predominantly from acidophilic thermophiles specifically the genera *Sulfolobus* and *Acidianus*. They have also been found to sparsely infect neutrophilic *Thermoproteus* and *Pyrobaculum*. Crenarchaeal viruses identified thus far have shown extraordinary diversity in their morphology and genomes so as to warrant their classification into novel families. Based on their unique characteristics they have been classified into seven different families (Figure 1-3) , which include the spindle-shaped *fuselloviridae* [19-24], rod-shaped *rudiviridae* [25, 26], filamentous *lipothrixviridae* [27-31], two-tailed *bicaudaviridae* [32], droplet shaped *guttaviridae* [33], spherical shaped *globuloviridae* [34, 35] and bottle-shaped *ampullaviridae* [36], while the unclassified icosahedral Sulfolobus Turreted Icosahedral Virus, STIV [37] and large *Sulfolobus Tengchongensis* Spindle-shaped virus 1, STSV1 [38] still need to be categorized.

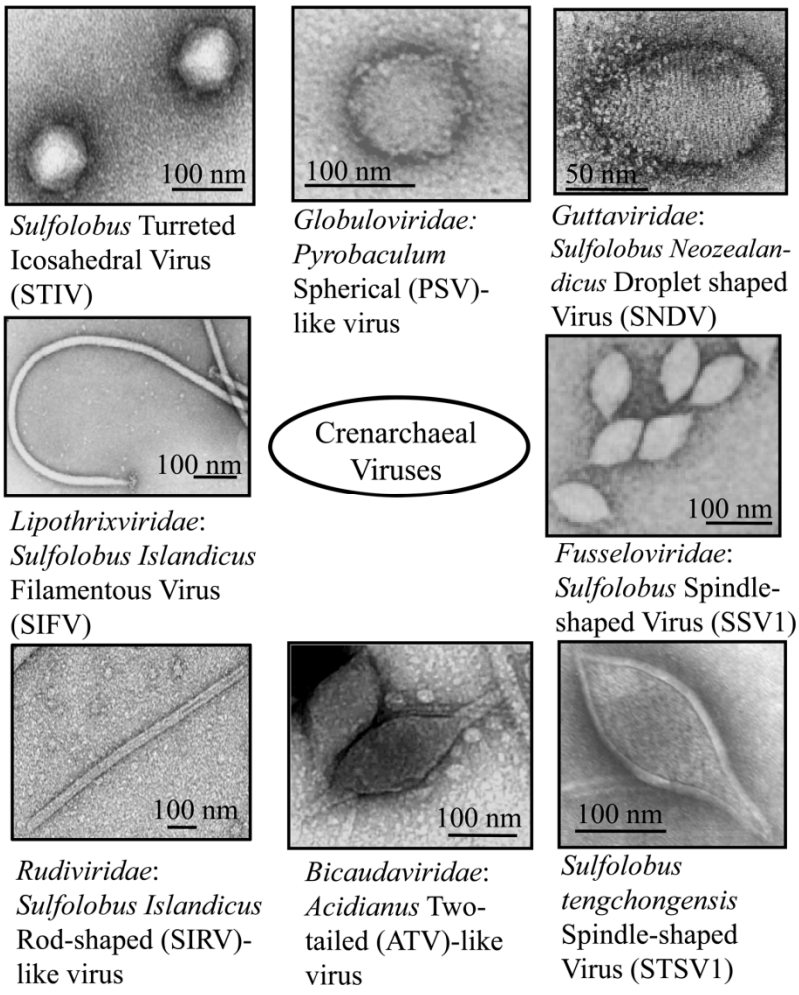


Figure 1-3. Morphological Diversity in Crenarchaeal Viruses.
Beginning at upper left follow clockwise: STIV [37], a PSV-like virus, SNDV [33], SSV1 [19], STSV1 [38] an ATV–like virus, an SIRV-like virus, and SIFV [39]. Micrographs of PSV-like, SIRV-like and ATV-like viruses from Yellowstone National Park (courtesy of Mark Young) while STIV was isolated from rabbit creek (YNP), SNDV was from New Zealand, SSV1 was from Japan, STSV1 was from Tengchong, China and SIFV was isolated from solfataric fields at Hveragerdi in southwest Iceland.

Most crenarchaeal viruses, apart from the *rudiviridae* family, are enveloped. These viruses encapsidate dsDNA genomes ranging in size from ~15 kbp to 75 kbp, which encode for 35 to 72 open reading frames (ORFs). Out of the 23 completed genomes listed in the NCBI database, the majority of the genomes encoded by these viruses are circular except for the filamentous *lipothrixviridae*, rod shaped *rudiviridae*

and the spherical *globuloviridae* which encapsidate linear genomes. *Thermoproteus tenax* virus 1 (TTV1, *Lipothrixviridae), Acidianus* two-tailed virus (ATV, *bicaudaviridae)* and the uncategorized STIV are the only known crenarchaeal viruses to cause cell lysis [16, 27, 40]. While *Sulfolobus* spindle-shaped viruses (SSVs) [21], ATV [41], and the STSV1 [38] are known to code for an integrase gene, most of these viruses are thought to be stably maintained as low copy number plasmids in the host cell without integration. *Sulfolobus* spindle-shaped viruses or *fuselloviridae* are the focus of this study and will be discussed in further detail.

Fuselloviridae

The *Sulfolobus* Spindle-shaped Viruses (SSVs) of the *fuselloviridae* family are by far the most well studied of the archaeal viruses. The SSV viruses are commonly found in high temperature (>70^{o}C), acidic (pH $\leq$ 4) hot springs around the world. All members of the fusellovirus family infect the hyperthermophilic crenarchaea of the genera *Sulfolobus*. They exhibit a characteristic ~60x90nm spindle or lemon shaped morphology (Figure 1-3) with short tail fibers at one end [20] which are presumed to be involved in viral attachment to the host [22]. Six SSV isolates have been characterized from thermal environments around the world. SSV1 was isolated from Beppu-Japan [19], SSV2 from Iceland [23], SSVRH isolated from the Ragged Hills area in Yellowstone National Park [22], SSVK from Kamchatka-Russia [22], and SSV4 and SSV5 from the hot springs of Iceland [24]. SSV1 is the *type* virus for fuselloviruses and serves as a paradigm for the study of this crenarchaeal viral family.

SSV1

Sulfolobus Spindle-shaped Virus 1, SSV1, was originally isolated from *Sulfolobus shibatae* and is the most extensively studied archaeal virus. SSV1 contains a 15.5 kbp circular double stranded DNA genome and exhibits a low G+C content of ~39%, similar

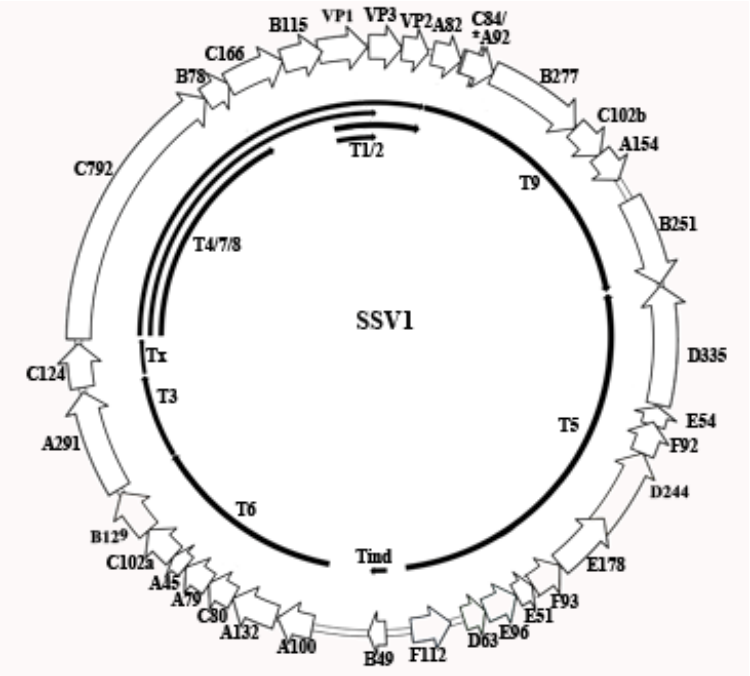


Figure 1-4. The SSV1 Genome.
The SSV1 Open Reading Frames are shown as open arrows, while transcript identities are shown in the interior (black lines). Figure was created with VectorNTI 10.1.1 (Invitrogen)

to its host. The SSV1 genome was found to be present in three different forms, as a provirus integrated into the host chromosome within the arginyl tRNA gene, as covalently closed supercoiled or relaxed episomal DNA, or as positively supercoiled DNA enclosed in the viral capsid [42]. The SSV1 genome was sequenced in 1991 by Palm et.al. and ten polycistronic transcripts (T1-T10) with seven different promoters and their terminators were identified [43]. Nine of the ten transcripts, T1-T9, appear to be

produced constitutively at low levels while the tenth transcript, Tind, was not found to be expressed in uninduced cells [43, 44]. Whereas on UV irradiation of the infected host cells, chronological transcription kicks in, beginning with an almost 16-fold increase in the amounts of the Tind transcript, which lacks a typical TATA-box like promoter [44]. This is followed by expression of T5, T6 and the slightly delayed T9 early transcripts. The remaining T1/2, T3 and T4/7/8 late transcripts are up regulated at a later time point during the SSV1 replication. The early termination of the T2 transcript results in transcript T1, which codes for the most abundant structural protein VP1 [43]. Similarly the leaky termination of transcript T8 results in the smaller transcript T7 [43]. Palm et al. noted another interesting feature, all the ORFs containing cysteines clustered in one half of the circular genome [20]. While, this trend is particularly noticeable in SSV1, where the cysteine containing proteins are encoded largely by the T5 and T6 transcripts [20], it is also apparent in the genomes of other *fuselloviridae* [23, 45]. The clustering of 14 cysteine containing Open Reading Frames (ORFs) in the SSV1 genome was believed to be a result of a genome fusion event [20].

Thirty-four ORFs were identified in the sequenced SSV1 genome [20]. The genes for three structural proteins VP1, VP2, and VP3 were recognized, as these proteins had been isolated from the purified viral particle and partial sequence determined by Edman degradation. However, for the remainders of these ORFs, statistically relevant similarities to sequences in the non-redundant NCBI database were not found. Thus despite complete genome sequence for SSV1, hypothetical functions based upon sequence homology could be suggested for only 3 of the putative gene products. With further genomic analysis ORF B251 was shown to be similar to DnaA, an ATP-dependant nucleotide binding

protein which promotes local unwinding of dsDNA [46], and D335 was identified as an integrase [47]. Interestingly, D335 was recently found to be non-essential for the survival of the virus [47, 48]. In the absence of the integrase however, no integration of the viral genome into host chromosome was seen [48].

More recent bioinformatics analysis carried out by Prangishvilli et.al [16] along with observations made by us in collaboration with M. Dlakic's lab (personal communication) has identified additional DNA binding motifs in the SSV1 genome. ORF E51 and C80 are predicted ribbon-helix-helix (RHH) proteins with weak similarity to the CopG family of proteins, while A45, A79, and B129 contain motifs suggesting the presence of C_2H_2 zinc fingers. Other putative functions for additional ORF encoded products have been proposed based on their transcriptional profiles from the microarray analysis done by Frols et.al. [44] . ORF C124 which is simultaneously expressed along with other capsid associated proteins at a late stage during the viral replication cycle suggests a similar structural role for the protein [44]. ORF B115, an ArsR-like putative transcriptional repressor, is also significantly up regulated at the end of the viral transcriptional cycle in both induced and non-induced cells which might suggest its possible role in the down regulation of the early SSV1 genes at the end of the viral cycle [44]. However 22 of 34 ORFs, ~65%, are still not reliably identified by these approaches.

As structural similarity between related proteins persists longer than similarity in their primary amino acid sequences, recent biochemical and structural analysis to determine such functional and evolutionary relationships were recently undertaken by our laboratory. These crystallographic studies reveal structural similarity between SSV1 D63 and Repressor of Primer (ROP) [49], an adaptor protein that serves to regulate colE1

plasmid copy number in E. coli [50]. D63 may thus play a similar role in regulating replication of the circular viral genome. Similarly, structural and biochemical characterization of F93 reveal a homodimeric winged-helix protein which may serve as a transcription factor or a replication terminator protein [51]. Similarly, the structure of B129 reveals two tandem C_2H_2 zinc finger domains (Eilers et al., unpublished) and an N-terminal dimerization domain that intertwines about itself to form a left-handed double superhelix. The resulting homodimer thus contains a two-fold symmetric array of 4 zinc finger domains. In contrast, the structure of E96 does not reveal structural homology to any protein with known function although it does show a strong and unexpected structurally similarity to ORF14 gene product from SIFV (Eilers et al., unpublished, [52]). Structure for SSV1 F112 was solved as a part of this work and will be discussed towards the latter part of this thesis.

SSVRH

Genomic sequence analysis of fusellovirus shows very limited similarity to other genomes in the database but demonstrates a close relationship with the other fusellovirus isolates. The *Sulfolobus* Spindle-shaped Virus Ragged Hills, or SSVRH, is a homolog of the SSV1 *type* virus that was isolated from the acidic thermal pools (pH = 3.2, T = 81^{o}C) in the Ragged Hills region of the Norris Geyser basin in Yellowstone National Park [22]. The complete genome sequence of SSVRH was determined by Wiedenheft et.al. [22]. SSVRH also exhibits characteristic 60x90nm spindle shaped morphology.

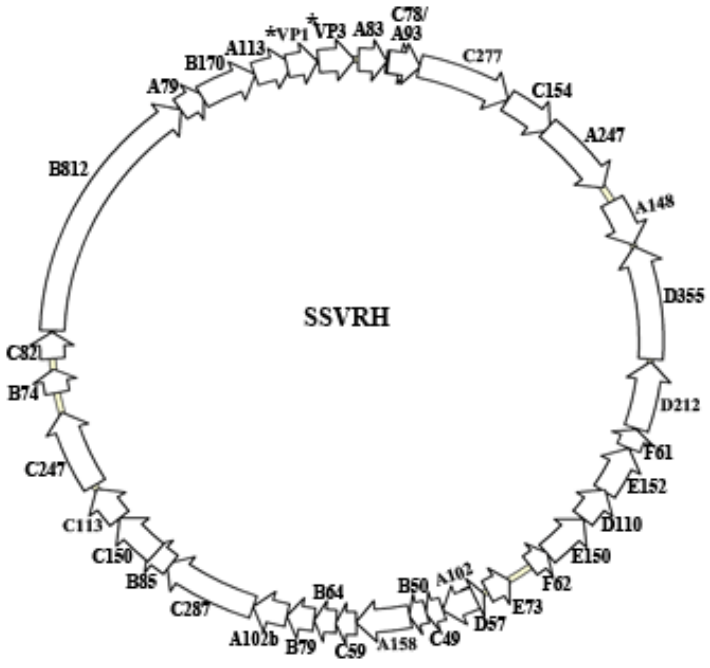


Figure 1-5. The SSVRH Genome.
The SSVRH ORFs are shown as open arrows, while those coding for particle associated proteins are labeled with an asterisk (*). Figure was created with VectorNTI 10.1.1 (Invitrogen).

The virion encapsidates a covalently closed positively supercoiled circular dsDNA genome similar to SSV1, with ~16.5kbp size genome that encodes for 39 ORFs. Table A-1 provides general information about these 39 putative gene products. The viral genome was shown to integrate into the leucyl tRNA gene of S*ulfolobus solfataricus* P2 genome [22].

In the case of SSVRH, functions for three of the putative gene products have been predicted. Two structural proteins VP1 and VP3 were identified by sequence homology to the SSV1 coat proteins [22]. However VP2, the third protein isolated from the mature SSV1 virion was found to be absent in SSVRH and the other SSV homologs. The viral integrase, D355, was identified through sequence similarities [22]. F61 with close sequence [49] and structural homology (Menon et.al. unpublished) to SSV1 D63 is

expected to have a similar Repressor of Primer (ROP) like adaptor protein function. ORFs E73 and B64, homologs of E51 and C80 from SSV1 respectively, are predicted to be ribbon-helix-helix (RHH) proteins with weak similarity to the CopG family of proteins. Based on sequence homology to SSV1 A79 and B129 respectively, ORFs B79 and C150 from SSVRH were found to contain zinc finger binding motifs. However, functional predictions for the remaining 29 ORFs of SSVRH are lacking. Structures for SSVRH D212 and SSVRH F61 were solved as a part of this work and will be discussed more extensively towards the end of this thesis.

Research Goals

On comparison of the archaeal viral proteins to protein sequences in the public database only very few proteins showed similarities to the sequences in the database. However, for few proteins similarities were found between the proteins from different archaeal viral families [16]. Such limited similarity could be a result of the extreme environments in which these viruses survive or due to the evolutionary distance from their bacterial and eukaryotic counterparts or a combination of both. Because protein tertiary structure is generally more conserved than the primary amino acid sequence of protein structural studies have the potential to reveal distant evolutionary relationships that are not discernible from the primary sequence alone. In turn these structural relationships may suggest a functional hypothesis. Indeed structural work in our laboratory [49, 51, 53, 54] and others [55, 56] have confirmed this, showing that structure can provide significant insights into protein function and providing strong validation for

the function from structural approach. Our initial effort to functionally annotate the SSV1 proteome was initiated by Dr. Paul Kraft, who cloned and expressed 34 ORFs from SSV1 with a C-terminal 6X His tag. However, a subset of these gene products failed to yield soluble proteins. Changing the growth conditions, the construct or the expression systems are some of the commonly used techniques to solve this problem. Therefore we initiated this study to determine the effect of changing the gene construct on soluble protein yields from the SSV1 genome. A test set made up of 11 ORFs from SSV1 were selected. New expression vectors were constructed in which the 6XHis tag relocated from the C-terminal to the N-terminal of the gene products for all the ORFs belonging to this test set. Table A-2 provides general information about these 11 SSV1 ORFs.

Two additional fuselloviral genomes, SSVK and SSVRH were sequenced by Wiedenheft et.al. [22]. This allowed a comparative genomics study between the four sequenced fuselloviral genomes available at that time, SSV1, SSV2, SSVRH and SSVK. This resulted in the identification of a core set of 18 genes, including the viral coat proteins VP1 and VP3 and the integrase, that are common to all 4 of these SSVs [22]. Thus although functions for a majority of the predicted proteins encoded by these SSV genomes remain undetermined these eighteen ORFs may provide a good starting point for an extensive investigation into a detailed understanding of the SSV life cycle at a molecular level. Again sequence analysis did not suggest a role for most of the predicted proteins from the SSVK and SSVRH genomes. Hence, no preference could be established for working with protein from one genome or the other. Therefore, we also decided to concentrate our cloning and structural efforts on this shorter list of 18

conserved proteins from SSVRH with the idea that the SSVRH gene product might complement our work on SSV1. These are the underlying basis for the work described in this thesis. In particular, this work details structural annotation based function for two different proteins SSV1 F112 expressed with an N-terminal His-tag and SSVRH D212. Collectively this body of work provides additional insights into the increasingly recognizable proteomescapes of the *fuselloviridae*, further advancing SSV1 and SSVRH as important model systems for the study of archaeal viruses.

Table A-1. General SSVRH ORF Information								
ORF #	**ORF name**	**NT seq. in genome**	**MW (Da)[a]**	**MW+6XHis (Da)[a]**	**Theoretical pI + His tag [a]**	**% Met (with His-tag, minus N-term Met.)**	**Predicted Function [b]**	**Homolog in SSV1**
1	* A83 (ttg)	34-285	9390.1	10230.1	6	3, 3.4 %		A82
2	* C78	282-518	9414	10254	11.73	1, 1.2%		C84
3	* A93	298-579	10824.8	11664.83	8.94	4, 4%		A92
4	* C277 (gtg)	580-1409	31668	32508	9.9	6, 2.1%		B277
5	* C154	1410-1874	17591.8	18431.76	5.69	3, 1.9%	coiled coil protein	A153
6	* A247 (gtg)	1846-2589	29044.7	29884.74	9.28	8, 3.2%	ATPase invovled in replication intiation	B251
7	A148	2737-3183	16692.3	17532.31	8.85	2, 1.3%		-
8	* D355 (gtg)	4269-3202	41727.6	42567.6	9.43	6, 1.7%	integrase	D355
9	* E23 @	4310-4239	-	-	-	-		-
10	* D212 (gtg)	4932-4291	25055.5	25895.48	8.98	0, 0	Possible nuclease	D244
11	E152	5585-5127	17151.9	17991.89	8.99	2, 1.3%		-
12	D110	5898-5566	12806.3	13646.28	10.87	2, 1.7%		-
13	E150	6356-5904	17843.2	18683.24	5.36	1, 0.6%		-
14	* F61	5125-4940	7124.14	7964.14	6.71	1, 1.5%	Repressor Of Primer adaptor like protein	D63
15	F62 (ttg) +	6538-6353	7175.3	8015.3	8.98	2, 3%		-
16	* F69 (ctg) @	-	-	-	-	-		-
17	* E73 (ttg)	6989-6768	8607	9447	8.91	2, 2.5%	RHH transcriptional regulator	E51
18	* A102a	7111-7419	12038.6	12878.58	6.24	1, 0.9%		22% seq identity to A100 by pairwise p-p balst
19	D57	7197-7024	6674.57	7514.57	8.05	1, 1.6%		-
20	C49 (ttg)	7416-7565	6092.6	6932.6	5.56	3, 5.5%		-

Table A-1 Continued General SSVRH ORF Information								
ORF #	**ORF name**	**NT seq. in genome**	**MW (Da)[a]**	**MW+6XHis (Da)[a]**	**Theoretical pI + His tag [a]**	**% Met (with His-tag, minus N-term Met.)**	**Predicted Function [b]**	**Homolog in SSV1**
21	B50 (gtg)	7562-7714	6233.5	7073.5	4.82	0,0		-
22	A158	7711-8187	18756.4	19596.35	8.47	3, 1.8%		-
23	C59	8184-8363	6755.56	7595.56	5.29	1, 1.5%		-
24	* B64	8360-8554	7665.91	8505.91	9.68	2, 2.9%	RHH transcriptional regulator	C80
25	* B79	8555-8794	9614.43	10454.43	9.68	3, 3.5%	Zn finger protein	A79
26	* A102b (gtg)	8797-9105	11887	12727.04	9.37	3, 2.8%		C102a
27	C287 (gtg)	9104-9965	33692.6	34532.6	6.94	9, 3.1%		-
28	B85 (gtg)	9929-10186	9899.45	10739.45	8.82	2, 2.2%		-
29	* C150 (gtg)	10101-10553	17218.9	18058.92	9.53	2, 1.3%	two C2H2 Zn finger	B129
30	C113 (gtg) +	10557-10895	13202.7	14042.7	6.38	5, 4.2%		-
31	* C247	10932-11675	26424	27264.04	7.22	3, 1.2%		21% seq identity to A291 by pairwise allignment
32	B74	11861-12085	8461.94	9301.94	9.82	3, 3.8%		-
33	C82 (ctg)	12183-12431	9391.62	10231.62	8.33	2, 2.3%		-
34	* B812	12446-14884	88416.2	89256.22	7.76	5, 0.6%		C792
35	* A79	14881-15120	8211.91	9051.91	7.2	4, 4.7%		B78
36	* B170 (gtg)	15101-15613	18341.8	19181.83	8.29	2, 1.1%		C166
37	* A113	15610-15951	12958.9	13798.9	9.59	1, 0.8%		B115
38	* VP1 (A89)	15902-16171	9453.43	10293.43	10.16	1, 1.1%	Capsid protein	VP1
39	* VP3 (C96)	16183-16473	10565.2	11405.24	6.96	1, 1.0%	Capsid protein	VP3
[a] Calculated by ProtParam (ExPASy tool Server).								
() Alternative start codon. @ No longer considered ORFs.								
* ORFs common between SSVRH and SSV1.								
[b] Function predictions based on literature and from structural studies.								

Table A-2. General Information for the Test Set of 11 ORFs Targeted for N-terminal 6X His Tag Constructs.							
ORF #	**ORF name**	**NT seq. in genome**	**MW (Da)[a]**	**MW+6X His (Da)[a]**	**Theoretical pI +His tag[a]**	**% Met (with His-tag, minus N-term.)**	**Predicted Function [b]**
1	B251	191-943	29514.4	30355.2	9.51	6, 2.8%	ATPase invovled in replication intiation
2	D335	1968-964	38841.9	39682.7	9.47	3, 0.89%	integrase
3	E54	2147-1986	6020.3	6861.1	11.71	2, 3.7%	-
4	F92	2398-2123	10759.7	11600.5	5.36	1, 1.1%	-
5	D244	3129-2398	28497.7	29338.5	8.55	3, 1.2%	Possible nuclease
6	E178	3575-3042	20378.5	21219.3	6.07	0, 0	-
7	F93	3862-3584	10947.8	11788.6	9.37	1, 1.1%	winged helix-turn-helix DNA binding protein
8	E51	4025-3873	5944	6784.8	8.31	1, 2.0%	RHH transcriptional regulator
9	E96	4316-4029	11157.6	11998.4	4.76	1, 1.0%	-
10	D63	4509-4321	7450.8	8291.6	6.15	2, 3.2%	Repressor Of Primer adaptor like protein
11	F112	4960-4625	13186	14026.8	6.72	2, 1.8%	winged helix-turn-helix DNA binding protein
[a] Calculated by ProtParam (ExPASy server).							
[b]Function predictions based on literature and from structural studies.							

CHAPTER 2

CLONING, EXPRESSION, PURIFICATION, AND CRYSTALLIZATION OF THE SSV1 AND SSVRH GENE PRODUCTS

Introduction

In our effort to functionally annotate the proteomes of SSV1 and SSVRH we initiated a systematic structure determination study of the predicted gene products from both the genomes. This chapter first provides a brief overview of the various steps that were carried out for cloning and expression trials for a large number of SSV1 and SSVRH constructs. Preliminary purification and crystallization studies that were carried out for the well behaved proteins are also described in this chapter. In particular, however, this chapter focuses on the five proteins that are the major focus of this thesis, specifically SSV1 F112, SSVRH D212, SSVRH F61, SSVRH E73 and SSVRH B64 and details the methods associated with protein expression, purification, crystallization, crystallographic data collection, structure determination, biochemical and biophysical characterization, and bioinformatics used in their study.

Cloning

Gateway Cloning Technology (Invitrogen Corporation) is a site-specific recombination system which facilitates the rapid expression of tagged proteins in various expression systems. Polymerase Chain Reaction (PCR) is used to amplify the gene of interest sandwiched in between additional sequences (Figure 2-1) that assists in

subcloning the DNA fragment into desired vectors using a ligase free site-specific recombination. This PCR product is then inserted into a destination vector via the BP and LR site-specific recombination reactions yielding an expression vector (Figure 2-2).

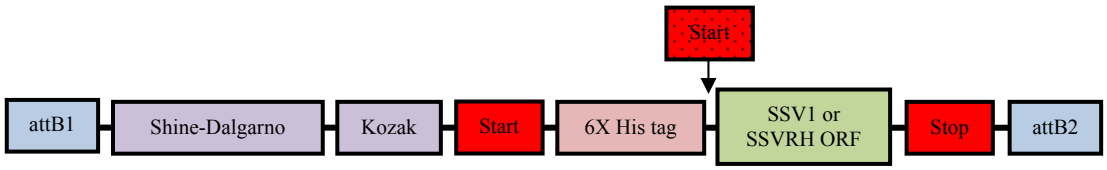


Figure 2-1. Schematic representation of the desired PCR product.
The different features of the PCR product to help in cloning, expression and purification are color coded. A second start codon at the position shown by red spotted box may also be present in the primer.

PCR Design

The first step towards the construction of expression clones for the proteins of interest was to amplify the desired gene fragments from the SSV1 and SSVRH genomes using PCR. In the case of SSV1, either the purified SSV1 DNA [22, 23] or pKMSD48 [13] a pBluescript based SSV1 vector was used, while for SSVRH purified viral DNA [22] was used as template for PCR amplification reactions. All DNA samples used were generously provided by Dr. Mark Young's laboratory at Montana State University. Nested PCR employs two successive amplification reactions with two sets of primers. The first set of primers were designated as "Universal primers" (1-3 and 1-4) and were designed to incorporate the attB1 site, a 6X His tag, and the Shine-Dalgarno and Kozak sequences on the 5'(primer 1-3) end of the gene while the attB2 and stop codon were placed at the 3' prime end (primer 1-4). The attB1 and attB2 sequences are specific recombination sites required for the correct integration of the PCR product into the desired vector. The Shine-Dalgarno and Kozak are signal sequence which direct the initial binding of the mRNA to the ribosome and in turn initiate translation in prokaryotes

and eukaryotes respectively. The 6X His tag which can be placed on either the N- or the C-terminal of the gene of interest, ensures rapid and efficient purification of soluble expressed proteins using immobilized metal-affinity chromatography (IMAC).

UNIVERSAL PRIMERS (UP):

1-3 = ATTB1 + SD + Kozak + MET + 6X HIS

1-3 5' GGGG ACAAGTTTGTACAAAAAAGCAGGCT TCGAAGGAGATAGA ACC ATG CATCACCATCACCATCAC

1-4 = ATTB2 + STOP

1-4 5' GGGG ACCACTTTGTACAAGAAAGCTGGGTC CTA

The second set of primers, designated as “Internal primers” (ORF # -1 and ORF# -2) consist of two parts, a 5’ portion that overlaps with the sequence from the universal primers and a 3’ portion that is gene specific. The internal primers were labelled according to the ORF numbers with the -1 or -2 indicating the 5’ and 3’ primer respectively as shown below for ORF#1 from SSVRH (A83)

INTERNAL PRIMERS:

1-1 = overlap with 1-3 UP + overlap with the N-terminal of ORF A83

1-1 5’ ATGCATCACCATCACCATCAC ATGAGTGCGTTAGGCGATGT

1-2 = overlap with 1-4 UP + RC of the C-terminal end of ORF A83

1-2 5' GTACAAGAAAGCTGGGTCCTA TATTCTTTCCTCCTTAAGCC

PCR was carried out as described in the Gateway manual [57] with KOD HiFi polymerase from Novagen, a high fidelity and high processivity polymerase. The only exceptions were in the variation of annealing temperature for the different primers in the

PCR reaction which was adjusted depending on the melting temperature (Tm) of each set of primers. All primers used in this study were purchased from Sigma-Genosys and their sequences are presented in Table A-3 and A-4 for SSVRH and SSV1 respectively. The first round of PCR involved a 10-cycle PCR reaction using the internal primers with the relevant viral genome as template. While in the second round of PCR, a 25-cycle reaction carried out with the universal primers and the PCR product from the first reaction as the template. For a more detailed description please refer to the Gateway manual [57] and Dr. Larson's PhD thesis [58].

Construction of the Entry and Expression Clones

The PCR products obtained as discussed above were then gel-purified using a QIAquick gel extraction kit (Qiagen, catalog # 28704). The purified DNA fragments which contained the genes of interest were then inserted into *E.coli* donor vector pDONR201 (Invitrogen) to create entry clones via the BP reaction. Site-specific recombination between the attB1 and attB2 sites present on the PCR products and the attP1 and attP2 sites on the donor vector were performed as described in the Gateway manual [57] driven by the BP clonase enzyme. For the BP reaction, the PCR fragments were mixed with the donor vector and the clonase enzyme at $1/4^{th}$ the recommended volume and incubated overnight at room temperature (25 ºC). 1 µl of Proteinase K was then added to mixture and the reactions were incubated for 1-2 hours at 37 ºC. The products were then either, frozen and stored at -20 ºC (for up to 1 week) or were immediately transformed into Dh5α competent cells (Invitrogen).

Table A-3. Primer Sequences Used for Cloning the SSVRH ORFs.

Internal ORF ID #	ORF name	NT seq in genome	Primer Name	Primer Sequence (5'-3'), the space is at the interface between universal primer overlap and ORF overlap
1	A83	34-285	1-1	ATGCATCACCATCACCATCAC ATGAGTGCGTTAGGCGATGT
			1-2	GTACAAGAAAGCTGGGTC CTA TATTCTTTCCTCCTTAAGCC
2	C78	282-518	2-1	ATGCATCACCATCACCATCAC ATGAAGGTGATACACAATGA
			2-2	GTACAAGAAAGCTGGGTC CTA AGGTATTCCCAAAAACTTTC
3	A93	298-579	3-1	ATGCATCACCATCACCATCAC ATGATATACGTTTCAGAAAA
			3-2	GTACAAGAAAGCTGGGTC CTA CTTTTGGATCACCATTCTCA
4	C277	580-1409	4-1	ATGCATCACCATCACCATCAC GTGAGTGATGGGAAACTACT
			4-2	GTACAAGAAAGCTGGGTC CTA GCTCACCCCTTTTGAGGAGT
5	C154	1410-1874	5-1	ATGCATCACCATCACCATCAC ATGGCTAAGAAGACGGAAGT
			5-2	GTACAAGAAAGCTGGGTC CTA ACCTTCTTCTTTGGCATCCA
6	A247	1846-2589	6-1	ATGCATCACCATCACCATCAC ATGCCAAAGAAGAAGGTTAA
			6-2	GTACAAGAAAGCTGGGTC CTA GGCACTTGGTACTTTTAACG
8	D355	4269-3202	8-1	ATGCATCACCATCACCATCAC ATGGGTAATAAAGTTTTCAC
			8-2	GTACAAGAAAGCTGGGTC CTA CAGTAAAACGTATTTTGTCA
10	D212	4932-4291	10-1	ATGCATCACCATCACCATCAC GTGACCGAAACTGATTTTAA
			10-2	GTACAAGAAAGCTGGGTC CTA CTTATTACTCCTAGCTAAAT
14	F61	5125-4940	14-1	ATGCATCACCATCACCATCAC ATGGAAGACTTAGATTCTTT
			14-2	GTACAAGAAAGCTGGGTC CTA CCTCAGCACCTCTTGTTTCT
17	E73	6989-6768	17-1	ATGCATCACCATCACCATCAC TTGGTTGAAAGTAAAAAGAT
			17-2	GTACAAGAAAGCTGGGTC CTA TTTTGCGGGAAAAAACCGCT
18	A102a	7111-7419	18-1	ATGCATCACCATCACCATCAC ATGGTTGAACCCCAATACCA
			18-2	GTACAAGAAAGCTGGGTC CTA ACTTACACGCGAAAGCAGTA
24	B64	8360-8554	24-1	ATGCATCACCATCACCATCAC ATGAAGGCAAGAGTGGAATA
			24-2	GTACAAGAAAGCTGGGTC CTA CTGCACCTCTTCTAGATATT
25	B79	8555-8794	25-1	ATGCATCACCATCACCATCAC ATGTATAAATGCCCACTCTG
			25-2	GTACAAGAAAGCTGGGTC CTA TTCCACCTCTAATTTGCGTT
26	A102b	8797-9105	26-1	ATGCATCACCATCACCATCAC ATGAACGAAACTTTACTAGA
			26-2	GTACAAGAAAGCTGGGTC CTA CCTCACCTTCTTACGCCTAT
29	C150	10101-10553	29-1	ATGCATCACCATCACCATCAC ATGAGGGACTCAGGCAGTGA
			29-2	GTACAAGAAAGCTGGGTC CTA ACTTACGCAAATATTATGTT
31	C247	10932-11675	31-1	ATGCATCACCATCACCATCAC ATGAGGAAGACCCTTCTAGC
			31-2	GTACAAGAAAGCTGGGTC CTA CCTGATTCGGCCATTCATAG
34	B812	12446-14884	34-1	ATGCATCACCATCACCATCAC ATGAAGTGGGGACTATCTTT
			34-2	GTACAAGAAAGCTGGGTC CTA TTCCTCCTCACTCCCCATAA
35	A79	14881-15120	35-1	ATGCATCACCATCACCATCAC ATGACAGACGCAATTAGTTT
			35-2	GTACAAGAAAGCTGGGTC CTA GTACCCATCTTTCCCACCTA
36	B170	15101-15613	36-1	ATGCATCACCATCACCATCAC GTGGGAAAGATGGGTACTAA
			36-2	GTACAAGAAAGCTGGGTC CTA TGGCCCGTAACCCCTTATCA
37	A113	15610-15951	37-1	ATGCATCACCATCACCATCAC ATGACTTATAACGCGAATGA
			37-2	GTACAAGAAAGCTGGGTC CTA GCCCTTCTTAGTTTCTTTGG
38	VP1 (A89)	15902-16171	38-1	ATGCATCACCATCACCATCAC ATGAGGGCAGTCGCGGTATT
			38-2	GTACAAGAAAGCTGGGTC CTA ATCTTTGTAGATTTTATACG
39	VP3 (C96)	16183-16473	39-1	ATGCATCACCATCACCATCAC ATGGAGTCAAATGTTAAGGC
			39-2	GTACAAGAAAGCTGGGTC CTA CTCACTCTTGTACAATCTGT

Table A-4. Primer Sequences Used for Cloning the SSV1 ORFs with N-terminal 6xHis tag.

Internal ORF ID #	ORF name	NT seq in genome	Primer Name	Primer Sequence (5'-3'), the space is at the interface between universal primer overlap and ORF overlap
1	B251	191-943	1-1	CATCACCATCACCATCAC GTAAGGAACATGAAGAT
			1-2	CAAGAAAGCTGGGTC CTA TCTTGTTACGTCGTCAA
2	D335	1968-964	2-1	CATCACCATCACCATCAC ACGAAAGATAAGACGCG
			2-2	CAAGAAAGCTGGGTCCTA GACCCCTTTTAGCCATT
3	E54	2147-1986	3-1	CATCACCATCACCATCAC GCTAGAATACAAGGTGG
			3-2	CAAGAAAGCTGGGTCCTA TGGTTGCCAAAACGGGC
4	F92	2398-2123	4-1	CATCACCATCACCATCAC ATATTCAACTCAGAGGA
			4-2	CAAGAAAGCTGGGTCCTA TTTGTCCACCTTGTATT
5	D244	3129-2398	5-1	CATCACCATCACCATCAC GATTGTAATAGGAATTG
			5-2	CAAGAAAGCTGGGTCCTA TTTTCCGCCCCTCGCTA
6	E178	3170-3042	6-1	CATCACCATCACCATCAC GATCAAAGAAGAGTTGA
			6-2	CAAGAAAGCTGGGTCCTA CCATATTGAAACGTGGG
7	F93	6819-6639	7-1	CATCACCATCACCATCAC AAAAGTACACCATTCTT
			7-2	CAAGAAAGCTGGGTCCTA CCTCTTTTTGATTTGTT
8	E51	6982-7135	8-1	CATCACCATCACCATCAC ATAGAGAAGAGGAAAGA
			8-2	CAAGAAAGCTGGGTCCTA CTCCTCAGCCTCTAATA
9	E96	7273-6986	9-1	CATCACCATCACCATCAC ATGATGGTGCAAGTTCC
			9-2	CAAGAAAGCTGGGTCCTA GGTCACCTTGTTTATTG
10	D63	7466-7286	10-1	CATCACCATCACCATCAC AGTAAAGAAGTCTTAGA
			10-2	CAAGAAAGCTGGGTCCTA GCTCACCTTAATGAGCT
11	F112	7917-7582	11-1	CATCACCATCACCATCAC GCACAAACTCTAAATTC
			11-2	CAAGAAAGCTGGGTCCTA CTGCTTTGCCTTTATAA

The transformed cells were plated on Luria Broth (LB) plates supplemented with 50 µg/ml kanamycin. The pDONR201 vector used for the BP reactions contains a ccdB killer gene along with kanamycin resistance. The recombination reaction removes the ccdB gene and inserts the gene of interest in its place into the donor vector to create the entry clone. The CcdB protein selectively targets the E.coli DNA gyrase and kills the cells. Thus only those cells with vectors in which the ccdB gene has been replaced by recombination will survive. For each clone, three colonies were selected and colony PCRs was used to confirm the presence of the gene of interest using the gene specific internal primers. Only clones that gave positive results with the correct size by colony PCR were sent off for sequencing to ensure a mutation free clone. ABI BigDye Terminator cycle sequencing was performed at the DNA Facility, Iowa State University

(Ames, Iowa) for all the entry clones and the results were analyzed using DNAStar (Lasergene Inc, Madison, WI).

The expression clones were generated through subsequent recombination reaction performed between the attL-containing entry clones and attR-containing destination vectors. Different destination vectors can be selected to carry out the LR recombination reaction based on the choice of the expression system such as *E.coli*, yeast or mammalian cells. As the affinity tag was already designed into the PCR construct, pDest14 a non fusion destination vector was used in this study. The LR reactions were also performed according to the Gateway manual instructions [57] with certain modifications similar to the BP reaction. The reaction products were then transformed into DH5α competent cells (Invitrogen) and plated on selection plates supplemented with 100 µg/ml ampicillin. Two to three colonies were picked and colony PCR with gene specific internal primers was carried out to confirm the presence of the desired gene sequence in the expression clone. The BP and LR clones with the proper insert sequence were grown overnight in LB media containing the correct anitibiotics and the clones were purified and extracted in either Tris-EDTA buffer or water by the QIAprep spin miniprep kit (Qiagen). Expression clones for 5 ORFs from SSV1 and 20 ORFs from SSVRH have been constructed successfully and are stored at -20 ºC (Table A-5 and A-6).

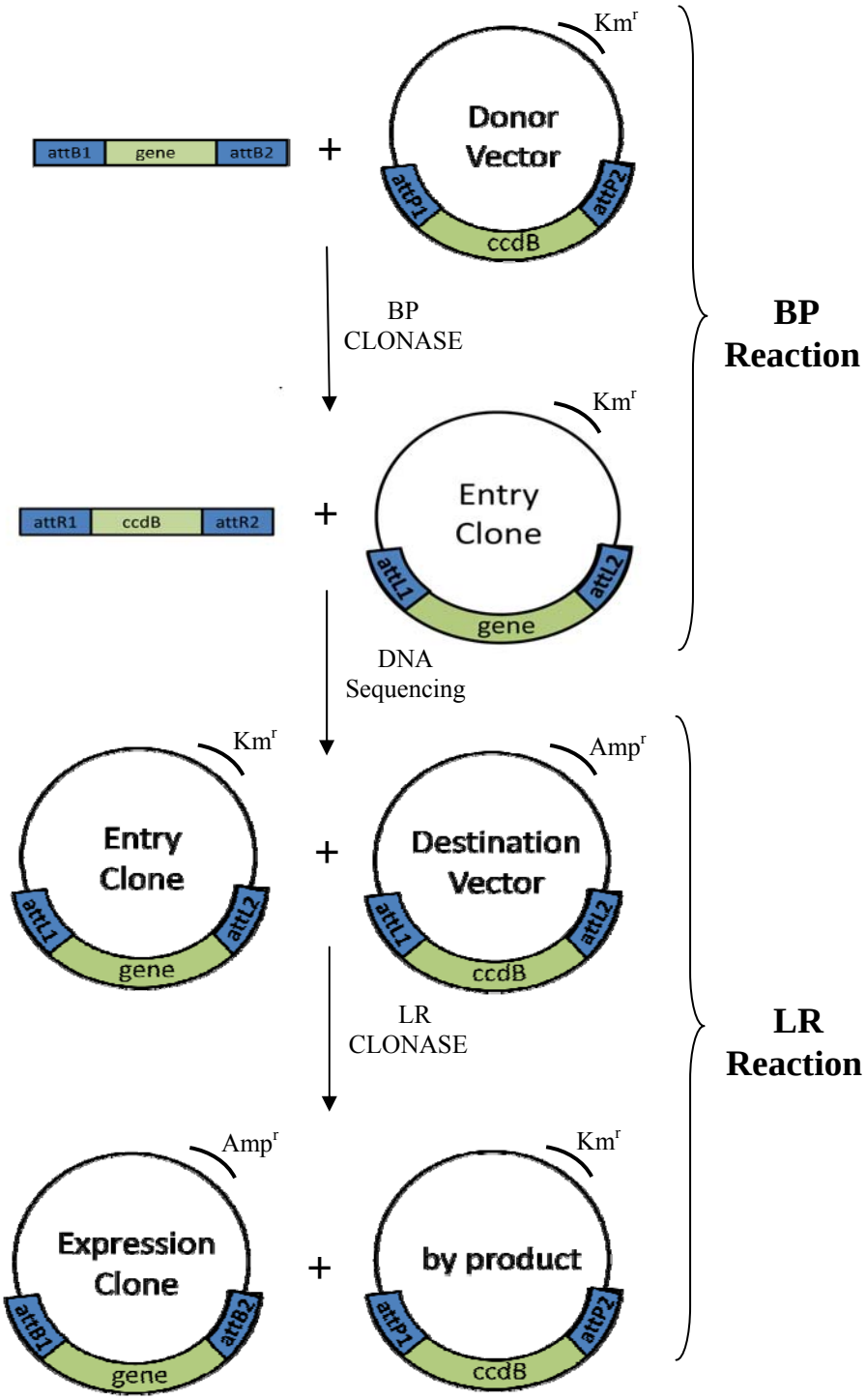


Figure 2-2. Schematics for the Gateway Cloning site-specific recombination reactions.

Expression and Purification

Small scale expression trials for 25 of the successfully cloned ORFs from both SSV1 and SSVRH were carried out in a bacterial expression system. For protein expression, BL21-CodonPlus-(DE3)-RIL cells (Stratagene) were transformed with the pDEST14 vector containing the gene of interest. A single colony of each clone was used to inoculate 5 ml of LB media with 100 μg/ml ampicillin (AMP) and 34 μg/ml chloramphenicol (CAM) and cultured for 5-6 hours at 37 ºC. These cultures were then used to inoculate 25 ml overnight LB+AMP+CAM cultures. These inoculums were further expanded to a 1 litre culture where cells were grown to an OD_{600} (Optical Density at 600 nm) of about 0.6-0.8. Protein expression was then induced using 1 mM IPTG (isopropyl-ß-D-thiogalactopyranoside) for 4-6 hours. The cells were pelleted by centrifugation at 6,000xg and were stored at -20 ^{o}C. For the well-behaved proteins the expressions were scaled up to batches of 6 x 1-L each in shaker flasks.

The 6X His tag on our proteins facilitated their easy purification using the immobilized metal-affinity chromatography (IMAC) system. Columns packed with HIS-Select™ Nickel-NTA Affinity media (Sigma-Aldrich) were used and protein purification was tried with both high salt and low salt buffer conditions. Cell pellets were thawed and resuspended at 5ml/g of wet cell paste in either high salt lysis buffer (10 mM Tris, pH 8.0, 20 mM $NaH_2PO_4.H_2O$, 1 M NaCl and 5 mM imidazole) or low salt lysis buffer (10 mM Tris, pH 8.0, 300 M NaCl and 5 mM imidazole) along with freshly prepared phenyl methyl sulfonyl fluoride (0.1 mM PMSF). The resuspended cells were then passed through a French Press (American Instrument Co., Inc., Silver Springs, MD) at least

twice to ensure complete lysis of the cells. The cell lysates were incubated at 65 °C for 20 min to denature contaminating *E. coli* proteins, and clarified by centrifugation at 22,000 ×*g* for 20 min. The supernatants were applied to a column containing a 3–5-ml bed volume of HIS-Select™ Nickel Affinity Gel (Sigma-Aldrich). The column was washed with 10 column volumes of high salt wash buffer (10 mM Tris, pH 8.0, 20 mM $NaH_2PO_4.H_2O$, 1 M NaCl, and 10 mM imidazole) or low salt wash buffer (10 mM Tris, pH 8.0, 300 mM NaCl, and 10 mM imidazole). The final elution of the proteins were also done either using high salt elution buffer (10 mM Tris, pH 8.0, 20 mM $NaH_2PO_4.H_2O$, 300 mM NaCl and 200 mM imidazole) or low salt elution buffer (10 mM Tris, pH 8.0, 50 mM NaCl, and 10 mM imidazole). The purified proteins were analyzed by SDS-Poly Acrylamide Gel Electrophoresis (SDS-PAGE) to determine their levels of expression and purity. Other chromatographic techniques such as size exclusion chromatography (SEC) or ion-exchange chromatography (IEC) were also used as required and are described below for individual proteins. The final step of purification before the proteins were subjected to crystallization trials always involved a buffer exchange step using SEC, so as to get the protein into a minimal but well defined buffer. Protein concentration was determined by Bradford assay using bovine serum albumin (BSA) as a standard. The purity and molecular weight of the proteins were also, at times, confirmed by MALDI-TOF mass spectrometry performed with a Bruker Biflex III with a choice of different matrices. Dynamic light scattering (DLS) measurements to confirm the oligomeric state of some of these well-behaved target proteins was also done on a Brookhaven Instruments ZetaPals (phase analysis light scattering) particle size/ζ potential analyzer.

The purified protein in minimal buffer was concentrated to ~10 mg/ml using amicon spin concentrators (Amicon ultra centrifugal unit; Millipore) of the appropriate molecular weight cutoff. A pH versus salt concentration gradient coarse screen was setup for each of the protein to identify the buffer condition where the protein was the best behaved. If another buffer condition was identified where the protein was found to be more soluble then the final buffer exchange was done into this new buffer before the final concentration step. Three of the six ORFs from SSV1 and nine of the twenty ORFs from SSVRH gave soluble protein expression (Table A-5 and A-6). The well behaved proteins were carried onto the next step through crystallization trials while those that did not behave well were left behind.

Crystallization

Common techniques used to grow protein crystals are by hanging drop vapor diffusion, sitting drop vapor diffusion, dialysis buttons, or microbatch techniques. In the hanging drop vapor diffusion method a few microliters (2µl in our study) of protein solution is mixed with an equal amount of reservoir solution containing the precipitants. A drop of this mixture is put on a glass slide which covers the reservoir and it is left to equilibrate. As the precipitant concentration in the protein- precipitant mixture in the drop is less concentrated than in the reservoir solution, water vapor diffuses from the drop into the reservoir. As a result both the protein and precipitant concentration in the drop slowly increases to a level optimal for crystal formation.

In this study, all proteins that gave soluble protein expressions were first screened for crystallization conditions using the microbatch under oil technique at the Hauptman-

Woodward Institute (HWI), New York, a high throughput crystallization facility. From the 1536 conditions screened at HWI the conditions which gave crystals were then repeated at MSU to try and reproduce the crystals. Because the HWI uses a microbatch method to grow crystals, while here in our lab the hanging- or sitting- drop vapor diffusion technique is used the conditions were not directly transferable between the two techniques. To overcome this problem, we frequently needed to use reduced precipitant concentrations often by the factor of 2 compared to the microbatch method in order to duplicate the crystals seen at HWI by microbatch under oil. Once the initial crystallization conditions were identified from the microbatch method, a finer grid screen was setup around that condition with small changes in the pH or salt or precipitating agent concentrations. Addition of various "additives" to the hanging drop can also affect the quality of the crystals formed. Additives such as detergents, cations, anions, reducing agents, chaotropic agents, nucleotides and other such potential ligands were also used for the optimization of crystallization condition (Hampton Additive Screen). Temperature variations from 18ºC to 27ºC were also screened for identification of the optimal growth temperatures. Results from the various crystallization trails on the different proteins from SSV1 and SSVRH is listed in Tables A-5 and A-6. Proteins where diffraction quality crystals were obtained which resulted in a structure are discussed in greater details.

Table A-5. The Progress Towards Structural Characterization of SSVRH.

Internal ORF ID #	ORF name	Cloned and Sequenced	Vector names (entry, expression clones)	Soluble Expression	Crystallization trials	Crystals	Diffraction	Structure	Polyclonal Antibodies[c]	Comments
1	A83	yes	BP clone-A83 LR clone-A83	no	no	no	no	no	no	
2	C78	yes	BP clone-C78 LR clone-C78	no	no	no	no	no	no	
3	A93	yes	BP clone-A93 LR clone-A93	no	no	no	no	no	no	
4	C277	yes	BP clone-C277 LR clone-C277	no	no	no	no	no	no	
5	C154	yes	BP clone-C154 LR clone-C154	yes (HS)	no	no	no	no	no	on a gel, the band size smaller than expected
6	A247	yes	BP clone-C247 LR clone-C247	no	no	no	no	no	no	
8	D355	no	-	no	no	no	no	no	no	
10	D212	yes	BP clone-D212 LR clone-D212	yes (HS)	HTS	yes	yes	yes (2W8M)	yes	
14	F61	yes	BP clone-F61 LR clone-F61	yes (LS)	HTS	yes	yes	yes	yes	
17	E73	yes	BP clone-E73 LR clone-E73	yes (LS)	HTS	yes	no	no	no	small crystals from HTS screen, not able to reproduce at home
18	A102a	yes	BP clone-A102a LR clone-A102a	yes (LS)	no	no	no	no	no	low expression for protein
24	B64	yes	BP clone-B64 LR clone-B64	yes (HS)	HTS	yes	no	no	no	small crystals from HTS screen
25	B79	yes	BP clone-B79 LR clone-B79	yes (LS)	no	no	no	no	no	on a gel, the band size smaller than expected
26	A102b	yes	BP clone-A102b LR clone-A102b	no	no	no	no	no	no	
29	C150	yes	BP clone-C150 LR clone-C150	yes (HS)	HTS	yes	no	no	no	crystals from HTS screen, not able to reproduce at home
31	C247	no	-	no	no	no	no	no	no	base insertion in the BPclone
34	B812	no	-	no	no	no	no	no	no	
35	A79	yes	BP clone-A79 LR clone-A79	no	no	no	no	no	no	
36	B170	yes	BP clone-B170 LR clone-B170	no	no	no	no	no	no	
37	A113	yes	BP clone-A113 LR clone-A113	yes (HS)	no	no	no	no	no	proteolysis, 2 bands seen after purification
38	VP1 (A89)	yes	BP clone-VP1 LR clone-VP1	no	no	no	no	no	no	
39	VP3 (C96)	yes	BP clone-VP3 LR clone-VP3	no	no	no	no	no	no	

Table A-6. The Progress Towards Structural Characterization of N-terminal 6X His Tag SSV1 proteins.										
Internal ORF ID #	**ORF name**	**Cloned and Sequenced[b]**	**Vector names (entry, expression clones)**	**Soluble Expression**	**Crystallization trials**	**Crystals**	**Diffraction**	**Structure**	**Polyclonal Antibodies[c]**	**Comments**
1	B251	no	-	no	no	no	no	no	no	
2	D335	no	-	no	no	no	no	no	no	
3	E54	yes	BP clone-E54 LR clone-E54	no	no	no	no	no	no	
4	F92	no	-	no	no	no	no	no	no	
5	D244	no	-	yes	no	no	no	no	no	
6	E178	yes	BP clone-E178 LR clone-E178	yes	no	no	no	no	no	impurities, used SSV1 DNA for amplification as gene interrupted in pKMSD48 genome
7	F93	yes	BP clone-F93 LR clone-F93	no	no	no	no	no	no	crsytal structure (1TBX) from C-terminal 6X his tag protein, Paul's work
8	E51	no	-	no	no	no	no	no	no	
9	E96	yes	BP clone-E96 LR clone-E96	yes	yes	yes	yes	yes	no	extra ATG on the N-terminus, Brians solved the structure
10	D63	yes	BP clone-D63 LR clone-D63	yes	no	no	no	no	no	crsytal structure (1SKV) from C-terminal 6X his tag protein, Paul's work
11	F112	yes	BP clone-F112 LR clone-F112	yes	yes	yes	yes	yes (2VQC)	yes	

Materials and Methods for the Structure Determination of F112

Cloning

ORF F112 was amplified using a nested Polymerase Chain Reaction (PCR) as described above. The internal forward and reverse primers were 5'-ATGCATCACCATCACCATCACATGAGTGCG TTAGGCGATGT-3' and 5'-GTACAAGAAAGCTGGGTCCTATATTCTTTCCTCCTTAAGC C-3', while the external forward and reverse primers were 5'-GGGGACAAGTTTGTACAAAA AAGCAGGCTTCGAAGGAGATAGAACCATGCATCACCATCACCATCAC-3' and 5'-GGG GACCACTTTGTACAAGAAAGCTGGGT CCTA-3', respectively. The entry clone (BPclone-F112) was then transferred into pDEST14 (Invitrogen) with a second site-specific recombination reaction, yielding the expression vector pDest14_N-F112.

Expression and Purification

For native F112 protein expression, BL21(DE3)-RIL *E.coli* cells (Stratagene) were transformed with pDest14_N-F112. Typically a single colony was used to inoculate 5 ml of Luria Broth (LB) + 100 μg/ml ampicillin with growth for 5-6 hours, followed by expansion to a 25 ml overnight culture. The overnight culture was then scaled up to 1 L and cells were grown to an OD_{600} between 0.6 and 0.8. Protein expression was induced using 1 mM IPTG (isopropyl-ß-D-thiogalactopyranoside) for 4-6 hours. Cells were harvested by centrifugation at 3,000 ×g (IEC PR-7000M) for 20 minutes and the pellets were stored at -20 ^{o}C.

For F112 purification, cell pellets were thawed and resuspended at 5 ml/gram of cell pellet in lysis buffer (10 mM Tris, pH 8.0, 20 mM $NaH_2PO_4.H_2O$, 1 M NaCl and 5

mM imidazole) along with freshly prepared phenyl methyl sulfonyl fluoride (0.1 mM PMSF). The cell suspension was vortexed vigorously to homogenize the mixture. The homogenate was then passed through a French Press (American Instrument Co., Inc., Silver Springs, MD) to ensure complete lysis of the cells. The cell lysate was incubated at 65 °C for 20 minutes to denature contaminating *E. coli* proteins, and clarified by centrifugation at 22,000 ×g for 20 minutes. The supernatant was applied to a column containing a 3-5 ml bed volume of HIS-Select™ Nickel Affinity Gel (Sigma-Aldrich). The column was washed with 10 column volumes of wash buffer (10 mM Tris, pH 8.0, 20 mM $NaH_2PO_4.H_2O$, 1 M NaCl, 10 mM imidazole) and F112 was eluted in 10 mM Tris, pH 8.0, 20 mM $NaH_2PO_4.H_2O$, 300 mM NaCl and 200 mM imidazole. F112 was then applied to a calibrated Superdex S-75 (Amersham-Pharmacia) column equilibrated with 10 mM Tris (pH 8.0) and 50 mM NaCl. Protein concentrations were determined by Bradford assay [59] using Protein Assay Reagent (Bio-Rad) and BSA as a standard. The purity and molecular weight of F112 were confirmed by SDS-PAGE and MALDI-TOF mass spectrometry. MALDI-TOF MS was performed with a Bruker Biflex III in the matrix alpha-cyano-4-hydroxycinnamic acid. Dynamic light scattering (DLS) measurements to determine the oligomeric state of F112 were made on a Brookhaven Instruments ZetaPals (phase analysis light scattering) particle size/ζ potential analyzer. DLS was measured at 90^o using a 661 nm diode laser and the correlation functions were fit using a constrained least-squares analysis as described [60].

For production of selenomethionine-labelled protein (SeMet-F112), the methionine auxotroph B834 (DE3) *E.coli* (Novagen) was transformed with pDest14_N-F112. A single colony was used to inoculate a 5 ml overnight LB/AMP culture. The

culture was centrifuged at 500 ×g to pellet the cells, and the pellet washed twice with 0.5 ml of Milli-Q-H_2O, and resuspended in 1L of SeMet medium [61]. Subsequent expression and purification of SeMet-labelled F112 was identical to that described above for the native protein. The yield of SeMet-labelled F112 was generally 1-1.5 mg per gram of cell pellet.

Crystallization and Data Collection

Purified F112 was concentrated to 10 mg/ml with spin concentrators. The protein was crystallized using hanging drop vapour diffusion. Drops were setup at 24 ^oC using 2 µl of F112 and 2 µl of well solution consisting solely of 0.1 M MES at pH 6.0. Crystals up to 0.2 x 0.07 x 0.01 mm in size were obtained in 4 weeks time. The crystals were transferred to synthetic mother liquor containing 25% glycerol as a cryoprotectant for 30 minutes, and then flash-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 integrated with Mosflm (version 6.2.4) [62], and scaled and reduced in space group R3 to a resolution of 2.3 Å using SCALA [63].

Structure Determination and Refinement

SOLVE [64] was used to determine the positions of the selenium-substructure and to calculate initial phases. Two selenium sites per asymmetric unit were identified using SOLVE. RESOLVE [65] was used for density modification and initial model-building. Iterative model building with O [66] and refinement with REFMAC5 [67, 68] led to the final model. Because residues comprising the N-terminal His-tag were disordered, we

chose to number residues in the model such that they are consistent with the native, non-tagged protein sequence. Amino acids 1-3 at the N-terminal and 73-112 at C-terminal end were not modelled due to the lack of interpretable electron density. TLS parameters [69, 70] were included in the refinement, wherein the F112 monomer was divided into 15 TLS groups (1:4-8, 2:9-12, 3:13-16, 4:17-20, 5:21-25, 6:26-30, 7:31-35, 8:36-39, 9:40-43, 10:44-47, 11:48-51, 12:52-56, 13:57-61, 14:62-66, 15:67-72). The final R factors for the model were 19.9% / 17.3% (R_{free}/R_{cryst}) respectively.

Structural Analysis and Coordinates

Three dimensional structural homology searches were carried out using the DALI [71] and VAST [72] servers. Superposition of DP2 upon F112 was done with LSQKAB [67, 73]. Structural figures were generated with PYMOL [74]. Coordinates and structure factors have been deposited in the Protein Data Bank under accession code 2VQC.

Electrophoretic Mobility Shift Assays

Non-specific binding of F112 to DNA was characterized using electrophoretic mobility shift assays (EMSA). A randomly chosen fragment of the SSV1 genome was PCR amplified and used as template in the EMSA studies. This PCR amplified 153 base pair DNA template consisted of bases 5177-5323 of the SSV1 genome. The forward and the reverse primers used for amplification were 5'-GTGATGGTGATGGTGATGATGGGATGTGCAAAATC-3' and 5'- CAAGAAAGCTGGGTCCTAGAACAAATCATTTATTG-3', respectively. The amplified DNA (1μM final concentration when present) was incubated with or without F112 (5μM final concentration), at 65 °C for 30 minutes in 20 mM Bis-Tris pH 6.5, 50

mM KCl, 5% glycerol and 1 mM EDTA. Samples were electrophoresed in a 1.6% agarose gel containing ethidium bromide, using 25 mM histidine and 30 mM MOPS at pH 6.3 as the running buffer, for 40 min at 110 volts. DNA was visualized by UV illumination.

Similar mobility shift assays were also tried with PCR amplified 1000bp DNA fragments encompassing the entire SSV1 genome (Table S-1). For this assay 20nM of the amplified DNA was incubated with 200nM F112 protein at 65°C for 30 minutes in the same buffer mentioned above. The electrophoresis and visualization conditions used were also as described above.

Materials and Methods for Disulfide Induced Protein Stability

Assembly of the Metagenome and Predicted Subcellular Localization

Amino acid sequences for known and hypothetical proteins from 18 hyper-thermophilic viruses were used in the analysis. The full names of the viruses, their abbreviations and associated accession numbers are: Sulfolobus Spindle-shaped Virus 1 (SSV1), XO7234 [19-21]; Sulfolobus Spindle-shaped Virus 2 (SSV2), AY370762) [23]; Sulfolobus Spindle-shaped Virus 3 (SSV3) (K. M. Stedman and Y. Combet-Blanc, personal communication); Sulfolobus Spindle-shaped Virus Ragged Hills (SSVRH), AY388628 [45]; *Sulfolobus* Spindle-shaped Virus Kamchatka 1 (SSVK1), AY423772 [45]; *Thermoproteus tenax* Spherical Virus 1 (TTSV1), AY722806 [34]; *Thermoproteus tenax* virus (TTV1), X14855 [27, 29]; *Sulfolobus islandicus* Filamentous Virus (SIFV), AF440571 [28]; *Acidianus* Filamentous Virus 1 (AFV1), AJ567472 [31]; *Acidianus* Filamentous Virus 2 (AFV2), AJ854042 [30]; *Sulfolobus islandicus* Rod-shaped Virus 1

(SIRV1), AJ414696 [25]; *Sulfolobus islandicus* Rod-shaped Virus 2 (SIRV2), AJ344259 [25]; *Acidianus* Rod-shaped Virus 1 (ARV1), AJ875026 [26]; *Pyrobaculum* Spherical Virus (PSV), AJ635161 [35]; *Acidianus* Two-tailed Virus (ATV), AJ888457 [26]; *Sulfolobus* Turreted Icosahedral Virus (STIV), AY569307 [37]; *Sulfolobus tengchongensis* Spindle-shaped Virus (STSV1) AJ783769 [38]; *Acidianus* Bottle-shaped Virus (ABV) EF432053 [75].

Proteins known to be present in purified viral particles or exhibiting detectable bacterial or eukaryotic signal sequences as determined by the program SignalP [76] were considered to be extracellular, for the sake of this analysis. Putative integral membrane proteins were identified using the program TMHMM [77, 78] and also removed. Because even numbers of cysteine residues are frequently employed in metal binding, these motifs were identified using the program ScanProsite [79-81], or taken from the primary literature [16], and were also removed from the intracellular pool, as these residues are unlikely to be involved in disulfide formation and would have thus contributed false positives to the analysis. This filtered metagenome represents a genomic pool enriched in intracellular proteins.

Mutagenesis of Cysteine Residues of F112

The template for site-directed mutagenesis was the F112 entry vector described above (pDest14_N-F112). Single and double mutants were generated by QuickChange™ site-directed mutagenesis kit guidelines (Stratagene). Residues Cys51 and Cys58 were mutated to either Ala or Ser single or double mutants. Standard mutagenesis primers were

designed with the position of mutation indicated by bases listed in smaller case. The sequences of the primers are as follows:

Cysteine to Alanine :

1) 5'F112_C51A_C58A:
5'-GAGCATTGAAAGCTATTgccGAAAGACATCCAGACGAGgctGAAGTACAATACAAAAAC

2) 5'F112_C51A:
5'-GAGCATTGAAAGCTATTgccGAAAGACATCCAGACGAGTGTGAAGTACAATACAAAAAC

3) 5'F112_C58A:
5'-GAGCATTGAAAGCTATTTGCGAAAGACATCCAGACGAGgctGAAGTACAATACAAAAAC

Cysteine to Serine:

1) 5'F112_C51S_C58S :
5'-GAGCATTGAAAGCTATTtccGAAAGACATCCAGACGAGtctGAAGTACAATACAAAAAC

2) 5'F112_C51S:
5'-GAGCATTGAAAGCTATTtccGAAAGACATCCAGACGAGTGTGAAGTACAATACAAAAAC

3) 5'F112_C58S:
5'-GAGCATTGAAAGCTATTTGCGAAAGACATCCAGACGAGtctGAAGTACAATACAAAAAC

The QuickChange™ site-directed mutagenesis kit was used to carry out the mutagenesis reaction (Stratagene). 50ng of the F112 wild-type expression vector (pDest14_NF112) was used as a template for the mutagenesis PCR reaction along with the appropriate primers. 1µl of Dpn I enzyme at the supplied concentration was added to each amplification reaction. The reaction mixtures were then incubated at 37 ^{o}C for 1-2 hours. The sequence of the resulting expression clones (pDest14_N- F112_C51A, pDest14_N F112_C58A, pDest14_N- F112_C51AC58A , pDest14_N- F112_C51S, pDest14_N- F112_C58S, pDest14_N- F112_C51S C58S) were verified using ABI BigDye

Terminator Cycle sequencing. Expression and purification for the proteins with the correct mutations were carried out similar to the protocol described earlier for F112.

Circular Dichroism (CD) Spectroscopy

CD spectroscopy was performed on a Jasco-810 spectropolarimeter with constant N_2 flushing (Jasco, Inc., Easton, MD). A Lauda circulating water bath with Peltier thermal regulator was used to control the temperature of the optic cell chamber, where rectangular cells of 1 cm path length were used. For CD wavelength scans, a protein stock solution of each polypeptide (0.5mg/ml) in 10 mM Bis-Tris, pH 6.5, and 50mM NaCl was diluted (60-80uL) into buffer with or without 5 mM TCEP to achieve a normalized starting ellipticity of -14.0, at an approximate concentration of 0.25-0.33uM. Protein stability was measured at a wavelength of 222nm, indicative of secondary structure of α-helices, by thermal denaturation. For thermal melting experiments, data points were taken at 1°C intervals at a scan rate of 60°C per hour from 40°C to 92°C.

Thermal Melt Assays Using Fluorescence Measurements

Protein stability as a function of temperature was studied using a real-time PCR instrument (Rotor-Gene 3000 from Corbett Research, Australia) [82, 83]. Purified F112 at a final concentration of 600μg/ml (in 10mM Bis-tris, pH 6.5, 50mM NaCl) was mixed with the reporter dye SYPRO Orange (1:1000 dilution, Invitrogen - catalog # S6650). Studies under denaturing conditions were carried out in the presence of DTT (final concentration of 100mM) mixed with purified F112 (final concentration at 300μg/ml) and the reporter dye under the same buffer conditions as before. Samples were heated from 35-99^{o}C at a rate of 1^{o}C per minute. The excitation wavelength was 470 nm with

emission detected at 555 nm. Raw data from the PCR instrument was exported to MS Excel for data plots.

Materials and Methods for the Structure Determination of D212

Cloning

ORF D212 was amplified using a nested Polymerase Chain Reaction (PCR) as described above. The internal forward and reverse primers were:

5'- ATGCATCACCATCACCATCACATGACCGAAACTGATTTTAA -3' and:

5'- GTACAAGAAAGCTGGGTCCTACTTATTACTCCTAGCTAAAT -3'

respectively, while the external forward and reverse primers used for the reaction was same as used in Menon et.al. [50]. The entry clone (BPclone-D212) was then transferred into pDEST14 (Invitrogen) with a second site-specific recombination reaction, yielding the expression vector pDest14_N-D212.

Expression and Purification

BL21 (DE3)-RIL *E.coli* cells (Stratagene) were transformed with pDest14_N-D212 for protein expression. A single colony was used to inoculate 5ml of Luria Broth (LB) + 100µg/ml ampicillin with growth for 5-6 hours, followed by expansion to a 25ml overnight culture. The overnight culture was then scaled up to 1 L and cells were grown to an OD_{600} between 0.6 and 0.8. Protein expression was induced using 0.5mM IPTG (isopropyl-ß-D-thiogalactopyranoside) for 4-6 hours. Cells were harvested by centrifugation at 6,000 ×g for 20 minutes and the pellets were stored at -20^{o}C.

For D212 purification, cell pellets were thawed and resuspended in 5ml lysis buffer (20mM Tris, pH 8.0, 20mM $NaH_2PO_4.H_2O$, 1M NaCl and 5mM imidazole along with freshly prepared 0.1mM phenyl methyl sulfonyl fluoride (PMSF) per gram of cell pellet. The resuspended cells were then passed through a French Press (American Instrument Co., Inc., Silver Springs, MD) to lyse the cells. The cell lysate was incubated at 65°C for 20 minutes to denature contaminating *E. coli* proteins, and clarified by centrifugation at 22,000 ×g for 20 minutes. The supernatant was applied to a column containing a 3-5ml bed volume of HIS-Select™ Nickel Affinity Gel (Sigma-Aldrich). The column was washed with 10 column volumes of wash buffer (20mM Tris (pH 8.0), 20mM $NaH_2PO_4.H_2O$, 1M NaCl and 10mM imidazole) and D212 was eluted in 20mM Tris (pH 8.0), 300mM NaCl and 200mM imidazole. Size-exclusion chromatographic purification of D212 was performed using a Superdex^{TM} S-75 (Amersham-Pharmacia) column equilibrated with 10mM Bis-Tris (pH 6.5) and 100mM NaCl buffer. Protein concentrations were determined by Bradford assay [59] using Protein Assay Reagent (Bio-Rad) and BSA as a standard.

Crystallization and Data Collection

Purified D212 was concentrated to 10mg/ml using spin concentrators. The protein was crystallized using hanging drop vapour diffusion. Drops were setup at 24^{o}C using 2µl of D212 and 2µl of well solution consisting 0.1M HEPES (pH 7.5), 0.05M K_2HPO_4, 10% (w/v) PEG 8000. Crystals up to 0.1 x 0.05 x 0.02mm in size were obtained in 2 weeks time. The crystals were transferred to well solution containing 25% glycerol as a cryoprotectant for 10 minutes, and then flash-frozen in liquid nitrogen. Heavy atom

soaks were carried out for 1 hour in 10mM K_2PtCl_4 or 30 seconds in 0.5M NaBr in SML containing glycerol. A single wavelength dataset centered on the Br K-edge and a three wavelength Pt L-edge dataset were collected at the Stanford Synchrotron Radiation Laboratory (SSRL beamline 9-1 and 9-2). A native dataset was also collected on D212 crystals at SSRL beamline 9-1. Data were integrated, scaled and reduced in space group $P2_1$ to a resolution of 2.4Å using HKL2000 [84]. For the Mn^{+2} metal soaks native D212 crystals were transferred to a modified SML containing 0.1M HEPES (pH 7.5), 10% (w/v) PEG 8000, and 10mM of $MnCl_2$. As phosphates compete for binding the metal ions, the K_2HPO_4 present in original condition was removed from the SML mix. The crystals were soaked in this phosphate free buffer, overnight at 24^{o}C before they were flash-frozen. Data were collected at 100 K on a Cu-Kα edge at the home source and a MAR345 image plate detector. The 3.0Å data obtained were processed using HKL-2000.

Structure Determination and Model Refinement

The structure was solved using multiple isomorphous replacement with anomalous scattering (MIRAS). SOLVE [64] was used to determine the positions of the heavy atoms and to calculate initial phases. SOLVE [64] was used to identify the heavy atom substructure and to calculate initial phases. Using the relevant data sets two platinum sites and one bromine site were identified per asymmetric unit. RESOLVE [65] was used for density modification and initial model-building. Iterative model building with COOT [85] and refinement with REFMAC5 [67, 68] led to the final model. Translational/Libration/Screw (TLS) parameters [69, 70] were included in the refinement, wherein each D212 subunit was divided into 15 TLS groups. The final R

factors for the model were 20.9%/24.5% (R_{cryst}/R_{free}) respectively. Model quality was assessed with Molprobity [86]. Residues 1-17 from the N-terminus, 55-67 in the β1-β2 loop, 107-114 in the β3-β4 loop, 169-176 in the β7-β8 loop and 206-212 from the C-terminal are not modelled due to the lack of interpretable electron density. The residue Glu125 on chain B has zero occupancy for the side chain atoms. Residue Gln155 on chain A exhibits an alternate conformation for the side chain atoms and has been modeled both ways with half occupancies. Beyond residue Pro56 in chain A, the electron density becomes weak and ambiguous with two possible directions for the chain to continue and was left unmodeled. Three dimensional structural homology searches were carried out using the DALI [71], VAST [72] and SSM [87] servers. Structural figures were generated with PYMOL [74]. Coordinates and structure factors have been deposited in the Protein Data Bank under accession code 2W8M.

Synthetic DNA Substrates

Four-way junctions or Holliday junctions (Hj) were assembled as previously described in Birkenbihl et.al [88]. The fixed Hj (F) was assembled using the following oligonucleotides:

11: 5’-GCTCGCGTGCACGGATCCCCGCATGCGGACTGCAGCCAGTCGGATG-3’ (46nt),

12: 5’-AGAGGCGACGCGGTCTAGAGGGCGTTGCGGGGATCCGTGCACGCGAG-3’ (47nt),

13: 5’-ATGCCTCGAATTCAGGCGTACGCCCTCTAGACCGCGTCGCCTC-3’ (43nt),

14: 5’-CATCCGACTGGCTGCAGTCCGCAACGCCTGAATTCGAGGCATT-3’ (44nt).

While the mobile junction (M) was prepared using the following oligonucleotides:

01: 5’-GCTCGCGTGCACGGATCCGGCGTACGCCACTGCAGCCAGTCGGATG-3’ (46nt),

02: 5'-AGAGGCGACGCGGTCTAGAGGGCGTACGCCGGATCCGTGCACGCGAG-3' (47nt),

03: 5'-ATGCCTCGAATTCAGGCGTACGCCCTCTAGACCGCGTCGCCTC-3' (43 nt),

04: 5'-CCATCCGACTGGCTGCAGTGGCGTACGCCTTGAATTCGAGGCATT-3' (44nt).

A linear dsDNA control was prepared by annealing oligonucleotide 04 with its complement, 04'. Each oligonucleotide was labeled at the 5' end with ^{32}P using γ-^{32}P-ATP and T4 polynucleotide kinase. The junctions were assembled by hybridization of equimolar amounts of a single ^{32}P-labeled oligonucleotide with three complementary unlabeled oligonucleotides in TBE (80 mM Tris, 80 mM borate, 2 mM EDTA) containing 10 mM $MgCl_2$ buffer. The junctions were purified over a 6% native polyacrylamide gel and eluted in 10 mM Tris (pH 8.0) buffer and used for further assays.

Endonuclease Assay

The experimental conditions for the endonuclease activity assay were essentially as described by Birkenbihl et.al for the Hjcs from SIRV1 and SIRV2 [88] with minor modifications. Cleavage assays were performed by incubating, 10nM of assembled mobile or fixed (^{32}P)-radioactively labeled cruciform DNA with 1µM D212 protein in reaction buffer consisting of 10mM Bis-Tris pH 6.5, 100mM NaCl, 10mM $MgCl_2$, 1mM DTT, at 25ºC for 2 hours. In parallel reactions, T4 endonuclease VII was used as a positive control for Hjc activity. The reaction was stopped by adding 2µl of loading buffer (20% bromophenolblue, 25% glycerol in TBE) to the 10µl total reaction volume, and the mixture was analyzed on a 10% native TBE gel. Gels were visualized by autoradiography.

Electrophoretic Mobility Shift Assay

Gel retardation assays to study the binding of D212 to various types of DNA such as four way junctions, dsDNA or ssDNA were performed. 10nM ^{32}P labeled cruciform and dsDNA prepared as discussed above were incubated with D212 at various concentrations ranging from10-200nM. To analyze ssDNA binding activity of D212 the protein was incubated with 10nM labeled oligonucleotide h (see the 'Synthetic DNA Substrates' section for sequence details) at various concentrations of D212 similar to the previous binding experiment. The same buffer system was used for the endonuclease cleavage assay was also used for this experiment except that the 10mM $MgCl_2$ was substituted with 5mM EDTA. The assay was carried out for a maximum of 1 hour at room temperature. Samples were removed at 15 minute intervals and the reaction was stopped by adding the loading dye (20% bromophenolblue, 25% glycerol in TBE). The mixtures were immediately loaded onto an 8% native polyacrylamide gel and run at room temperature for 2 hours at 7V/cm in TBE buffer. Bands on the gel were visualized by autoradiography.

Similar mobility shift assays were also tried with PCR amplified 1000bp DNA fragments encompassing the entire SSV1 genome (see supplementary Table S-1). For this assay 20nM of the amplified DNA was incubated with 200nM D212 protein at 65°C for 30 minutes in the same buffer (10 mM Bis-Tris pH 6.5, 100 mM NaCl, 5 mM EDTA, 1 mM DTT). The samples were fractionated through 1.6% agarose gel electrophoresis with ethidium bromide and visualized under UV illumination.

Materials and Methods for the Structure Determination of F61

Cloning

ORF F61 was amplified using a nested Polymerase Chain Reaction (PCR) as described above. The internal forward and reverse primers were: 5'- ATGCATCACCATCACCATCAC ATGGAAGACTTAGATTCTTT-3' and 5'- GTACAAGAAAGCTGGGTCCTA CCTCAGCACCTCTTGTTTCT-3', while the external forward and reverse primers were 5'-GGGGACAAGTTTGTACAAAA AAGCAGGCTTCGAAGGAGATAGAACCATGCATCACCATCACCATCAC-3' and 5'-GGGGACCACTTTGTACAAGAAAGCTGGGTCCTA-3', respectively. The entry clone (pDONR201_N-F61) was then transferred into pDEST14 (Invitrogen) with a second site-specific recombination reaction, yielding the expression vector pDest14_N-F61.

Expression and Purification

For F61 protein expression, BL21(DE3)-RIL *E.coli* cells (Stratagene) were transformed with pDest14_N-F61. A single colony was used to inoculate 5ml of Luria Broth (LB) + 100µg/ml ampicillin with growth for 5-6 hours, followed by expansion to a 25ml overnight culture. The overnight culture was then scaled up to 1 L and cells were grown to an OD_{600} between 0.6 and 0.8. Protein expression was induced using 1mM IPTG (isopropyl-ß-D-thiogalactopyranoside) for 4-6 hours. Cells were harvested by centrifugation at 6,000 ×g for 20 minutes and the pellets were stored at -20 ^{o}C.

For F61 purification, cell pellets were thawed and resuspended at 5ml/gram of cell pellet in lysis buffer (10mM Tris, pH 8.0, 5mM imidazole, 300mM M NaCl) along with freshly

prepared phenyl methyl sulfonyl fluoride (0.1mM PMSF). The resuspended cells were then passed through a French Press (American Instrument Co., Inc., Silver Springs, MD) to ensure complete lysis of the cells. The cell lysate was incubated at 65°C for 20 minutes to denature contaminating *E. coli* proteins, and clarified by centrifugation at 22,000 ×g for 20 minutes. The supernatant was applied to a column containing a 3-5ml bed volume of HIS-Select™ Nickel Affinity Gel (Sigma-Aldrich). The column was washed with 10 column volumes of wash buffer (10mM Tris, pH 8.0, 10mM imidazole, 20mM $NaH_2PO_4.H_2O$, 1M NaCl) and F61 was eluted in 10mM Tris, pH 8.0, 200mM imidazole, 50mM NaCl. A contaminating *E.coli* protein with higher molecular weight was found to co-purify with F61. Washing the column with a high salt buffer containing phosphate was found to get rid of the contaminating protein. Hence the wash buffer used for the purification of F61 was different from the standard low-salt wash buffer used for other proteins. F61 was found to precipitate out in Tris-buffer (pH 8) after 2-3 days. A solubility profile tray was set up to find the best buffer condition for F61. The pH and salt concentrations were varied in this experiment. Unexpectedly, after 3-4 days rod shaped crystals for F61 were found at pH=4 and pH=5. Hence a new buffer, 10mM Na-acetate (pH 5.0) and 50mM NaCl, was applied to a calibrated Superdex S-75 (Amersham-Pharmacia) and was used for the final elution of F61. Protein concentrations were determined by Bradford assay [59] using Protein Assay Reagent (Bio-Rad) and BSA as a standard.

Preparation of Leucine to Methionine Mutants

As F61 did not have a native methionine, and molecular replacement failed to solve the crystallographic phase problem (below), mutants were made where specific Leu residues of the wild-type F61 were substituted to Met. Two sites, Leu11 and Leu50, were chosen to make the mutations depending on the position of the leucines using secondary structure prediction. Single mutants were generated by QuickChange™ site-directed mutagenesis kit (Stratagene) where residues Leu11 and Leu50 were mutated to methionine. Standard mutagenesis primers were designed with the position of mutation indicated by bases listed in smaller case. The sequences of the primers are as follows:
Leucine to Methionine:

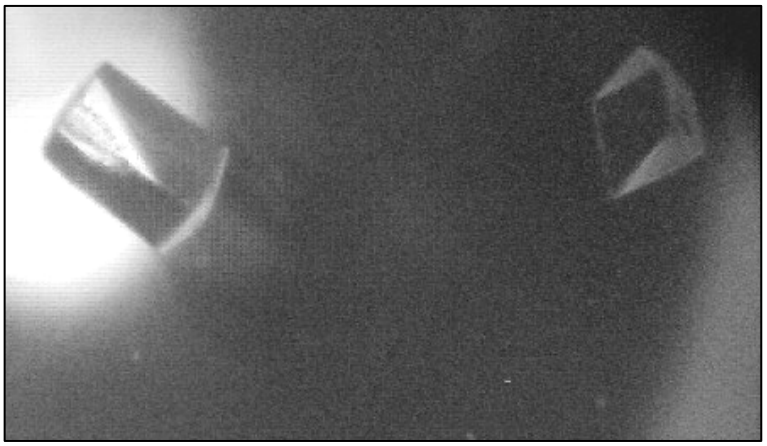
1) 5'- F61_L11M:
5'- TCTTTATCAAGGAAAaTGTTGAGTAAAGTCA

2) 5'- F61_L50M:
5'- GCTGAACTATACTCAaTGAAGTTAAAGATAG

Fifty ng of the F61 wild-type expression vector (pDest14_N-F61) was used as a template for the mutagenesis PCR reaction along with the appropriate primers. 1µl of Dpn I enzyme at the supplied concentration was added to each amplification reaction. The reaction mixtures were then incubated at 37^{o}C for 1-2 hours. The sequence of the resulting expression clones (pDest14_N-F61_L11M and pDest14_N-F61_L50M) were verified using ABI BigDye Terminator Cycle sequencing. For production of selenomethionine-labelled variant protein (SeMet-F61_L11M and –F61_L50M), the methionine auxotroph B834 (DE3) *E.coli* (Novagen) was transformed with pDest14_N-F61_L11M and pDest14_N-F61_L50M respectively. A single colony from each mutant

was used to inoculate two 5ml overnight LB/AMP culture. The culture was centrifuged at 500 ×g to pellet the cells, and the pellet was washed twice with 0.5ml of Milli-Q-H_2O, and resuspended in 1L of SeMet medium [61]. Expression and purification for the two single variants F61_L11M and F61_L50M were identical to that described above for the native F61 protein. The yield of SeMet-labelled F61 was approximately 20 mg per gram of cell pellet in the case of the L11M variant and 12 mg per gram of cell pellet for the L50M variant.

Crystallization and Data Collection



Purified F61_L11M and F61_L50M were concentrated to ~12 mg/ml with spin concentrators. The proteins were crystallized using hanging drop vapour diffusion. Drops were setup at 23 ^{o}C using 2 µl of F61 mutants and 2 µl of well solution consisting of 0.2 M MgFormate and 20% (w/v) PEG 3350. The crystals were transferred to synthetic mother liquor containing 22.5% PEG400 as a cryoprotectant for 30 seconds, and then flash-frozen in liquid nitrogen. The crystals of the F61_L11M variant gave good diffraction and 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 integrated, scaled and reduced in space group C2 to a resolution of 2.0 Å using HKL2000 [84].

Structure Determination and Model Refinement

SOLVE [64] was used to determine the positions of the selenium-substructure and to calculate initial phases. Six selenium sites per asymmetric unit were identified using SOLVE. RESOLVE [65] was used for density modification and initial model-building. The current R factors for the model are 32.3%/31.3% (Rfree/Rcryst) respectively. Further cycles of manual building with COOT [85] and refinement with REFMAC5 [67, 68] are currently under way.

Materials and Methods for the Cloning and Purification of E73

Cloning

ORF E73 was amplified using a nested Polymerase Chain Reaction (PCR) as described above. The internal forward and reverse primers were: 5'- ATGCATCACCATCACCATCACTTGGTTGAAAGTAAAAAGAT-3' and 5'-GTACAAGAAAGCTGGGTCCTATTTTGCGGGAAAAAACCGCT-3', while the external forward and reverse primers were 5'-GGGGACAAGTTTGTACAAAA AAGCAGGCTTCGAAGGAGATAGAACCATGCATCACCATCACCATCAC-3' and 5'-GGGGACCACTTTGTACAAGAAAGCTGGGTCCTA-3', respectively. The entry clone (BPclone-E73) was then transferred into pDEST14 (Invitrogen) with a second site-specific recombination reaction, yielding the expression vector pDest14_N-E73.

Expression and Purification

For E73 protein expression, BL21(DE3)-RIL *E.coli* cells (Stratagene) were transformed with pDest14_N-E73. A single colony was used to inoculate 5ml of Luria Broth (LB) + 100 μg/ml ampicillin with growth for 5-6 hours, followed by expansion to a 25ml overnight culture. The overnight culture was then scaled up to 1 L and cells were grown to an OD_{600} between 0.6 and 0.8. Protein expression was induced using 1mM IPTG (isopropyl-ß-D-thiogalactopyranoside) for 4-6 hours. Cells were harvested by centrifugation at 6,000 ×g for 20 minutes and the pellets were stored at -20^{o}C.

For E73 purification, cell pellets were thawed and resuspended at 5ml/gram of cell pellet in lysis buffer (10mM Tris, pH 8.0, 5mM imidazole, 300mM M NaCl) along with freshly prepared phenyl methyl sulfonyl fluoride (0.1mM PMSF). The resuspended cells were then passed through a French Press (American Instrument Co., Inc., Silver Springs, MD) to ensure complete lysis of the cells. The cell lysate was incubated at 65°C for 20 minutes to denature contaminating *E. coli* proteins, and clarified by centrifugation at 22,000 ×g for 20 minutes. The supernatant was applied to a column containing a 3-5ml bed volume of HIS-Select™ Nickel Affinity Gel (Sigma-Aldrich). The column was washed with 10 column volumes of wash buffer (10mM Tris, pH 8.0, 10mM imidazole, 300mM NaCl) and E73 was eluted in (10mM Tris, pH 8.0, 200mM imidazole, 50mM NaCl). E73 was then applied to a calibrated Superdex S-75 (Amersham-Pharmacia) column equilibrated with 10mM Tris (pH 8.0) and 50mM NaCl. Protein concentrations were determined by Bradford assay [59] using Protein Assay Reagent (Bio-Rad) and BSA as a standard.

Crystallization

Purified E73 was concentrated to ~10mg/ml and was sent for high throughput crystallization screening (HTS) at HWI, NY. At the one week time point there were several conditions which yielded very small and sometimes showers of crystals in the drop. All efforts to reproduce better crystals at home did not yield single crystals suitable for structural studies. Expression clones (LR clone-E73) were provided to Dr. Valerie Copie's lab so that further structural and dynamic studies of this protein with multidimensional nuclear magnetic resonance (NMR) could be pursued. The initial resonance assignments for the E73 backbone and side chain atoms have now been determined and a manuscript describing the assignments has been submitted [89].

Materials and Methods for the Cloning and Purification of B64

Cloning

ORF B64 was amplified using a nested Polymerase Chain Reaction (PCR) as described above. The internal forward and reverse primers were: 5'- ATGCATCACCATCACCATCACATGAAGGCAAGAGTGGAATA-3' and 5'- GTACAAGAAAGCTGGGTCCTACTGCACCTCTTCTAGATATT-3', while the external forward and reverse primers were 5'-GGGGACAAGTTTGTACAAAA AAGCAGGCTTCGAAGGAGATAGAACCATGCATCACCATCACCATCAC-3' and 5'-GGG GACCACTTTGTACAAGAAAGCTGGGTCCTA-3', respectively. The entry clone (pDONR201_N-B64) was then transferred into pDEST14 (Invitrogen) with a second site-specific recombination reaction, yielding the expression vector pDest14_N-B64.

Expression and Purification

For B64 protein expression, BL21(DE3)-RIL *E.coli* cells (Stratagene) were transformed with pDest14_N-B64. A single colony was used to inoculate 5ml of Luria Broth (LB) + 100µg/ml ampicillin with growth for 5-6 hours, followed by expansion to a 25ml overnight culture. The overnight culture was then scaled up to 1 L and cells were grown to an OD_{600} between 0.6 and 0.8. Protein expression was induced using 1 mM IPTG (isopropyl-ß-D-thiogalactopyranoside) for 4-6 hours. Cells were harvested by centrifugation at 6,000 ×g for 20 minutes and the pellets were stored at -20 ^{o}C.

For B64 purification, cell pellets were thawed and resuspended at 5 ml/gram of cell pellet in lysis buffer (10mM Tris, pH 8.0, 20 mM $NaH_2PO_4.H_2O$, 5mM imidazole, 1M M NaCl) along with freshly prepared phenyl methyl sulfonyl fluoride (0.1 mM PMSF). The cell suspension was vortexed vigorously to resuspend the cells. The homogenate was then passed through a French Press (American Instrument Co., Inc., Silver Springs, MD) to ensure complete lysis of the cells. The cell lysate was incubated at 65 °C for 20 minutes to denature contaminating *E. coli* proteins, and clarified by centrifugation at 22,000 ×g for 20 minutes. The supernatant was applied to a column containing a 3-5ml bed volume of HIS-Select™ Nickel Affinity Gel (Sigma-Aldrich). The column was washed with 10 column volumes of wash buffer (10mM Tris, pH 8.0, 20mM $NaH_2PO_4.H_2O$, 10mM imidazole, 1M NaCl) and B64 was eluted in (10mM Tris, pH 8.0, 20mM $NaH_2PO_4.H_2O$, 200mM imidazole, 300mM NaCl). B64 was then applied to a calibrated Superdex S-75 (Amersham-Pharmacia) column equilibrated with 10mM Tris (pH 8.0) and 150mM NaCl. A very broad elution pattern was seen for B64 on a SEC indicative of either a non-homogenous protein sample containing a mixture of B64

protein in various oligomeric states or due to a prolonged interaction of the protein with the column. The protein elutes were centered around 8-10ml of elution volumes, suggestive of a dimer for B64 in solution. Protein eluted with a Protein concentrations were determined by Bradford assay [59] using Protein Assay Reagent (Bio-Rad) and BSA as a standard.

Crystallization

Purified B64 was concentrated to ~10mg/ml and was sent for high throughput crystallization screening (HTS) at HWI, NY. The protein did not result in any crystals in the ~1500 conditions that were scanned at HWI. The protein size was suitable for solution NMR studies. Hence expression clones for B64 were also provided to Dr. Valerie Copie's lab to pursue structural and dynamic studies of this protein with NMR.

Summary

The initial sequence analysis of SSV1 and SSVRH genomes did not reveal significant sequence homology between most of its predicted gene products and proteins in the database with known functions. In order to arrive at functional hypothesis, we initiated a systematic structure based functional annotation for these virus encoded proteins. Towards this effort I have successfully cloned a subset of genes from both viruses. Six ORFs from SSV1 and 20 ORFs from SSVRH were cloned with an N-terminal 6X His tag, into expression vectors suitable for the *E.coli* bacterial expression system. Soluble protein expression was obtained for a number of proteins, and depending on the amount of soluble protein expression and the purity of the obtained product,

several of these were chosen for additional study. Among others, soluble protein was obtained for ORFs E178, E96, and F112 from SSV1. The yields for E178, F112 and E96 were ~0.2 mgs, 1-2 mgs and ~ 4 mgs per gram of cell pellet respectively. The structure of F112 was completed as a part of this study [50]. Brian Eilers, a research assistant from our lab has recently determined the structure of E96. Although E178 did show some soluble protein expression, the protein was not of the purest form and also the yield of the protein obtained after Ni-NTA based IMAC purification was very low. Further efforts to increase the yield and purity of the protein are needed. In case of the SSVRH ORFs, C154, D212, F61, E73, A102a, B64, B79, C150 and A113 all gave soluble proteins. D212, F61, E73, B64 and C150 yielded pure proteins at concentrations ~ 5 mg, 12 mg, 12 mg, 5 mg and 13 mg per gm of cell pellet respectively and hence were ideal for crystallization trials. While for C154, A102a, B79 and A113 protein yields were ~0.5 mg, 0.2 mg, 1 mg and 1.5 mg per gram of cell pellet. Initial crystallization conditions were identified for all these proteins from high throughput crystallization screening (HTS) at HWI. Good diffraction quality crystals were obtained in house for D212 and F61 by optimization of the initial condition from HTS and structures of both these proteins have been solved as a part of this initiative. E73 and B64 showed great potential as microcrystals were seen in several different conditions from the HTS screen. But even after several trials and scanning many different crystallization conditions diffraction quality crystals could not be obtained for these proteins. Clones for both proteins have been provided to Dr. Copie's lab and NMR based structure-function annotation for the proteins are underway. We were also unable to reproduce the HWI crystallization conditions for C150. C150 is well conserved among all fusseloviruses and is predicted to

be a Zn-finger DNA binding protein. Brian Eilers has determined the structure of SSV1 B129, a homolog of SSVRH C150 (Eilers et.al., unpublished). This result further emphasizes the importance of working with homologous proteins from various *Fuselloviridae*, as one protein homolog might be better behaved than the other. The yield of soluble protein for ORF A102a was ~0.2 mg/ gm of cell pellet and hence further optimization of its protein expression is needed. For ORF A113 purification protocol needs to be further optimized as proteolysis was observed after purification by SEC. For ORF C154 and B79 purification using standard protocols resulted in proteins with sizes smaller than expected on a SDS-PAGE gel analysis. Mass spectrophotometric studies of protein fragmentation patterns might help us to identify whether this band simply shows anomalous mobility in SDS-PAGE, or if this smaller sized band is in fact a cleaved product of the correct protein, or an *E.coli* impurity. Proteins are sometimes mis-folded and form inclusion bodies, the analysis of all cell pellets on PAGE gels did not reveal any such potentially mis-folded proteins in our case.

This study was initiated to identify a subset of well-behaved proteins that are ideal candidates for structure based functional annotation studies from these unique fuselloviridae genomes using conventional methods. Little time was spent optimizing conditions for those proteins that were not well behaved. Further expression trials to identify optimal expression conditions for these proteins are needed and are likely to identify conditions that will allow expression of additional proteins suitable for structural studies. Conditions to explore include temperature, nature of the construct, use of alternative N-and C-terminal fusions, and the use of alternative expression systems such as yeast or bacullovirus. One such example where changing the expression vector by

relocating the 6X His tag from the C- to the N-terminus has resulted in increased protein expression and crystallization conditions for SSV1 F112. Other tricks such as altering the media in which the cells are propagated or co-expression with chaperonins that assist in correct folding of the proteins can also be tried [90]. Auto-induction of proteins [91] has also reportedly increased soluble protein yields for several proteins being currently under study in our lab and hence may also be productive. On the other hand, the auto-induction protocol and IMAC purification frequently results in the copurification of several contaminating proteins with molecular weights between 20 and 25 kDa. Two of these have been identified by Dr. George Gauss as glycerol kinase, and chloramphenicol acetyl transferase. These proteins, which are frequently difficult to remove by other chromatographic techniques, are not seen with IPTG induced protein expression. Currently, only salt concentrations in the lysis, wash and elution buffer were altered in the purification protocols, thus further optimization of purification conditions by changing either the pH, or sampling more salt concentrations or adding denaturant needs to be analyzed.

This current endeavor resulted in the determination of structures for three proteins F112 (SSV1), D212 and F61 (SSVRH). All three structures were found to have folds similar to other proteins in the database. Hence structure based functional annotation for these proteins were possible providing a significant increase in our understanding of this unique family of viruses.

CHAPTER 3

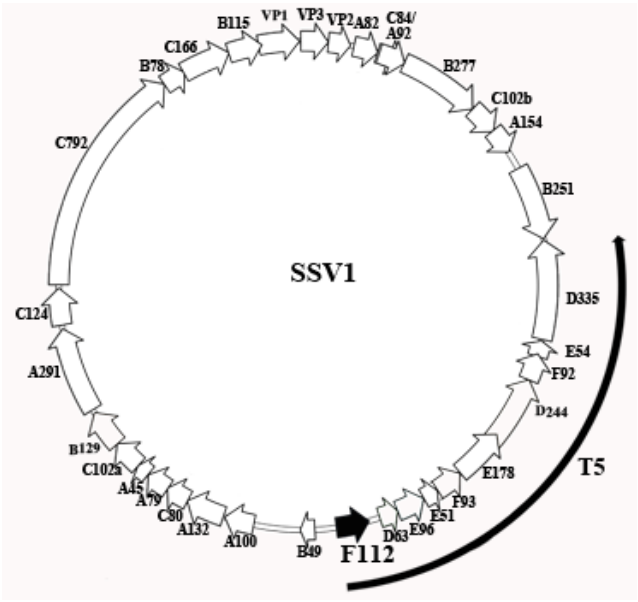
CRYSTAL STRCUTURE OF SSV1 F112; A DNA BINDING WINGED HELIX-TURN-HELIX PROTEIN.

Introduction

F112 is a 112 amino acid protein encoded in reading frame F of the SSV1 genome [22, 23]. The F112 gene product is unique to the SSV1 virus and no homologs were detected in the other members of the *fuselloviridae* family. F112 shows no similarity to proteins with known function and only weak similarity is detected to ORF70, a hypothetical protein with unknown function, from Acidianus Two-tailed Virus [41] with an E-value of 10^{-3} using a PSI-BLAST search [92]. Even with relaxed search parameters that allow an E-value below the default threshold, only weak similarity was established to a putative phospholipase (E-value of 3.7) and an AMP-dependant synthetase and ligase (E-value of 5.5). Similarly, the most recent, comprehensive bioinformatics study of archaeal viral genomes by Prangishvili et.al [16] predicted putative functions for a number of archaeal viral proteins, however F112 was not among those.

In this chapter we report on the initial biochemical characterization and structural analysis of SSV1 F112. The structure reveals similarity to a winged helix transcription factor, suggesting a possible function for F112 as a transcriptional regulator. Another interesting aspect was an intramolecular disulfide bond found in the structure of F112. The implication of the disulfide bond and cysteine distribution in the genomes of SSV1 and other hyperthermophilic viruses in general is further discussed in detail in chapter 4.

A



B

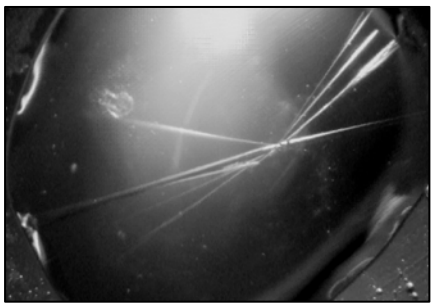


Figure 3-1. ORF F112 from SSV1.
(A) SSV1 genome with open arrows representing different ORFs from SSV1. The filled arrow shows the location of F112 in the SSV1 genome. The position and orientation of transcript T5 is shown in as a black arrow outside the genome. (B) Crystals of SeMet derivative of F112 grown in 0.1M MES (pH 6.0).

Results

The F112 construct used in this study encodes a total of 118 residues, 112 corresponding to the native protein plus an N-terminal His_6-tag, with a total calculated mass of 14,026 Da. Analytical runs on a calibrated Superdex™ S-75 column suggest a molecular weight of 20 kDa (data not shown), intermediate between the expected sizes of a monomer and dimer. Dynamic light scattering, however, indicates a particle diameter of 29.6 Å. A spherical protein with a molecular weight of 14 kDa (monomer size for F112), is predicted to have a particle diameter of 31.8 Å while for a 28 kDa (dimer size for F112) the particle diameter is expected to be 39.6 Å. This suggests that F112 is present in solution as a monomer.

Structure of F112

The protein crystallizes in space group R3 with a single molecule in the asymmetric unit. The structure was determined at 2.3 Å resolution using multiwavelength anomalous diffraction [93] with crystals of selenomethionine incorporated protein [94]. Detailed statistics on data collection and structure refinement are presented in Tables A-7 and A-8. Interpretable electron density begins at the N-terminus with residue 4 of the native sequence and ends at residue 73. Residues 74 through 112 were not visible in the electron density, and accordingly not modelled.

Table A-7. Data collection for F112

Data Set	Se-Edge	Se-Peak	Se-Remote
Wavelength (Å)	0.979320	0.979180	0.891940
Space Group	R3 (Hexagonal Setting)		
Cell Constants (a,b,c; Å) $\alpha=\beta=90.00^{o}$, $\gamma=120.00^{o}$	87.72 , 87.72 , 34.49		
Resolution Range[a] (Å)	25.53-2.3 (2.36-2.30)		
Unique Reflections[a]	4410 (330)	4386 (331)	4411 (333)
Average Redundancy[a]	7.7 (7.4)	7.6 (7.3)	7.7 (7.3)
I/σ[a]	21.5 (12.7)	16.2 (10.3)	20.6 (12.3)
Completeness (%)	99.7 (100)	99.7 (100)	99.8 (100)
R_{sym}[a,b] (%)	6.6 (11.1)	8.8 (12.6)	6.9 (11.3)

[a]Numbers in parenthesis refer to the highest resolution shell.

[b]$R_{sym}=100*\Sigma_h\Sigma_i|I_i(h)-\langle I(h)\rangle|/\Sigma_h I(h)$ 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.

The structure of the N-terminal domain reveals a winged helix-turn-helix motif (wHTH). This motif typically contains a right-handed three helix bundle that defines the HTH motif, flanked by a two- or three- stranded antiparallel ß-sheet that constitutes the wing (Figure 3-2). In the case of F112, the polypeptide runs from the N-terminus,

through helix α1, strand β1, helices α2 and α3, and then adds strands β2 and β3 to complete the wing. Proteins displaying this fold commonly insert the third helix, or recognition helix, into the major groove of their target DNA to form specific interactions with the bases. Residues on the recognition helix also commonly interact with the phosphodiester backbone of the DNA. In addition, residues at the N-terminus of helix α1 and the residues in the wing are also predicted to interact with the ribose-phosphate backbone. In some cases, residues at the tip of the wing are also found to insert into the minor groove of the target DNA [95].

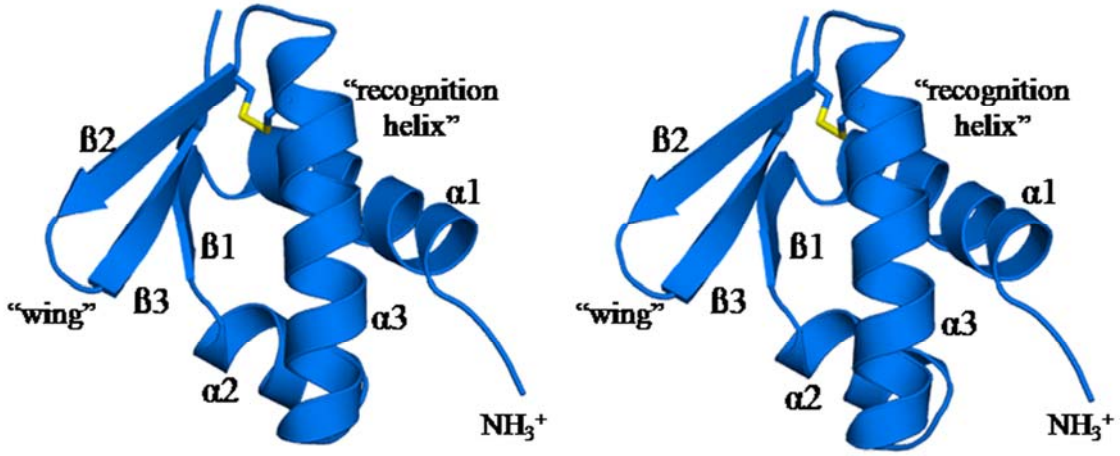


Figure 3-2. Structure of F112.
Stereo image of the F112 structure. The polypeptide traces from the N-terminus ($^{+}H_3N$) through helix α1, strand β1, helices α2 and α3, strands β2 and β3to the C-terminus (COO^{-}). The α3-helix is the recognition helix, which typically binds the major groove of the DNA. The reverse turn between strands, β2 and β3 forms the wing of the wHTH motif. Generally, the wing also contacts the DNA. The intramolecular disulfide bond between Cys51(α3)-Cys58(β2) is depicted as a stick model, with sulfur atoms in yellow.

Not surprisingly, a search for structural homologues using the DALI [71] and VAST [72] servers returns a multitude of proteins containing winged-helix domains similar to that of F112. This includes DP2 (PDB 1CF7-B, DALI Z-score = 6.0, RMSD = 2.6 Å with 11 %

identity over 61 aligned residues), a winged helix DNA binding eukaryotic transcription factor that forms a heterodimeric complex with a second winged helix protein, E2F4 [92]. The DP2:E2F4 complex plays a central role in the eukaryotic cell cycle where it promotes transition from G_1 to S phase. As expected, the recognition helix (α3) of DP2 is found within the DNA major groove, with the wings contacting the ribose-phosphate backbone. As shown in Figure 3-3A, similar interaction is easily modeled for F112 by superposition of the DP2-DNA complex [PDB ID 1CF7, [92]] upon F112 using lsqkab [67]. The docking suggests that Asn^{42}, Gln^{44} and/or Arg^{45} of F112 are likely to contribute to base specific interactions with the target DNA sequence (Figure 3-3 A). Further, the calculated electrostatic potential on the surface of F112 reveals areas of significant positive potential that are appropriately positioned for interaction with the ribose-phosphate backbone (Figure 3-3 B).

Table A-8. Model Refinement for F112	
R_{cryst}[c] (%)	17.3
R_{free}[c] (%)	19.9
Real Space CC[d] (%)	94.8
Mean B Value (overall; Å^2)	23.5
Coordinate Error (based on maximum likelihood, Å)	0.121
RMSD from ideality:	
Bonds (Å)	0.007
Angles (°)	0.923
Ramachandran Plot[e]:	
Most Favored (%)	90.8
Additional Allowed (%)	9.2
PDB Accession Code	2VQC
[c]$R_{cryst} = \Sigma\|\|F_o\|-\|F_c\|\|/\Sigma\|F_o\|$ 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 PROCHECK[96].	

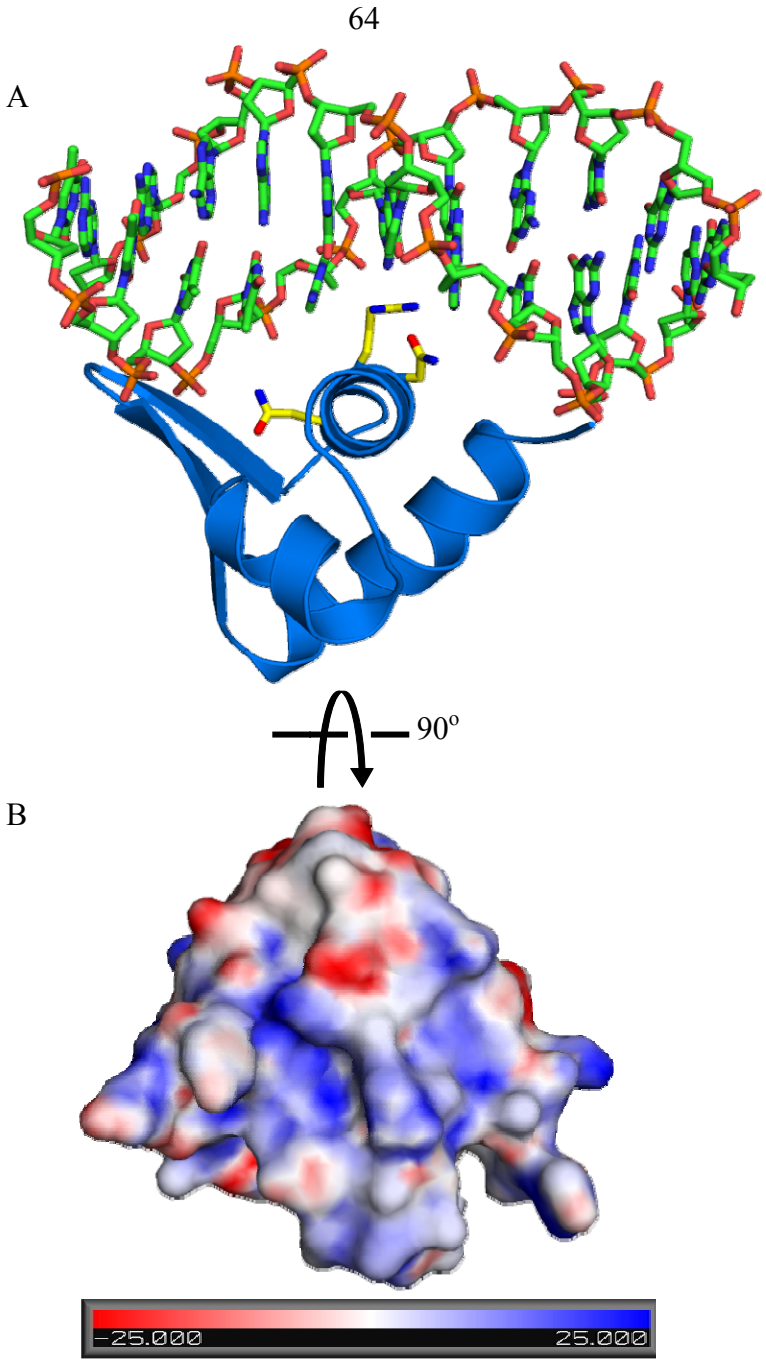


Figure. 3-3. A model for F112 interaction with DNA.
(A) The DNA orientation was obtained from superposition of the DP2-DNA complex onto F112 where the F112 monomer is depicted in blue. The orientation of F112 is such so that the recognition helix (α3) is perpendicular to the plane of the paper. The model suggests that the N-terminal residues, the recognition helix and the wing of F112 together facilitate in DNA recognition. The residues Asn^{42}, Gln^{44} and/or Arg^{45} in the recognition helix are depicted as sticks and are likely to be involved in making base specific interactions with the target DNA. Superpositions were prepared with lsqkab. (B) Electrostatic potential mapped on the putative DNA binding surface of F112. The surface potential reveals a predominantly positive surface conducive to DNA binding. Images were rendered in Pymol [74].

Structural similarity to a number of prokaryotic wHTH proteins is also seen. One example is SarR (PDB 1HSJ-A, DALI Z-score = 5.5, RMSD = 2.2 Å with 10 % identity over 61 aligned residues), a homodimeric transcription factor that regulates production of the SarA transcript in *Staphylococcus aureus* [97]. The structural similarities between F112 and proteins like DP2 and SarR suggests that F112 will also function as a transcriptional regulator. However, the winged-helix motif has also been adapted for other purposes. It is often found in larger proteins where it provides a DNA binding activity that complements nucleic acid associated enzymatic activities. It is also found less frequently in nucleic acid independent roles; for example, as a domain involved in mediating protein-protein interactions [95]. The small size of F112 would, however, seem to rule out an intrinsic enzymatic activity, while the disordered C-terminus would seem the more likely candidate for any protein-protein interactions. Thus a DNA binding function would seem to be the most likely role for the N-terminal winged-helix domain of F112.

To test this prediction, we looked for interactions between F112 and double stranded DNA using an electrophoretic mobility shift assay (EMSA). While demonstration of a high affinity protein-DNA interaction requires isolation and/or knowledge of a specific recognition sequence, most sequence specific DNA binding proteins also show non-specific interactions with DNA at elevated DNA or protein concentrations [98]. Indeed, non-specific interactions are clearly seen in EMSAs using F112 at low micromolar concentrations and small randomly chosen, DNA fragments, from the SSV genome (Figure 3-4).

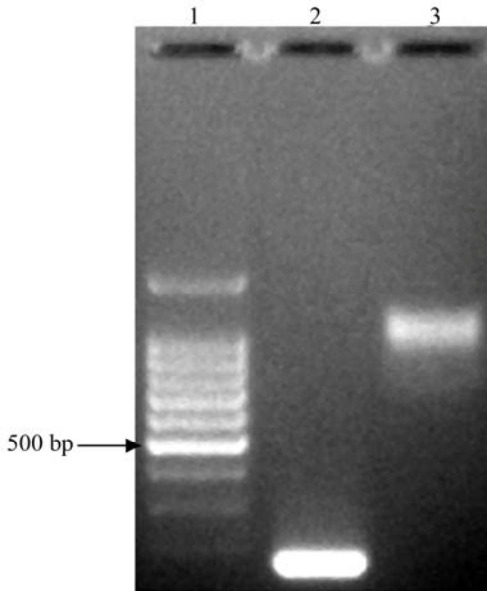


Figure 3-4. Non-specific binding of F112 to double-stranded DNA
Protein-DNA interactions were monitored with agarose gel electrophoresis. DNA was stained with ethidium bromide and visualized by UV illumination. Lane 1 contains a 100-bp DNA ladder. Lane 2 contains the 153-bp DNA fragment in the absence of F112. Lane 3 contains DNA in the presence of a 5-fold molar excess of F112 resulting in a shift in the mobility of the DNA.

However, a search for a high affinity interaction using a library of PCR amplified DNA fragments covering the SSV1 genome was unsuccessful, although this approach has worked nicely before for STIV F93 (Eilers and Lawrence, unpublished). The search for a high affinity binding site may be more successful if an F112 heterodimer can be identified.

Disulfide Bonds

Another notable aspect of the F112 structure is the presence of an intramolecular disulfide bond. Clear electron density is observed between Cys^{51} in helix α3 and Cys^{57} in strand β1, with the 2.0 Å sulfur-sulfur distance typical of a disulfide (Figure 3-5). At first glance, this might seem surprising, as DNA binding proteins typically reside in a

reducing intracellular environment in most organisms, one that is generally not conducive to disulfide formation [99].

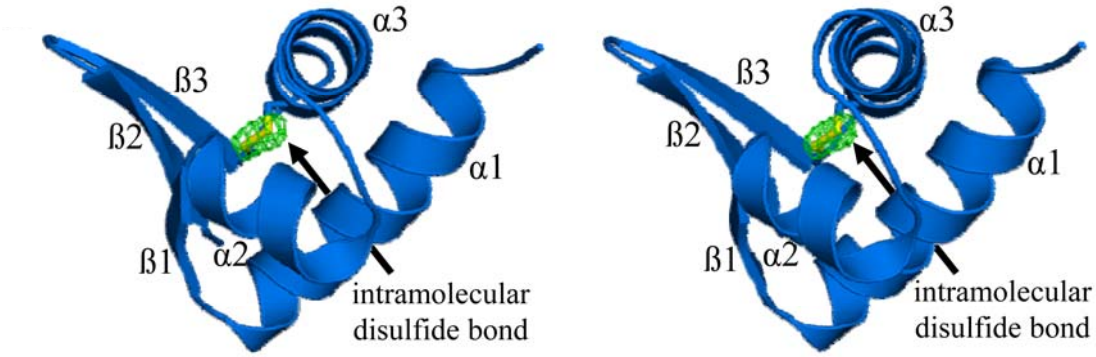


Fig. 3-5 Intramolecular Disulfide Bond of F112.
Relative to the orientation in panel A, F112 is rotated 90° about the depicted horizontal axis. The secondary structural elements are labeled as in Fig3-2. The view is now along the axis of the recognition helix. The disulfide bond is depicted as a stick model surrounded by the green mesh of difference electron density (contoured at 5σ) in which the sulfur atoms were omitted from model refinement.

Interestingly, however, crystal structures of SSV1 E96 (unknown function) and SSV1 B129, a zinc-finger DNA binding protein, also reveal disulfide bonds (Eilers et al. and Lawrence, unpublished). It is possible, though, that SSV1 might encapsidate these proteins within the viral particle; thus, these proteins may not be restricted to a strictly intracellular environment. In order to address whether these proteins might be packaged within SSV1, the Bothner and Young laboratories analyzed the purified SSV1 viral particle by mass spectrometry. However, no evidence was found for the incorporation of B129, E96 or F112 in the SSV1 particle. While this does not prove their absence from the viral particle, it is consistent with their predicted intracellular distribution. Further, genomic analysis demonstrates that disulfides are probably common in the intracellular proteins of many hyperthermophilic organisms including STIV [98]. The effect of the

disulfide bond and its physiological relevance in the proteins from thermophilic viruses is characterized in depth and discussed in chapter 4.

Discussion

A recent, comprehensive bioinformatics study [16], concluded that the genomescapes of crenarchaeal viruses remain relatively barren of recognizable features, and that only a small fraction of the genes are conserved in other viruses or cellular life forms. Although several members of the helix-turn-helix family were identified through this study, F112 was not among those. The F112 structure thus demonstrates that additional proteins belonging to the helix-turn-helix family remain to be identified in these crenarchaeal viral genomes, and suggests the frequency of the HTH motif in these viral genomes may be a substantial underestimate. Thus, once again the usefulness of structural annotation for these unique viral genomes has been nicely demonstrated.

The structure of F112 revealed a winged helix-turn-helix motif with closest structural homology to a number of prokaryotic and eukaryotic wHTH transcription factors. The HTH domain is considered to be a paradigm to study DNA-protein interactions and the wHTH motif is the most commonly found HTH domain in the archaea [100]. The wHTH domain of F112 was found to be closely related to eukaryotic transcription factor DP2, prokaryotic transcription factor SarR and a more recently to a eukaryotic Z-DNA binding protein, double-stranded RNA-specific adenosine deaminase (ADAR1, PDB ID 3F23-C, DALI Z-score = 7.1, RMSD = 2.3 Å with 14 % identity over 63 aligned residues).

The heterodimeric DP2-E2F [92] binds DNA and mediates transcriptional regulation through the binding and release of ancillary proteins of the retinoblastoma (Rb) family and cyclin dependant kinases (Cdk). During gene suppression, Rb binds E2F and expression of the target gene is suppressed while in the late G1 phase Rb is phosphorylated by Cdk proteins, and released from E2F resulting in the expression of the target gene. Subsequently, E2F and DP2 are also phosphorylated resulting in their release from DNA and again a suppression of target gene expression. The affinities of the specific transcription factors for their target sequences on DNA were found to be dependent on interactions with effector molecules, which bound to them, or phosphorylation and other such post-transcriptional modifications [95]. Similar effects on the transcriptional activity of F112 should be considered as we search for the F112 target sequence. F112 is also similar in structure to transcriptional regulator SarR from *S.aureus* which displays dual functionality as an activator or repressor depending on the cellular environment. SarR acts as an activator and directly binds to the *hla* promoter thus increasing the expression of α hemolysin gene. Several regulatory mechanism have been put forth to explain the repressor like function of SarR [97]. SarR binds to the *sarA* P1 promoter region and by simple competitive displacement mechanism switches off the transcriptional activation of SarA activator protein. Another suggestion to explain its repressor activity is the formation of a heterodimer of SarR with SarA thus interfering with the activator function of the SarA homodimer. The structural similarities suggest that F112 may also function as a transcriptional regulator. F112, based on its time of expression in UV induced cells [44] might either act as an activator to initiate the transcription of the early T5 transcript or be a repressor of the UV induced T_{ind} transcript.

F112 also shows close structural similarity (SSM rmsd of 2.4 Å) to a Z-DNA binding human protein. Left-handed Z-DNA is found generally found in regions near the transcription start sites. Thus this further suggests a transcriptional activator like function for F112.

Along with the DNA binding domain, winged-helix proteins also commonly contain other domains involved in signal reception, activation or nucleic acid associated activity such as such as DNA repair and RNA metabolism. The HTH domain can also be adapted as a structural unit of a larger enzymatic domain or to mediate protein-protein interactions [95]. The missing C-terminal domain of F112 is too small to act as an enzymatic domain by itself; hence it might be involved in functions like oligomerization, signal reception or activation. The missing C-terminal residues from the current model are predicted to form an α-helix (~ 20 residues, Ile^{72} to Phe^{93}) and the remaining residues are predicted to be in a random coil by the protein secondary structure prediction algorithm *JPred* from the ExPASy Proteomics tools web-server [101]. Interestingly, a C-terminal helix is found at the heart of the dimer interface for SSV1 F93 and STIV F93, both of which are homodimeric wHTH proteins.

An unexpected intramolecular disulfide bond was observed in the crystal structure of F112. F112 is encoded by transcript T5 which also encodes for a number of apparent intracellular proteins that includes D355 (the viral integrase) [47, 102] and D63 [49], a structural homolog of ROP (an adaptor protein that regulates plasmid copy number) [103, 104] and another winged helix-turn-helix transcription factor F93 [51]. F112 is also most likely to be an intracellular protein as it is not found in the purified virion. Additional evidence that intracellular disulfide bonds are frequently used in the cellular proteins of

hyperthermophilic crenarchaeal viruses to provide thermostability is presented in the next chapter (Chapter 4) [105].

Conclusions

In conjunction with the previous studies of SSV1 D63 and F93 by Kraft et.al [49, 51] this work demonstrates the value of structural annotation in addressing the functions associated with these unique viral genomes. F112 is encoded by the early T5 transcript and hence may play a role as transcriptional regulator during the initial replication process of the SSV1. F112 is suggested to be a transcriptional regulator from this study while another protein F93 is suggested to be a transcription factor or replication terminator protein [51] based on structural homologies. Unlike F93, however F112 is not present among the other members of the *fuselloviridae* family and hence may play a role that is unique to SSV1, as opposed to a more general function relevant to all *fuselloviridae.* As SSV1 is the only UV-inducible *fuselloviridae* [44], F112 might play a regulatory role in the UV-specific induction of the viral genome.

F112 is a monomer in the crystal structure and a possible monomer in solution as evident from the DLS studies. However, transcriptional regulators are known to function as homo or hetero dimmers. The missing 40 residues at the C-terminus may be involved in the DNA induced dimerization or may interact with some other SSV1 or host protein to form a heterodimer. F112 was expressed as both N and C-terminal 6X His tag recombinant protein. Although good expression was seen for both constructs crystals suitable for structural analysis were only found for N-terminal His-tagged protein.

F112 could also interact with another viral or host protein and in-directly regulate or recruit additional proteins to the bound DNA. To identify binding partners pull-down assays with the whole cell lysate from infected cells could be employed. Immobilized recombinant His-tagged F112 can be used to identify other bound proteins that will elute along with the F112. Other affinity tags such as Glutathione-S-transferase (GST) or biotin are also commonly used and sometimes preferred over the His tags. Mass spectrometry or N-terminal sequencing could then be used to identify the binding partner. Frols et.al [44] suggested a possible activation of the T5 and T6 transcripts by the Tind encoded activator protein in UV induced SSV1 infected host cells. As F112 is one of the first T5 encoded ORFs to show an increase in expression on UV induction it is likely that B49, the Tind gene product, may interact with F112. However we have no preliminary data suggesting this to be true, hence this possibility should also be tested.

Structural homology strongly suggests that F112 interacts with DNA. In the current project preliminary non-specific interaction studies have already been carried out to confirm this possibility. However a biologically relevant binding site has yet to be identified. Chromatin Immunoprecipitations using microarrays **(**ChIP-CHIP) are one method that could be used to identify such a site. Similar studies to characterize intermolecular contacts between Ssh10b DNA-binding protein and DNA, in Sulfolobus shibatae, have been examined by Xue H.et.al. [106]. The ChIP-CHIP assay is an in-vivo experimental method which helps to determine whether a protein (e.g. transcription factor) binds to a particular region on the endogenous chromatin of living cells or tissues. The protein bound DNA complex is immunoprecipitated using antibodies against the target protein. The antibodies used for the ChIP-CHIP assay should be highly specific for

the protein of interest to minimize any non-specific interactions with other proteins. Although polyclonal antibodies have a higher rate of success in ChIP assays than monoclonals because polyclonal consist of a pool of antibodies against different epitopes, the batch-to-batch variation during the production of polyclonal antibodies could be a problem. Many companies are now selling “ChIP-grade” antibodies and this could be tried for F112 if the polyclonal antibodies donot work. Using high-throughput microarrays will help screen a large set of possible DNA binding sites. Optimization of the experimental conditions like the buffers used, the time of incubation, temperature at which reaction is carried out might be required. Once specific DNA binding has been established and the particular segment of DNA to which F112 binds to is identified, further DNA footprinting assays to determine the exact binding site could be tried.

CHAPTER 4

ROLE OF DISULFIDES IN HYPERTHERMOPHILIC PROTEIN STABILITY

Introduction

A unique clustering of the cysteine containing gene products in one fraction of the SSV1 genome was first noted by Palm et al. [20] where the cysteine containing proteins are encoded largely by the T5 and T6 transcripts [20]. This has led to the proposition that the genomes of *fuselloviridae* arose from a genome fusion event [20, 23]. Such an asymmetric distribution is also apparent in the genomes of other *fuselloviridae* [23, 45] and STIV [107]. The structure of SSV1 F112 [50], discussed in chapter 3, unexpectedly revealed the presence of an intramolecular disulfide bond. Intriguingly, structural studies on other proteins from our lab, SSV1 B129 and E96 (Eilers et.al. unpublished), STIV B116 [107], and STIV F93 [107] also revealed the presence of additional inter- or intramolecular disulfide bonds. These proteins including F112, are predicted to be DNA binding proteins (except E96, function unknown) and accordingly are presumed to be intracellular. Additional evidence for these proteins being intracellular includes gel-based mass spectrometry analysis of the STIV and SSV1viral particles and signal sequence analysis, both of which is discussed in greater detail in sections “Evidence of disulfide bonds from the crystal structure” and “Metagenome analysis” in this chapter. The consistent observation of intracellular disulfide is contradictory to the classical view, which states that the reducing environment of a cell is not very conducive to the formation of such disulfide bonds and hence, such interactions are generally restricted to

only extracellular or compartmentalized proteins. Combined, these observations prompted us to reinvestigate the distribution of cysteines in the SSV1 genome and subsequently, the available genomes of all other crenarchaeal viruses. Towards the goal of characterizing the cysteine distribution in the available crenarchaeal viral genomes, proteins from 18 different crenarchaeal viruses analyzed. In addition, these genomes were also pooled together and analyzed as a crenarchaeal viral metagenome. Strong evidence for an abundance of disulfide bonds in the cellular proteins of these viruses was obtained. This result further extends, the findings of Beeby and colleagues [108] of a widespread yet tightly controlled pattern of disulfide-bond utilization in intracellular proteins across thermophilic prokaryotes, to the viruses infecting some of these prokaryotes.

Results

Evidence of Disulfide Bonds from the Crystal Structure

The structure of F112 revealed the presence of a notable intramolecular disulfide bond. Clear electron density is observed between Cys^{51} in helix α3 and Cys^{58} in strand β1, with the 2.0 Å sulfur-sulfur distance typical of a disulfide (Figure 4-1). At first glance this might seem surprising, as F112 is predicted to be a DNA binding protein and hence is expected to be intracellular and environment not conducive to disulfide formation. However, the crystal structure of SSV1 B129, a zinc-finger DNA binding protein, and E96 also reveal the presence of a disulfide bond (Eilers et al. and Lawrence, unpublished). This observation has recently been extended to STIV [107], a hyperthermo-acidophilic virus isolated from Yellowstone National Park [37]. Structural studies on STIV carried out by Dr. Larson [58] in our lab also identified the presence of

an intramolecular disulfide bond in B116 which is a predicted DNA binding protein [107] and an intermolecular disulfide bond in F93 which is a winged helix DNA binding protein [98].

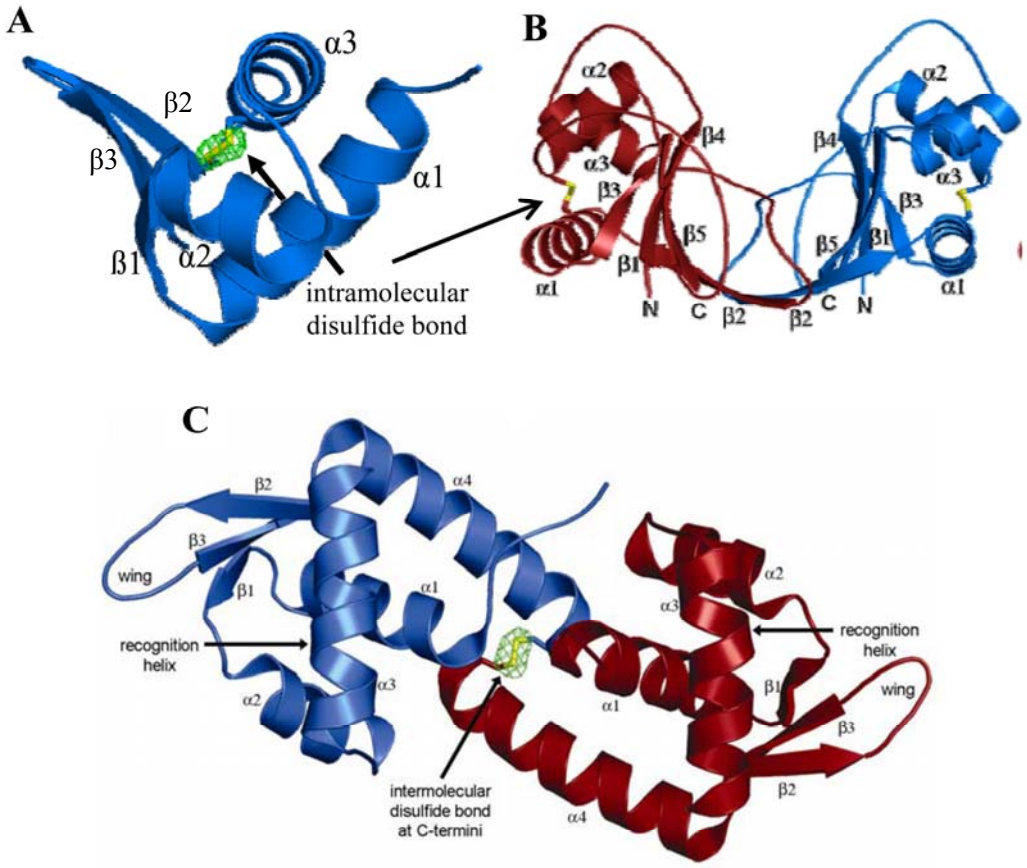


Figure 4-1. Structure of SSV1 F112 [50], STIV B116 [107] and STIV F93[98].
Cartoon representation of F112, B116 and F93 with the secondary structural elements labeled as in Fig 2B [50], Fig 1A [107], Fig1B [98]. The disulfide bond is depicted as a stick model. For SSV1 F112 and STIV F93 the disulfide stick representation is also surrounded by the green mesh of difference electron density (contoured at 5σ) in which the sulfur atoms were omitted from model refinement.

Further, biochemical and genomic studies also suggest that disulfides are in fact very common in the intracellular proteins of many hyperthermophilic organisms, where their frequencies correlate with increased growth temperatures, suggestive of their

contribution to protein thermostability [108, 109]. Thus, the disulfide bonds observed in these putative intracellular hyperthermoacidophilic viral proteins (Figure 4-1) may be physiologically relevant and might contribute to the thermostability of these proteins. Alternatively, these proteins might reside within the viral particle. The latter possibility has been addressed by the Bothner and Young laboratories at Montana State University. They analyzed the purified SSV1 and STIV viral particles by mass spectrometry [50, 98] and found no evidence for the incorporation of F112, B129 or E96 into SSV1, nor incorporation of F93 or B116 into STIV. While this does not prove their absence from the viral particle, it is consistent with an intracellular distribution.

Disulfide Conferred Thermostability of Proteins

F112 was used as a model to study the effect of the disulfide bond on the thermostability of the protein. Two different techniques were employed to follow the change in the melting temperature of F112 +/- reductant. Tris [2-carboxyethyl] phosphine (TCEP) at a final concentration of 5mM was used as the reducing agent in both experiments to break the disulfide bonds. In the first study, the ability of the disulfide bond to contribute to the thermostability was followed using fluorescence emission spectroscopy. This method involves the use of an environmentally sensitive dye (e.g. SYPRO Orange or ANS) as an indicator of protein melting. As a result of the conformational change induced in the protein at higher temperatures an increase in the fluorescence was observed as the SYPRO orange dye binds to the molten globule like state of the protein. Thermal melt assays of F112 were done both in the absence and presence of (inflection point of the curve) in the absence of a reducing agent, however in the presence

TCEP (Figure 4-2B). The melting temperature (T_m) for F112 is found to be at 87 ^oC of TCEP, the T_m is reduced to 76 ^oC, a decrease of 11 ^oC.

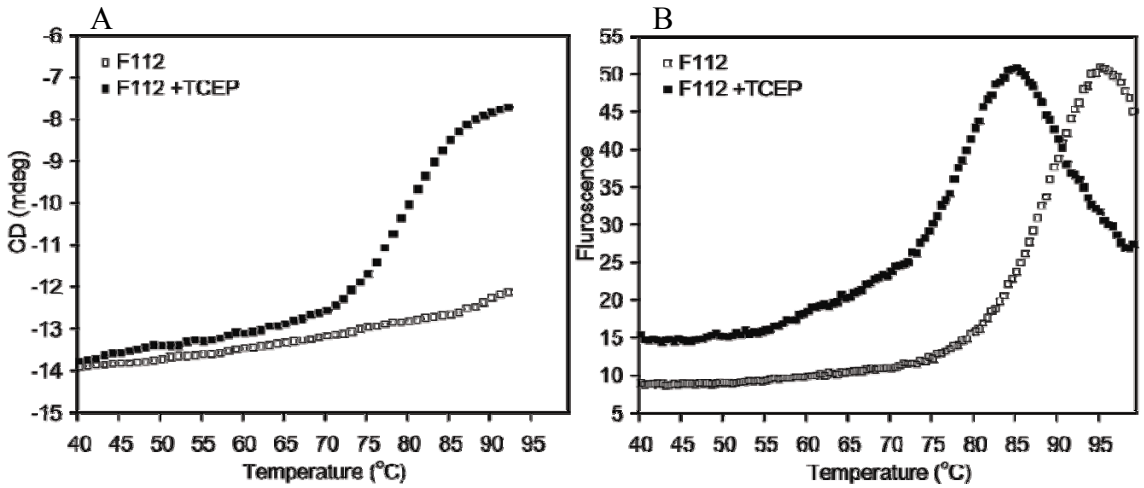


Figure 4-2. Stability of F112 +/- TCEP.
Thermal melt assays to study the effect of disulfide on the melting temperature (Tm) of F112 was followed using far UV CD spectroscopy and fluorescence emission spectroscopy. (A) The change in ellipticity at 222nm was measured as a function of temperature and a melting point (T_m) was obtained from the inflection point of the spectra. The T_m for F112 in the absence of TCEP (□) could not be determined but in the presence of 5 mM TCEP, the T_m falls to 78^{o}C (■), a decrease of more than 14^{o}C. (B) For the fluorescence emission spectroscopy SYPRO orange was used as a reporter dye which apparently binds to the exposed hydrophobic residues in the protein. As a result an increase in the fluorescence signal was observed as a function of protein denaturation. The Tm for each sample was obtained from the inflection point of the melt curves. Tm for F112 (□) was observed to be around 87^{o}C while in the presence of a reducing agent like TCEP, a 11^{o}C decrease in the Tm was measured at 76^{o}C (■).

As a confirmative study, we used a second technique to look at thermal melts of F112. Thermal melts from 40 to 92 °C were followed as a function of ellipticity at 222 nm using far Ultra Violet – Circular Dichroism (far UV-CD) spectroscopy (Figure 4-2A). This technique measures changes in the secondary structure of a protein with an increase in temperature. The CD spectroscopy experiments were performed by Dr. Stan Kwok at University of Colorado - Fitzsimmons, Aurora, CO. We were unable to measure a melting temperature for F112 (T_m > 90 ^oC) over the temperature range of the instrument (40 ^oC – 90 ^oC) in the absence of TCEP, presumably due to protein thermostability. However, in the

presence of 5mM TCEP the melting temperature decreased to 78 ^{o}C (inflection point of the thermal melt curve), suggesting a ΔT_m greater than 14 ^{o}C. Initial wavelength scans at 40 ^{o}C for the protein +/- TCEP have super imposable spectra indicating well folded protein in both cases (Figure 4-3).

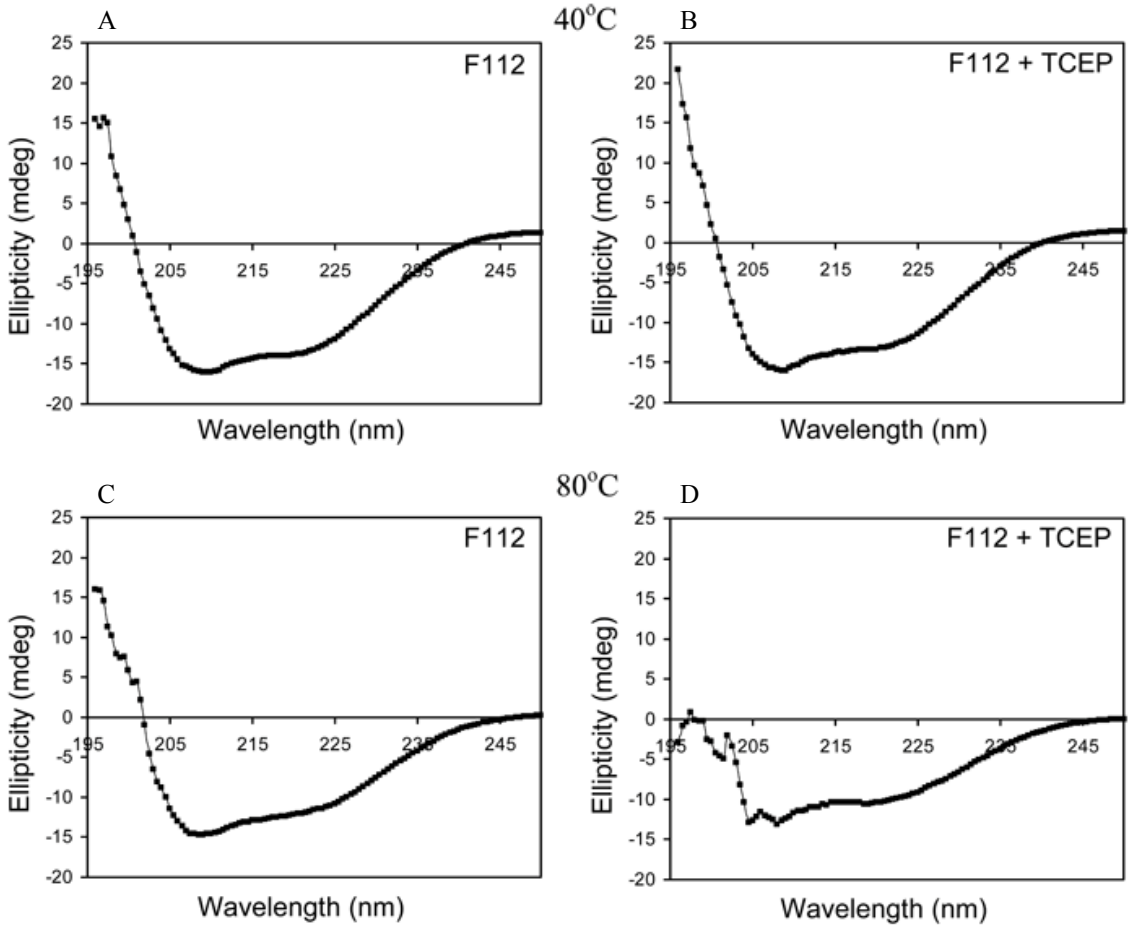


Figure 4-3: Far-UV CD spectra +/- TCEP at different temperatures.
Nearly identical spectra are seen for F112 +/- 5mM TCEP at 40 ^{o}C (panels A and B). At 80 ^{o}C (panels C and D), there is a pronounced loss of secondary structure in the presence of TCEP (panel D).

Even though, the individual temperatures obtained as T_m from both methods are slightly different, the overall significant loss in thermostability seen for F112 in the presence of TCEP (ΔT_m of 11 ^{o}C using fluorescence emission and ΔT_m of 14 ^{o}C using far

UV-CD spectroscopy) are in agreement. This suggests that at least in-vitro the disulfides play a key role in the stabilization of these thermophilic proteins at high temperatures. The magnitude of the stabilization *in vivo* is an open question, and will certainly depend upon the redox potential of the intracellular environment. To our knowledge, little is known regarding the identity of the major small molecule thiols or the intracellular redox potential in *Sulfolobus solfataricus*.

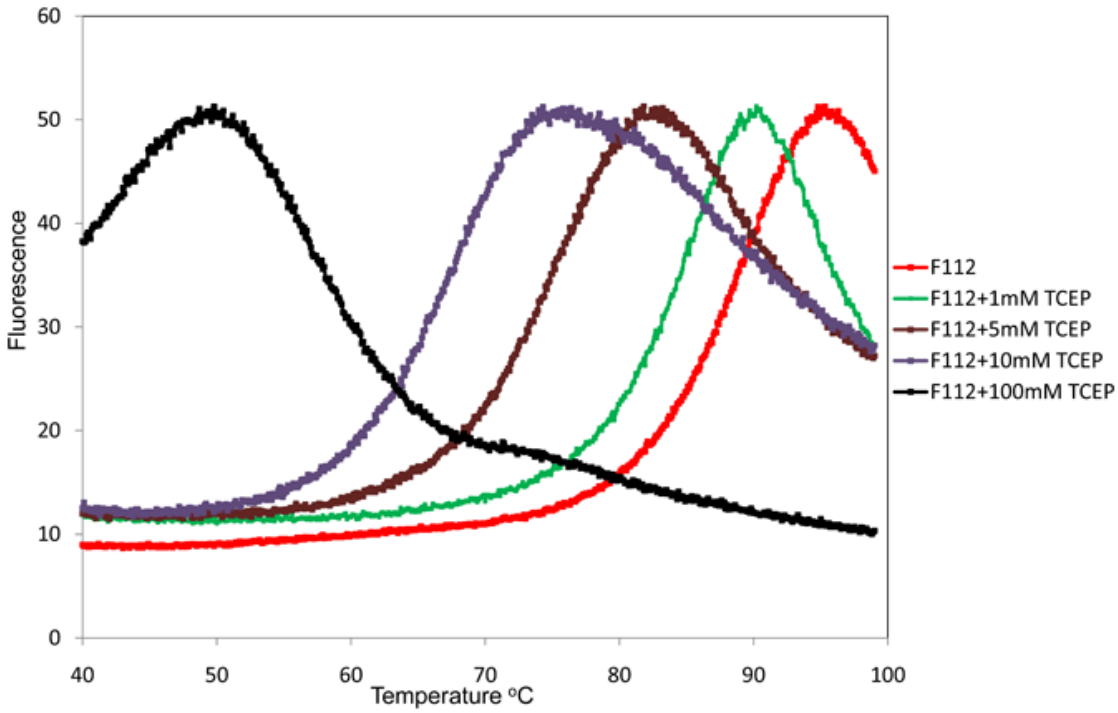


Figure 4-4: Stability of F112 in the presence of various TCEP concentrations.
Thermal melt assays to study the effect of various concentrations of reducing agent (TCEP) on the disulfide conferred thermostability was done using a fluorescence dye (SYPRO orange). A decrease in the melting temperature indicative of reduced thermostability is observed as the reducing agent concentrations are increased (from 1mM to 100mM TCEP). A complete denaturation in the presence of very high concentration of TCEP is seen even at very low temperatures (Tm ~ 45^{o}C).

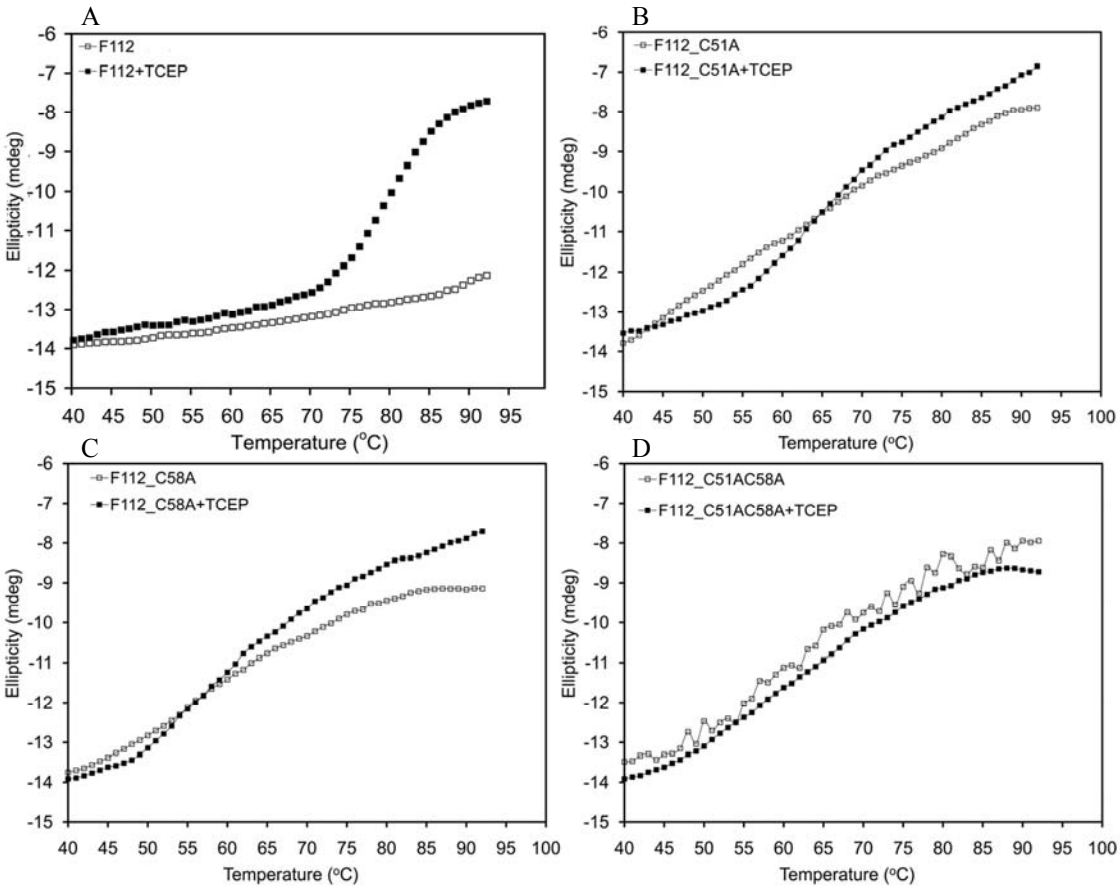


Figure 4-5. Stability of F112 and its mutants +/- TCEP.
Thermal melt assays to study the effect of disulfide on the melting temperature (Tm) of F112 and its mutants was followed using far UV CD spectroscopy. The change in ellipticity at 222nm was measured as a function of temperature and a melting point (T_m) was obtained from the inflection point of the spectra. The T_m for F112, F112_C51A, F112_C58A, F112_C51AC58A in the absence of TCEP (□) and in the presence of 5mM TCEP (■) are shown in panels A, B, C, D respectively. Panel A is same as Figure 4-2A while panels B, C, and D indicates a non-cooperative melting of the mutants.

However, even in a reducing environment, disulfide bonds can be quite stable [110]. Further, the stability of the disulfide and the stability of the protein are directly correlated [110]. Thus, a particularly stable disulfide is expected to confer protein thermostability, even in a reducing environment. Similar effects were seen for F112, where

increasing concentrations of the reductant TCEP decreased the ability of the protein to with stand thermal denaturation (Figure 4-4).

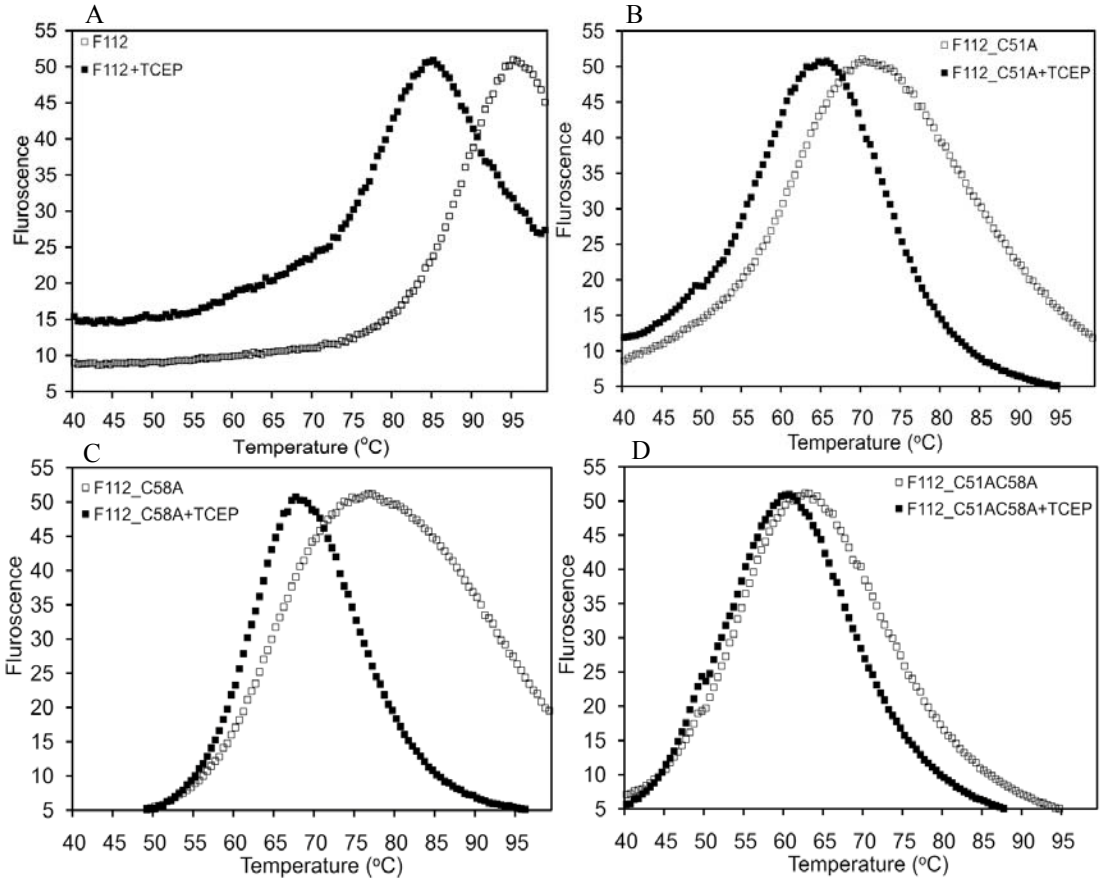


Figure 4-6. Stability of F112 and its mutants +/- TCEP.
Thermal melt assays to study the effect of disulfide on the melting temperature (Tm) of F112 and its Cys mutants was followed using fluorescence dye. SYPRO orange dye was used as a reporter dye. An increase in the fluorescence signal was observed as more dye bound to the hydrophobic residues as a function of protein denaturation. The T_m for F112, F112_C51A, F112_C58A and F112_C51AC58A in the absence of TCEP (□) and in the presence of 5mM TCEP (■) is shown in panels A, B, C, D respectively. Panel A is same as Figure 4-2B while panels B, C, and D indicates the melting of the mutants. The reduction in the T_m for the mutants indicates that these protein constructs are far less stable as compared to the wild-type protein.

Mutational Studies

The structure of F112 was the first protein structure where the presence of a disulfide bond was documented from SSV1. As the expression and purification of this

protein is fairly straightforward (see section of "Materials and methods for F112" in chapter 2) it served as the good candidate to examine the contribution of cysteines and disulfide bonding for thermostability of the protein using site-directed mutagenesis.

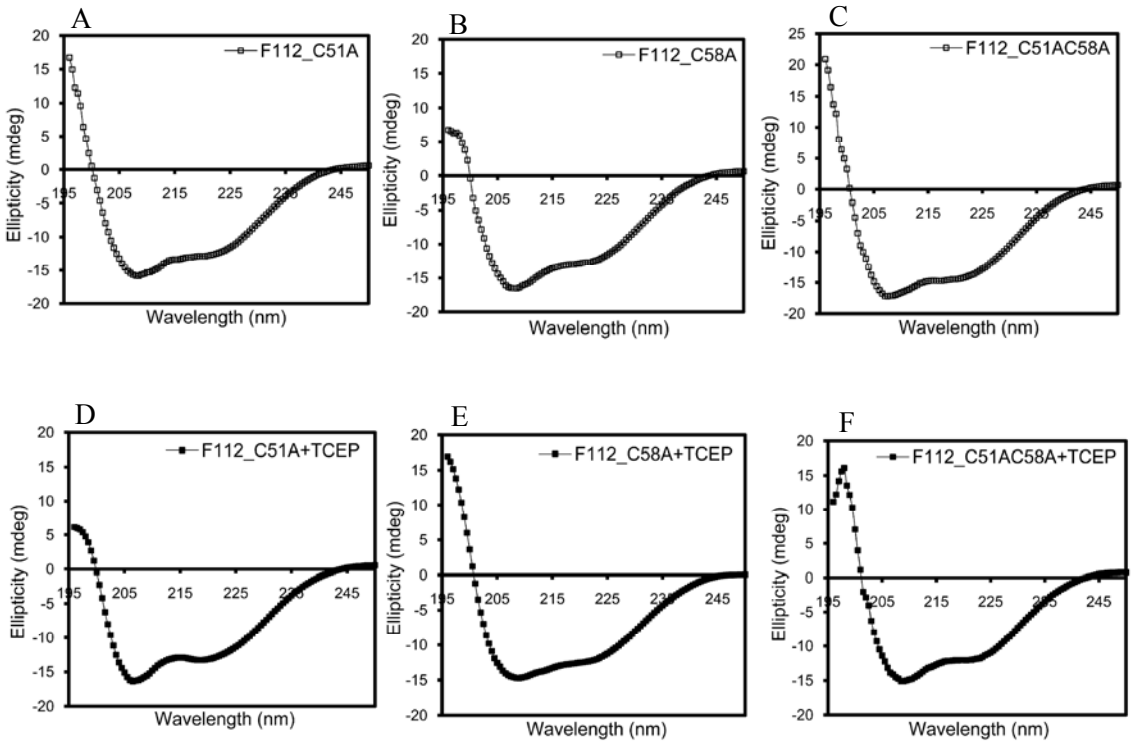


Figure 4-7: Far-UV CD spectra F112 mutants +/- TCEP at 40 ^{o}C.
Predominantly α-helical secondary structures are seen for all three mutants F112_C51A, F112_C58A, F112_C51AC58A (panels A, B, and C) at 40^{o}C similar to the wild-type F112 (Figure 4-3 A and B). The core structure is maintained in all three mutants at this lower temperature even after the addition of a reducing agent such as TCEP (panels D, E, and F).

To study this effect, single point variants targeting Cys^{51} and Cys^{58} as well as a double mutant where both the cysteines were substituted were prepared. The cysteine residues were mutated to either serines (F112_C51S, F112_C58S and F112_C51SC58S) or alanines (F112_C51A, F112_C58A and F112_C51AC58A). Though some efforts were made, we were unable to construct the C51SC58S mutant. For the other mutants, purification conditions for the resulting variants were similar to that for wild-type F112.

The yield of F112 variants were generally between 2-4 mg per gram of cell pellet. However, only the alanine mutants (F112_C51A, F112_C58A and F112_C51AC58A) were used for the subsequent thermal melt studies.

Far ultraviolet CD revealed a predominantly α-helical secondary structure at 40 ^{o}C and gave super imposable wavelength scans for both the wild-type F112 and the three alanine mutants in the presence and absence of a reducing agent (Figures 4-6 and 4-7). The unfolding for all three Cys to Ala mutants using far UV-CD spectroscopy experiments, was found to be non-cooperative rather than cooperative giving linear rather than sigmoidal curves. In addition, a T_m could not be established by CD spectroscopy as aggregation of the mutant proteins were seen at higher temperatures. However by fluorescence emission spectroscopy (SYPRO orange test) a drop in the melting temperature for the mutants was seen where the T_m for the mutants were as follows, F112_C51A ~ 63^{o}C, F112_C58A ~ 63^{o}C and F112_C51AC58A ~ 54^{o}C (Figure 4-6).

The disulfide bond formed between Cys^{51} and Cys^{58} is responsible for stabilizing a tight loop between the α3 and β2 secondary structural elements of F112. Elimination of the disulfide bond in the case of the three F112 variants results in a less stable loop which leads to unfolding of the protein with even a small increase in temperature (Figure 4-6). Thus even though the core structure of the mutants are still maintained as shown by the CD spectra at low temperature (Figure 4-7) the variants are far less thermostable as compared to the wild type protein (ΔT_m ~24 ^{o}C for the single mutants and ~33 ^{o}C for the double mutant). The increased loss of stability seen in case of the cysteine double mutant as compared to the single mutant could be a result of the disruption of favorable interactions of the cysteine residues with the surrounding amino acids (Figure 4-6, panels B, C and D).

Metagenome Analysis

Consistent with the frequent occurrence of intracellular disulfides, the small intracellular proteins in *P.aerophilum* are found to favor even numbers of cysteine residues [109]. A similar observation has been made for the intracellular proteome of STIV [111]. However, in contrast to *P. aerophilum* and STIV, we find that the predicted intracellular SSV1 proteome does not show a pronounced preference for proteins with an even number of cysteine residues. We believe this to be analogous to the cysteine distributions in 9 thermophilic bacterial and archaeal organisms where sequence-structure mapping indicates an abundance of intracellular disulfides [108, 109], yet only those intracellular proteins of *P. aerophilum* which were 100-200 amino acids in size, have been shown to exhibit a distinct preference for even numbers of cysteine residues [109]. Our structural studies on the proteins from SSV1 and STIV provide further evidence to suggest that disulfides might also be common to the intracellular proteomes of thermophilic viruses in general. Further, proteomic analysis of a metagenome composed of the predicted intracellular proteins from 18 hyperthermophilic viral genomes [19-21, 23, 25-31, 34, 35, 37, 38, 45, 75] exhibits a similar pronounced abundance of proteins with even numbers of cysteine residues. Putative false positives were removed from the metagenome so as to enrich the metagenome with only potential intracellular proteins. To prevent false positives, the proteins with a detectable signal sequences as determined by the program SignalP [76], putative integral membrane proteins which were recognized using the program TMHMM [77, 78], proteins with a predicted metal binding motif which were identified using the primary literature [16] or the program ScanProsite [79-81], and the extracellular proteins identified by the analysis of the viral particles were

removed. A pronounced saw-tooth distribution due to the abundance of proteins with even numbers of cysteine residues (Figure 4-8) is obtained and a clear preference was established for intracellular proteins with even numbers of cysteine residues.

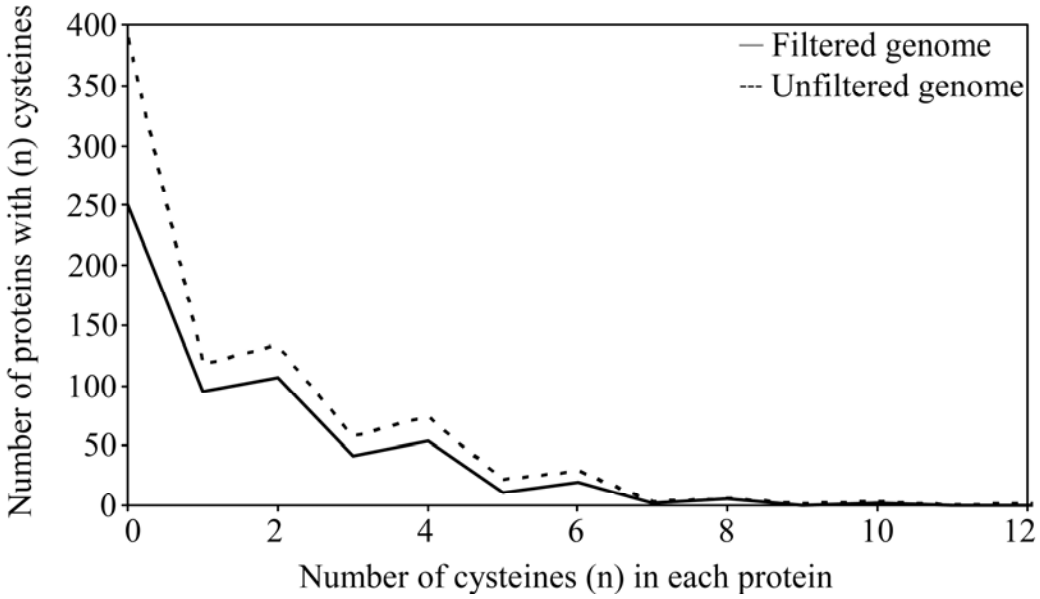


Figure 4-8. Distribution of cysteine in the hyperthermophilic viral metagenome.
Eighteen crenarchaeal viral genomes were combined to create the viral metagenome. False positive were minimized by filtering out known or predicted extracellular proteins, membrane proteins, and proteins exhibiting metal binding motifs, resulting in a "filtered" genome enriched for intracellular proteins. A clear preference for even numbers of cysteine residues in the predicted intracellular proteins of the metagenome is clearly seen (solid line), suggesting an abundance of intracellular disulfide bonds. The trend is also apparent in the entire viral metagenome (dashed line).

Similar distributions are seen for individual viruses from a variety of crenarchaeal viral families (Figure 4-9). These include TTSV1 [34] and PSV [35] (g*lobuloviridae)*, SIRV2 [112] (*rudiviridae)*, AFV1 [31] (*lipothrixviridae)* and STIV [98]. However, odd numbered cysteines can also participate in the formation of intermolecular disulfide bond as seen in the case of STIV F93 and such possibilities should not be ignored.

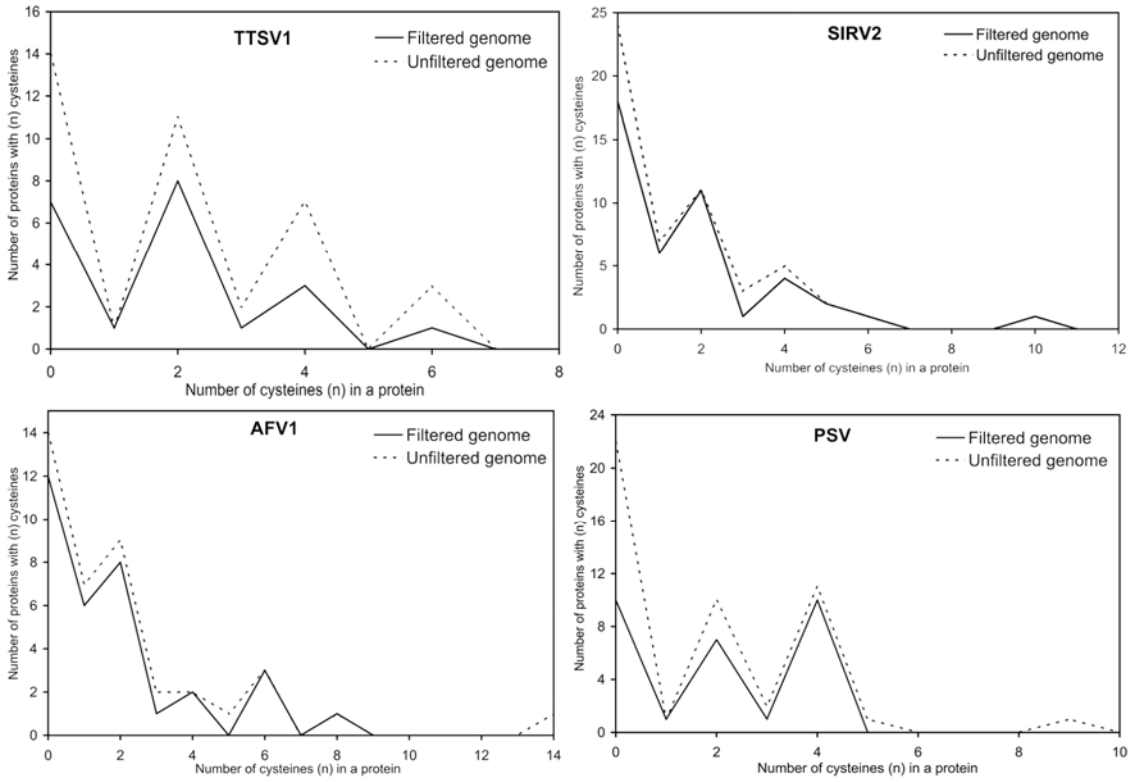


Figure 4-9. Analysis of the cysteines in the TTSV1, SIRV2, AFV1 and PSV genomes. The graph depicts number of proteins containing 'n' number of cysteines in the assorted genomes. The solid line corresponds to the filtered genome (enriched for intracellular proteins) and the dashed line represents the entire genome (unfiltered). The plot shows a clear preference for even number of cysteines as compared to odd, suggesting that these genomes might be abundant in intracellular disulfide bonds [50].

Cysteine in Virally Encoded Membrane Proteins

While examining cysteine distributions in hyperthermophilic viral genomes, we noticed a paucity of cysteine residues in the predicted membrane proteins [77, 78] of many viruses infecting *Sulfolobales*. For example, the predicted membrane proteins in STIV C67, B130, A55, C92, B174 and C557 lack cysteine altogether, while the non-membrane STIV proteins show a cysteine content of 0.95% [37]. Similarly, the cysteine content in predicted SSV1 membrane proteins is 0.08%, while the cysteine frequency in non-membrane proteins is 1.44%. This dichotomy in the cysteine content of membrane

versus non-membrane proteins is also apparent in the genomes of SSV2 [23] (0.09% membrane, 0.89% non-membrane), SSV3 (Stedman et.al., unpublished) (0.18%,0.98%), SSVK1 [22] (0.19%, 1.02%), SSVRH [22] (0.12%, 1.07%), SIRV1 [25] (0.10%, 0.87%), SIRV2 [25] (0.38% membrane, 0.96%), and STSV1 [38] (0.11%, 0.60%). In fact, the only outlier among viruses infecting *Sulfolobales* is SIFV [28], where the cysteine content in predicted membrane proteins is 1.77%, and that in the non-membrane proteins is 1.13%. with the exception of Acidianus Two-tailed Virus [32] (ATV, 0.31%, 0.63%), and Pyrobaculum Spherical Virus [35] (PSV, 0.50%, 1.37%), the reduced cysteine content in predicted membrane proteins appears to be restricted to viruses infecting the *Sulfolobales*.

Discussion

Intracellular Disulfides

In contrast to the intracellular proteomes of mesophilic organisms, biochemical and genomic evidence, including sequence structure mapping, strongly suggest that intracellular disulfides are common in thermophilic genomes [108, 109]. Interestingly, in *P. aerophilum*, where sequence-structure mapping gives an extremely high cysteine-cysteine proximity score, proteins that are 100-200 residues in size show a clear preference for even numbers of cysteine residues [109]. Similarly, the intracellular proteome of STIV also favors proteins with even numbers of cysteines, and two available structures confirm the presence of disulfides [107]. The observation here that the two cysteines in F112 form a disulfide bond, and the observation of similar disulfides in SSV1 B129 and E96, suggests that intracellular disulfides are also common to the

intracellular SSV1 proteome [108, 109]. Combined with structural work on STIV and the metagenomic studies, it strongly suggests that intracellular disulfides may be common to hyperthermophilic viruses in general. Indeed, a preference for intracellular proteins with an even number of cysteines is seen in at least 3 additional viral families, and is not limited to viruses infecting *Sulfolobales*. These viral families include the r*udiviridae* (SIRV2), the *guttaviridae* (AFV1), and the g*lobuloviridae* (PSV and TTSV1) (Figure 4-9), but not in SSV1 or other *fuselloviridae*. However, as discussed above, only genomes with the highest cysteine-cysteine proximity scores (> 40%) are expected to show a preference for an even number of cysteines residues in their intracellular proteome [108, 109]. Thus as seen for F112, genomes that lack a clear preference for proteins with an even number of cysteines are still likely to contain protein with disulfide bonds. In this light, the clear preference for proteins with even numbers of cysteines in STIV, SIRV2, AFV1, PSV, TTSV1 and the crenarchaeal metagenome are remarkable. Not a single intracellular prokaryotic genome has been shown to contain this pattern.

We believe the physical basis for the prevalence of even numbers of cysteines in the intracellular proteins of thermophilic viruses is two-fold. First, if disulfide bonds confer a selective advantage, i.e., through thermostability, one would expect evolutionary pressure to maintain cysteine residues in pairs. Second, the presence of an intracellular protein disulfide oxidoreductase (PDO) in a select group of thermophiles, including *Sulfolobus solfataricus*, the host for SSV1 and STIV, suggests they have cytoplasmic activities that catalyze disulfide bond formation [108]. Thus, unpaired, surface exposed cysteines might be selected against, as they could form inappropriate intermolecular disulfides, leading to nonspecific aggregation and loss of activity. Importantly, though,

while a preference for an even number of cysteine residues at the genomic level may indicate intracellular disulfides are common [107, 109], it does not imply the presence of disulfide bonds in a specific protein, even one that contains an even number of cysteines. This is in direct contrast to sequence-structure mapping which can make predictions for a specific protein [113]. Thus, significant numbers of cysteines in the intracellular proteome may remain as the free thiol, or mediate interactions with metals. In addition, proteins with only a single cysteine may form intermolecular disulfide bonds, as observed in STIV F93 [107].

The eukaryotic Pox viruses are large dsDNA viruses that replicate within the cytoplasm of the infected cells. Interestingly, cytosolic disulfide bond formation is a critical step in Pox virus maturation and generally the machinery required to carry out this task is encoded by the virus itself [114-120]. Similar machinery also appears to be also present in the genomes of the related asfarviruses, iridoviruses and phycodnaviruses [121]. However, like SSV1, the pox virus genome does not show a preference for even numbers of cysteine residues. Thus, formation of disulfide bonds in the cytosol of the host may represent yet another distant similarity between crenarchaeal viruses and the large double-stranded DNA viruses of eukaryotes.

Protein Disulfide Oxidoreductase (PDO) is a key enzyme involved in the intracellular shuffling of the disulfide bonds [122]. Interestingly, the presence of cytosolic PDO has been identified in numerous hyperthermophiles including the *S.solfataricus*. More recently, it has been reported that PDOs are exclusively found only in thermophilic organisms with highest predicted intracellular disulfide content [123]. Because viruses generally encode a minimum set of proteins required for propagation and

rely heavily on the host machinery to provide for its other needs, it is likely that thermophilic viruses will utilize the host PDO whenever possible in order to catalyze disulfide bond formation in the viral proteins. However, the TTSV1 genome which shows a clear preference for proteins with even number of cysteines (Figure 4-9), is predicted to encode for its own thiol-disulfide isomerase [16]. Interestingly, *Thermoproteus tenax,* the TTSV1 host is a hyperthermophilic sulfur-reducing anaerobe that does not exhibit an abundance of disulfide bonds in its predicted intracellular proteome and hence most likely will not code for a PDO like protein. This suggests that TTSV1 may carry its own thiol-disulfide isomerase in order to catalyze the formation of viral disulfides in the absence of the host enzyme, a process that is directly analogous to that seen for eukaryotic pox virus.

Genome Organization and Viral Evolution

Most viruses infecting *Sulfolobales* are thought to be enveloped viruses (exceptions are SIRV1 and SIRV2). Thus, it is not surprising that many proteins identified in the SSV1 and STIV viral particles are thought to be membrane proteins. Interestingly, they are found to cluster in the cysteine poor regions of the viral genomes. In SSV1 (Figure 4-10), this clustering appears to mirror the polycistronic transcripts and the operon-like organization of the SSV1 genome [20]. Thus, consistent with the locations and cysteine content of membrane proteins that copurify with the virus particle (VP1, VP3 and C792), the cysteine poor regions in SSV1 are likely to contain viral coat proteins, proteins that direct late stages of viral assembly at the membrane, and membrane proteins involved in viral egress. In contrast, the cysteine rich regions encode

the viral integrase (D335) [47, 102], the ROP-like D63 [49], winged helix proteins such as F93 [51] and F112 [50], and zinc-finger proteins such as B129 [Eilers and Lawrence, unpublished, [16]], suggesting that cysteine rich regions of the genome are likely to encode cellular proteins involved in virus multiplication or lysogeny. Consequently, the cysteine rich regions are likely to encode early gene products, while the cysteine poor regions are more likely to represent late gene products [20, 44], consistent with the microarray work of Frols et al. [44]. Thus, one need not postulate a genome fusion event [20, 23] to understand the clustering of cysteine containing ORFs in SSV1 [20] and other *fuselloviridae* [23, 45]. Further, in viruses infecting *Sulfolobales* (STIV for example), the presence of cysteine rich and cysteine poor sectors can provide valuable insight into the functional organization of the viral genome and suggests which operons encode for the cellular viral machinery (cysteine rich) and which encodes for the viral structural proteins (cysteine poor).

These observations are in agreement with the pioneering studies by the Yeates group at UCLA [108, 109] on the effect of disulfide bond formation on protein stabilization in thermophiles. While the work of Yeates and colleagues rises to the genomic level several other structural and biochemical studies also provide evidence for the existence of stabilizing disulfide bonds in specific thermophilic proteins. A DNA glycosylase [124], an adenylosuccinate lyase [125] and a PaPDO from *P. aerophilum* [108] along with a hypothetical protein [ST1625p] from *Sulfolobus tokodaii* [126], and a 5'-deoxy-5'-methylthioadenosine phosphorylase from Sulfolobus tokodaii (To be published) are all examples of intracellular proteins with disulfide bonds. In each case the presence of a reducing agent has been found to significantly affect the thermostability of these proteins

[125, 127-132]. Hence these structural and genomic studies from several other laboratories provide substantial evidence for an abundance of intracellular disulfide bonds in hyperthermophilic organisms and our observation extends this concept to the hyperthermophilic viruses.

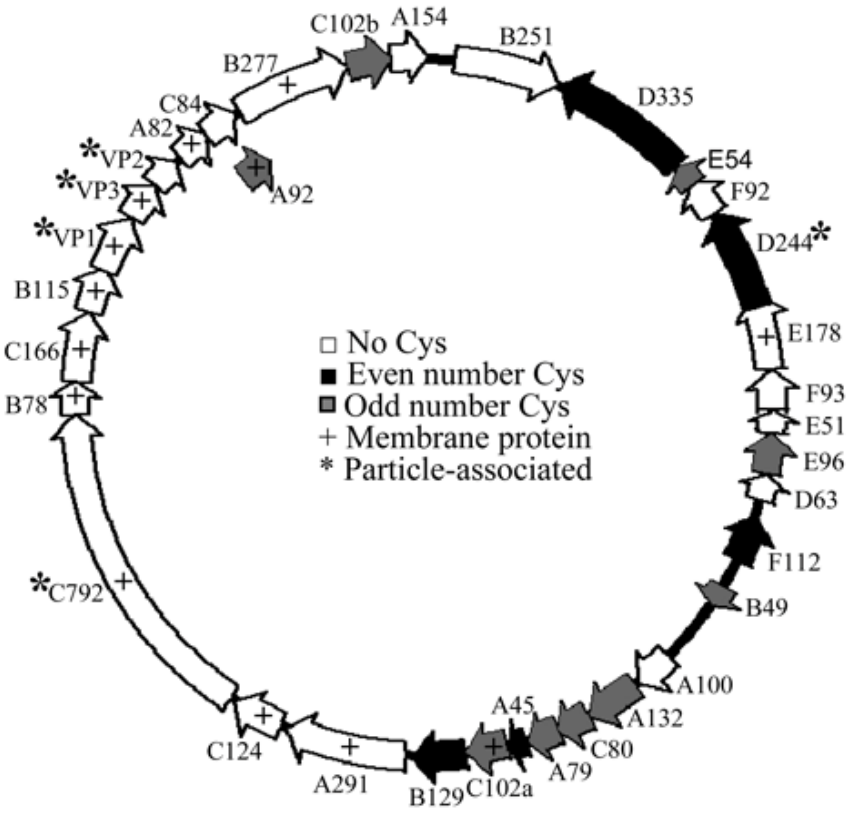


Figure 4-10. The SSV1 cysteine distribution genome map

Open reading frames that lack cysteine codons are in white, those with an even number of cysteine codons are in black, and those with an odd number of cysteine codons are in grey. Open reading frames coding for predicted membrane proteins are marked with a plus sign (+) while those coding for particle associated proteins are labeled with an asterisk (*). The right half of the viral genome is cysteine rich and encodes an abundance of putative intracellular proteins. In contrast, the left side is cysteine poor and contains many of the predicted membrane proteins (+). The map was made using Vector NTI Advance 10.1.1 (Invitrogen).

Conclusions

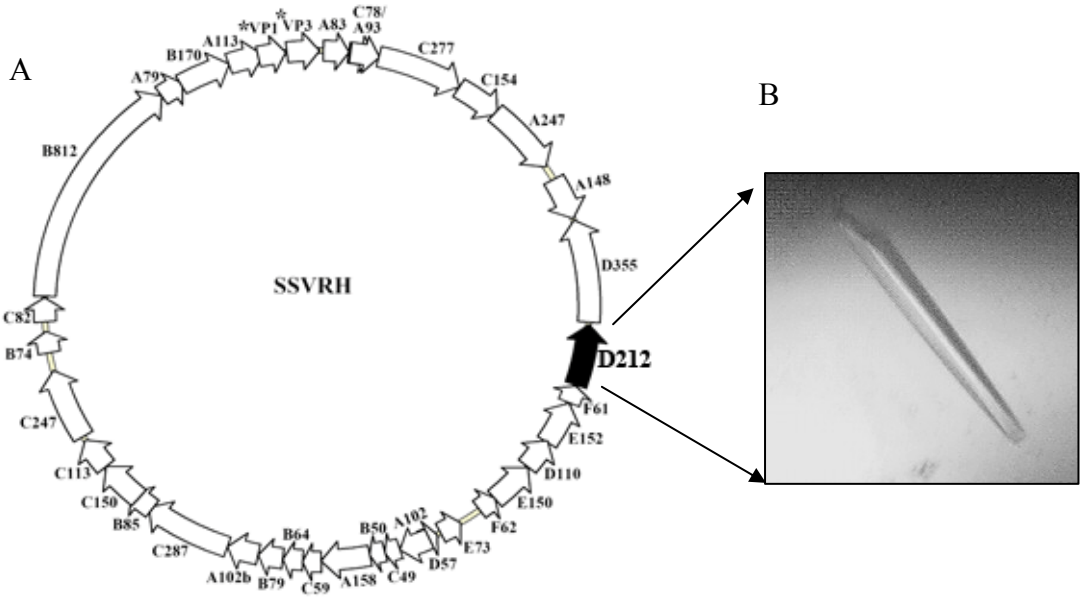
Hyperthermophilic viruses like SSV1 and STIV, along with their hosts, must be able to survive in extreme environments. In doing so, these organisms must not only be able to adapt, but also exploit these rigorous conditions to their advantage [133]. One method to identify factors influencing such adaptability and stability in hyper-thermophilic proteins is to compare structures of homologous proteins from thermophiles and mesophiles. Such comparative studies have identified several factors involved in the stabilization of these thermophilic proteins, which includes, tighter local packing, shortening or deletion of flexible loops, more extensive protein surface with fewer cavities, better hydrogen bonding, stronger salt bridges, increased helical content, reduced occurrence of thermolabile residues and disulfide bonds [102, 133-137]. Our structural work on the SSV1 and STIV proteins [50, 98, 107], mutational studies on F112 and the analysis of the crenarchaeal viral metagenome provides a remarkable demonstration of the importance of the role of disulfides in protein thermostability and their common occurrence in thermophilic viruses.

CHAPTER 5

CRYSTAL STRUCTURE OF SSVRH D212; A POSSIBLE FUSELLOVIRAL NUCLEASE.

Introduction

Open Reading Frame D212 in Sulfolobus Spindle-shaped Virus Ragged Hills (SSVRH) encodes a 212 amino acid gene product in reading frame D of the SSVRH genome [22]. Although similarities to hypothetical proteins from other archaeal viruses are often found, the lack of statistically significant similarity to protein sequences with known function has impeded functional annotation. D212 has well conserved homologs in most members of the *Fuselloviridae* reported to date [45]. This includes D244 from SSV1 with 79% sequence identity, E211 from SSV2 with 81% identity, ORF209 from SSV4 which shows 82% sequence identity and the homolog from SSV5 ORF214 which is 39% identical (Figure 5-1). Recent mass spectrometric analysis of the SSV1 viral particle has identified C792 and D244 as co-purifying with the SSV1 particle, along with the previously recognized particle associated proteins VP1, VP2 and VP3 [43, 50]. While, VP1, VP3 and C792 are predicted to be membrane proteins by TMHMM [77, 78], D244/D212 on the other hand is predicted to be a soluble protein. However, no functional predictions were made for D244 or its homologs [50] until this study.




```
SSVRH D212 ---------------------------------MTETD-FKLKYWGNQEQD-YILPTVGLGREYLVLGKLLISLSKWRAKGLIDFDVYLRPTGVGTLTNTINYEY 26
SSV1  D244 MDCNRNCYCKCVICNSISTDSHTSPRFNMVSEMAETN-FKLKYWGNQAED-YILPSTYLGREYLVFGKLLISLSKWRAKGLLDFDVYIRPTGVGTLTNVINEQY 58
SSV2   211 ------------------------------------MAD-FKLKYWGNQTED-YILPNFYLGREYLVLGKLLIKLAQWRVKGLIDFDVFVKITGVGTLTNVINEHY 24
SSV4   209 ------------------------------------MAD-FKLKYWGDQSED-YILPTTWLGREYLVLGKLIIKLAKWRKY--LDFDVYIRRSGVGTLTNTINYEY 24
SSV5   214 ---------------------------------MSETNSFKLKYWGDNPEG-LLKINPYTGRELEFLGKLMKKLS-----LHVNYDLYIMPSGIGTLDTSLLPNY 27
                                              *******             ***    ***             *     * **        *

SSVRH D212 YKGLEDKYDLTLYIRAK----DSYYPLLWIDITGSSWTEEQSKERYGESIYAILSVKVETAKKYDVLGRVFFIHYNDTEDKLKCISALQILNLERQNKIKKDK 86
SSV1  D244 YSGLQDKYDLTLYVKAK----TEYYPLIWIDITGSSWTEEQSKERYGESVYAILSTKVEVAKQNEVMGRVWFIHYSDAEDKLKCISALQILNLEKQGKIKKDK 118
SSV2   211 YEGLQDKYDLTLYVKAK----DSYYPLVWIDITGSSWTEEQSKERYGESVYAILSAKVEVAIKHDVMGRVWFIHYNDTEDKLKCISALQILNLERQGKIKKDK 84
SSV4   209 YKGLEDKYDLTLYVKAK----QQYYPLIWLDITGSSWTEEQSKERYGESVYAILSIKLEVAKKYDVMGRVWFIHYNDTEDKLKCINALQILNLERENKIKKDK 82
SSV5   214 YHDISDKYDFTIFTLVNLNGKKVKWPLVYIDVSGSKYSDKESEKRYGKPVDAILKTKVDLAFKYNVAGRVYFVLSIDKEDKFRVLTALQVMNLYKKGIAKLDS 86
           *    **** *                ** * **      * ***    *** * * * *** *    * ***     ***  **      * *

SSVRH D212 FERGAKSEYYLIPTSYWKNLTELRIALRGFYQSF-KEYLAR-SNK 212
SSV1  D244 FERDAVSYYYLIPLQYWKNLTDLRVSLKGFYQSF-KEYLAR-GGK 244
SSV2   211 FERDAKSEYYLIPVSMWKNLVELRVAIKGFYQSF-KEYLTRVSGK 211
SSV4   209 FERDAKSEYYLIPVSMWKNLTELRVSLRGFYQSF-KEYLARVSGK 209
SSV5   214 FENGAKTPYFLIPRDQWKPFRELIVGIKNWYSTISEEFMYRVSGK 214
           **  *   * ***   **    *            *   *   *
```

Figure 5-1. ORF D212 from SSVRH and sequence alignment.
(A) SSVRH genome with open arrows representing different ORFs from SSVRH while the filled arrow shows the location of D212 in the SSVRH genome. Picture made using Vector NTI Advance 10.1.1 (Invitrogen). (B) Picture of the D212 crystal used to solve the protein structure. (C) Sequence based alignments between SSVRH D212 [22] and its homologs in the other fuselloviruses the SSV1 D244 [20], SSV2 ORF 211 [23], SSV4 ORF 209 [24], SSV5 ORF 214 (Peng et.al., unpublished). The * indicates strictly conserved residues between the five members of the *fuselloviridae* family. The sequence alignments were CLUSTAL W [138].

Despite significant efforts to crystallize D244 for structural annotation all trials have been unproductive. We have thus turned to studies of the D244 homologs as they may be more amenable to structural studies. Indeed SSVRH D212 was successfully

purified and crystallized. X-ray diffraction studies of D212 reveal a structure with close similarity to Holliday junction cleavage (Hjc) enzymes. Structural homology also reveals similarity to other members of the nuclease superfamily, including type II restriction endonucleases and an atypical homing endonuclease. The functional implications are discussed.

Results

The D212 construct used in these studies consists of the 212 residues of the full length protein plus an N-terminal His_6–tag, giving a total calculated mass of 25,895 Da. D212 elutes from a calibrated Superdex^{TM} S-75 size-exclusion column with an apparent molecular mass of 50 kDa (data not shown), suggesting that D212 is present in solution as a homodimer.

Structure of the D212 Subunit

D212 crystallizes in space group $P2_1$ with two copies of the D212 polypeptide chain in the asymmetric unit and solvent content of 49.14%. The structure of D212 was determined at 2.4 Å resolution by multiple isomorphous replacement with anomalous scattering (MIRAS), using Br and Pt derivatives. Detailed statistics on data collection and structure refinement are presented in Tables A-9 and A-10. Consistent with size exclusion chromatography, the crystal structure reveals a homodimer with extensive contacts between the two subunits that are related to each other by a noncrystallographic 2-fold axis.

Table A-9. Data Collection[a,b] for D212					
Data Set	D212	D212-K_2PtCl_4 inflection	D212-K_2PtCl_4 peak	D212-K_2PtCl_4 remote	D212-NaBr peak
Wavelength (Å)	0.95369	1.07270	1.07203	0.82653	0.91163
Space Group	$P2_1$				
Cell Constants (a,b,c; Å) (α,β,γ;°)	56.68, 82.11, 60.29 90.0, 115.48, 90.0	56.55, 82.02, 60.18 90.0, 115.41, 90.0	56.31, 81.76, 59.93 90.0, 115.44, 90.0	56.49, 81.88, 60.10 90.0, 115.408, 90.0	56.57, 81.59, 60.53 90.0, 115.03, 90.0
Resolution Range (Å)	50-2.4 (2.49-2.40)	50-3.05 (3.16-3.05)	50-2.8 (2.9-2.8)	50-3.35 (3.47-3.35)	50-2.8 (2.9-2.8)
Unique Reflections	18,121 (1,868)	9,530 (958)	12,089 (1,207)	7,233 (708)	12,107 (1,186)
Average Redundancy	3.1 (3.0)	6.3 (6.4)	6.3 (6.3)	6.4 (6.4)	7.3 (5.9)
I/σ	32.6 (4.3)	41.4 (7.5)	45.2 (7.1)	29.06 (6.89)	28.6 (3.7)
Completeness (%)	92.9 (95.7)	99.9 (100)	99.2 (100)	98.8 (99.4)	99.5 (96.9)
R_{sym}[c] (%)	3.4 (32.6)	6.2 (29.3)	6.5 (30.8)	7.9 (31.7)	7.2 (33.2)
[a]Data were integrated, scaled, and reduced using the HKL-2000 software package.					
[b]Numbers in parenthesis refer to the highest resolution shell.					
[c]$R_{sym}=100*\Sigma_h\Sigma_i\|I_i(h)-<I(h)>\|/\Sigma_h I(h)$ where $I_i(h)$ is the i^{th} measurement of reflection h and $<I(h)>$ is the average value of the reflection intensity.					

Each D212 subunit is composed of a mixed α/ß structure with five α-helices and eight β-strands (Figure 5-2 A). Each protomer can be divided into two smaller subdomains. The larger central subdomain is a made up of a predominantly antiparallel mixed ß- sheet which consists of five ß-strands (β1, β2, β3, β5 and β6), with flanking α-helices on each face (α1/α6 and α2/α4 respectively); while the smaller subdomain is composed of a 3-stranded antiparallel β-sheet (β4, β7 and β8) with flanking helices (α3 and α5). A number of the connecting loops are disordered (shown as dotted lines in Figure 5-2 A and B), in chain A these include residues 1-16 at the N-terminus, residues 58-67 in the β1-α2 loop, residues 100-106 connecting β3 to α3 , residues 172-174 in the β7- β8 loop, and residues 205-212 at the C-terminus. While in chain B residues 1-18 at

the N-terminus, 57-67 in the β1-α2 loop, 108-112 between the β3 to α3, 170-175 in the β7- β8 loop, and 208-212 at the C-terminus were also disordered.

Table A-10. Model Refinement for D212	
R_{cryst}[c] (%)	20.9
R_{free}[c] (%)	24.5
Real Space CC[d] (%)	94.5
Mean B Value (overall; Å^2)	32.2
Coordinate Error (based on maximum likelihood, Å)	0.169
RMSD from ideality:	
Bonds (Å)	0.005
Angles (°)	0.698
Ramachandran Plot[e]:	
Most Favored (%)	98.1
Additional Allowed (%)	1.9
PDB Accession Code	2W8M
[c]$R_{cryst} = \Sigma\|\|F_o\|-\|F_c\|\|/\Sigma\|F_o\|$ 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 (5%) that were withheld from refinement.	
[d]Correlation coefficient (CC) is agreement between the model and $2mF_o$-DF_c electron density map.	
[e]Calculated using Molprobity.	

The dimer interface of D212 (Figure 5-2 B) is formed as a result of side by side stacking of the α1 helix with the symmetry related helix of the secong protomer in a side by side fashion. In addition, the N-terminal loop and strand β1 are also involved. Together, these elements form a largely hydrophobic interface that, upon dimer formation, buries ~1100 Å^2 (11.9%) of the subunit surface. Residues involved at the interface are largely hydrophobic and involved in non-covalent intersubunit interactions. These residues, as tabulated by PDBsum [139] are listed in Figure 5-2 panel C.

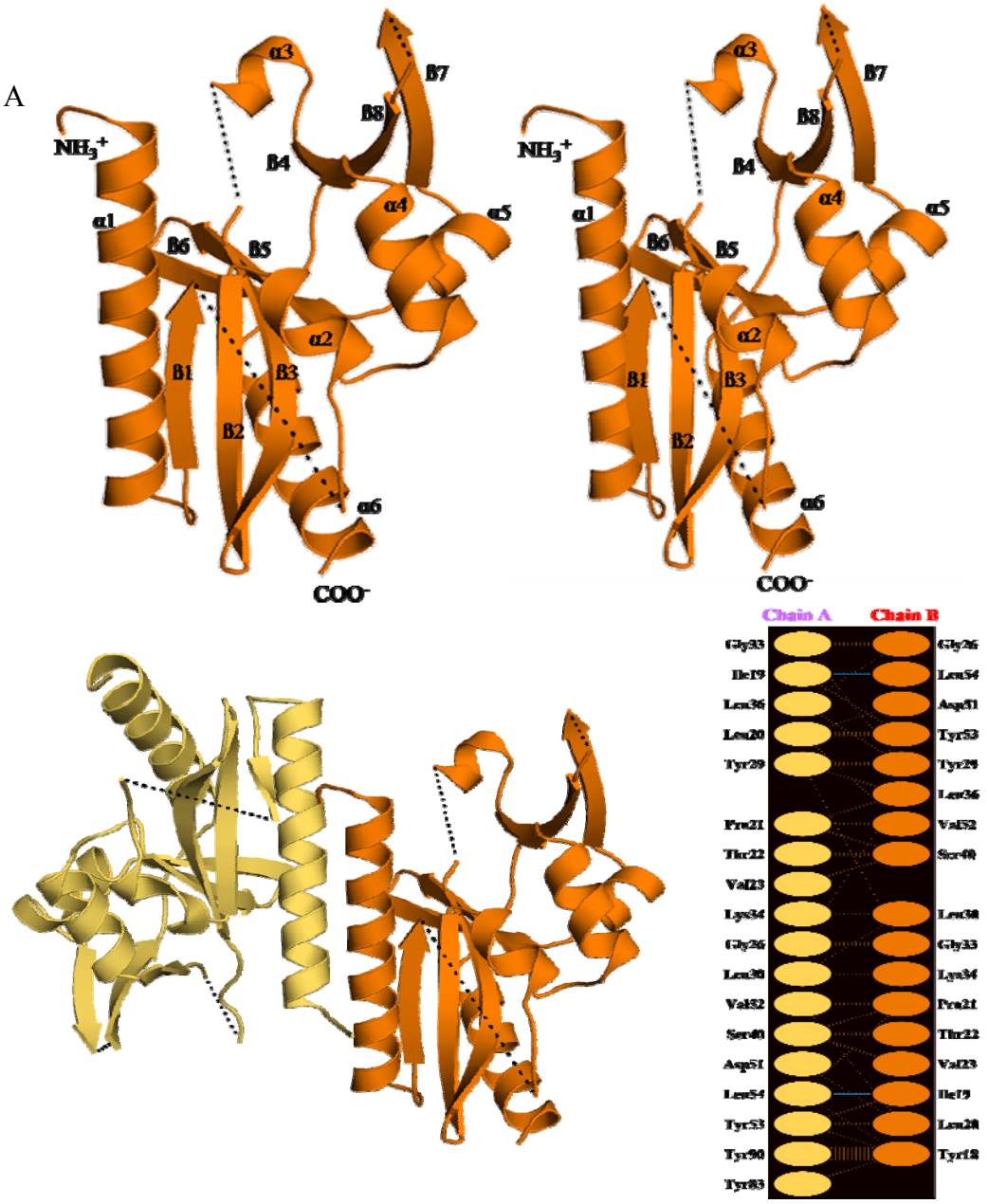


Figure 5-2: Structure of D212.
(A) Stereo image of the D212 subunit, chain A. The predominantly antiparallel mixed ß-sheet made up of five ß-strands (β1, β2, β3, β5 and β6) packed in between three α-helices (α1, α3 and α5) forms the central domain. While the other subdomain is made up of an additional anti-parallel three stranded ß sheet (β4, β7 and β8) along with two α-helices (α3 and α4) and an incomplete α-helix (α2). The dashed lines represent breaks in the chain. (B) The D212 dimer with the orientation of the subunit A identical to that in panel A. Helix α1is a prominent feature at the subunit interface (C) The interaction between the α1 helices of subunits A and B results in a primarily hydrophobic dimer interface with ~12% buried surface area per subunit. Many of the intersubunit interactions which are summarized in panel C are hydrophobic in nature.

Superposition of the two subunits show similar topology over most of the structure with an RMSD of 0.62 Å. The only notable difference between the two subunits is the absence or presence of the short α3 helix, which is missing from chain B.

Structural Homology

A search for structurally similar proteins using the DALI [71] , VAST [72] and SSM [87] servers identified members of the E-PD-(D/E)XK nuclease superfamily as the closest structural homologs (Figure 5-3). This diverse family of endonucleases was initially identified in structurally characterized type II restriction enzymes, but has subsequently been recognized in many other enzymes involved in DNA replication, repair, and recombination [140]. Consistent with the lack of functional annotation for D212 and its orthologs, members of the E-PD-(D/E)XK nuclease superfamily can exhibit extreme variability in their amino acid sequences, making it difficult to recognize highly diverged members. However, crystallographic studies show they share a common structural core [140]. The structure of D212 was found to be most similar to the Holliday junction cleavage enzyme (Hjc) from *Sulfolobus sulfotaricus* (PDB 1HH1), with a DALI Z-score = 9.0, and an RMSD = 3.0 Å with 12 % identity over 115 equivalent residues. Other proteins showing significant similarity includes several additional archaeal Hjcs, an atypical homing endonuclease (I-Ssp68031) and members of the type II restriction endonuclease. The atypical bacterial homing endonuclease (PDB 2OST) gives a DALI Z-score = 8.5, RMSD = 3.2 Å and 12% identity over 117 equivalent residues. While among the restriction endonucleases, HincII (PDB 2GIG) shows greatest similarity with a DALI

Z-score= 5.0, RMSD = 3.4 Å with 9 % identity over 116 equivalent residues. Importantly the structures of I-Ssp68031and HincII have been determined in complex with DNA.

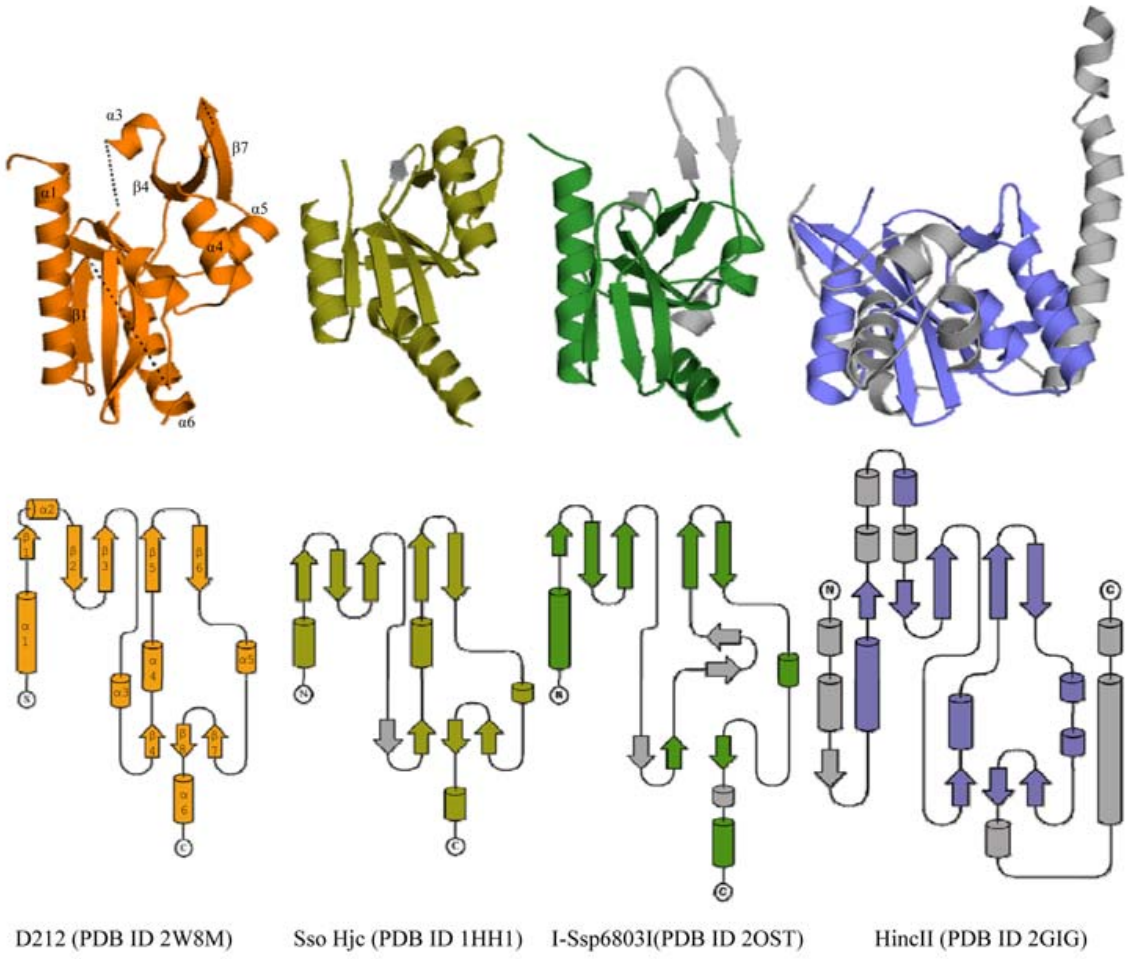


Figure 5-3. Topology comparison of D212 and its structural homologs
Cartoon and topology representations of D212 (orange), Sso Hjc (olive), Homing Endonuclease (I-Ssp68031 (green) and Restriction Endonuclease HincII (blue). Although the three endonucleases exhibit a similar core fold, there are significant differences. Secondary structural elements in Sso Hjc, the homing endonuclease and HincII which are not equivalent to D212 are colored in grey. Prepared using Pymol [74] and TOPDRAW [115].

Superposition of the D212 subunit onto Sso Hjc shows a very similar fold. In spite of the much larger protein size, 212 residues compared to ~145 residues in the Hjcs, the structure D212 does not reveal any additional secondary structural elements, although the α3 helix of D212 is replaced by a β4 strand in Sso Hjc (Figure 5-3). Instead the

increased size of D212 is largely a result of increased loop lengths in regions connecting the various secondary structural elements and an additional 10 residues at the N-terminus. Interestingly, the N-terminal residues are disordered in both D212 (18 disordered residues) and Sso Hjc (8 disordered residues). For the Hjcs, these residues have been implicated in DNA binding and are thought to be ordered only upon interaction with DNA [115]. Similarly, the homing endonucleases, 151 amino acids, is also largely similar, although consistent with the lower Z score, additional differences are seen (Figure 5-3). These are particularly noticeable in the second subdomain, involved in DNA recognition, where the homing endonucleases inserts two additional β-strands, and losses the α4 helix. Although HincII with 257 amino acids is much larger than D212, the secondary structural elements of D212 also match closely to a substructure of the restriction endonuclease HincII [141]. The major differences between the two structures are the addition of the two α helices and a β-strand at the N-terminal end, a small saposin-like domain that is inserted following the first β-strand, and 2 α helices at the C-terminal end. In addition, the last helix in D212 adopts a different orientation than that seen in HincII.

Active Site Architecture

The archaeal Holliday junction cleavage enzymes, homing endonucleases and type II restriction enzymes that show structural similarity to D212 are all members of a larger nuclease superfamily that contain a common E-PD-(D/E)XK active site motif although none of the individual residues in this motif is strictly conserved, particularly the first glutamate and the proline and lysine residues. Superposition of the Sso Hjc, the

atypical homing endonuclease and the HincII restriction endonuclease onto D212 identifies an equivalent active site motif in D212 (Figure 5-4 A). Specifically, Glu28, Asp79 and Asp97 in D212 correspond to residues Glu12, Asp42 and Glu55 in Sso Hjc, while in HincII, these residues correspond to Glu35, Asp114 and Asp127. Importantly, these residues are also strictly conserved in each of the D212 fuselloviral orthologs, including SSV1 D244. In contrast, the homing endonuclease makes several substitutions, utilizing Asp11, Asp36 and Gln49. This cluster of acidic active site residues generally function to co-ordinate a metal ion, either Mg^{2+} or Mn^{2+}, that itself facilitates an interaction with the negatively charged substrate DNA, or activates a water for nucleophilic attack. The lysine in this motif is proposed to correctly orient and stabilize the attacking hydroxyl nucleophile, or to play a role in transition state stabilization [142, 143]. While structure based sequence alignments do not show a strictly conserved lysine or proline residue in the equivalent position, the D212 structure does show that Lys123 of D212 is in close proximity to the three acidic residues of the active site, and might serve a similar role. Interestingly, Lys123 is also strictly conserved among the fuselloviral homologs, and a structurally equivalent lysine is also found in an identical position in the HincII structure. Further, migration of the lysine residue to other regions of the polypeptide to form the active site center has been observed in other restriction endonucleases [144]. And while the proline residue of the E-PD-(D/E)XK motif is present in the Sso Hjc, it is absent from HincII, D212 and the D212 orthologs. This proline residue is found to be substituted by an isoleucine, valine, asparagine or threonine residue in several endonucleases [143]. This substitution is also seen for D212 and HincII where the proline is replaced by a tyrosine and alanine respectively.

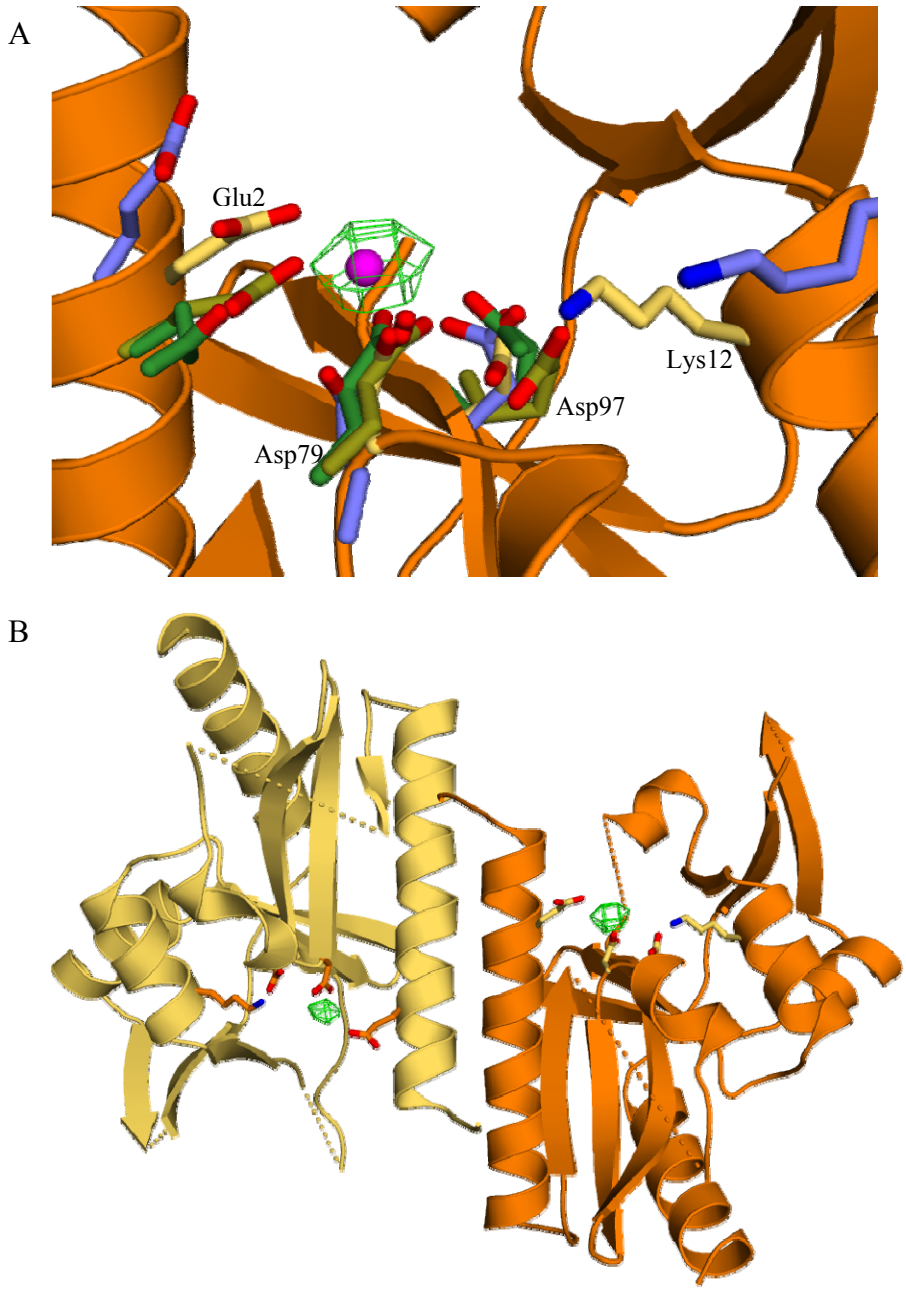


Figure 5-4. Active site residues of D212.

(A) Active site residues from Sso Hjc (olive), homing endonuclease (green) and HincII (blue) are superimposed on D212 (yellow). Glu28, Asp79, Asp97 and Lys123 are structurally conserved among D212 and its closest structural homologs. The predicted position of Mn^{+2} is depicted as a sphere (magenta), while the green mesh represents the 5σ difference electron density from the Mn^{2+} soaked crystals. (B) The D212 dimer shown in orange and yellow, with putative catalytic residues and the catalytic Mn^{2+} depicted as described for (A). The two Mn^{2+} ions are 24Å apart, similar to the distances seen in Holliday junction cleavage enzymes. However, this distance is significantly greater than that seen in restriction and homing endonucleases which bring to active sites to bear on a single piece of double stranded DNA. Difference electron density for Mn^{+2} and the active site residues are also depicted as described above.

The structural similarity to Sso Hjc and HincII suggested that Glu28 and Asp79 of D212 would also serve to coordinate Mg^{2+} or Mn^{2+}. We tested this prediction by soaking D212 crystals in 10 mM $MnCl_2$, and collecting a relatively low resolution data set to 3.5 Å on our home source X-ray diffractometer. Indeed, the difference density map for the Mn^{+2} soaked D212 crystals (Figure 5-4 panel B) confirmed the presence of manganese at the predicted metal binding site, further confirmation that these are the active site residues in D212. These putataive active site residues are situated at the N-terminal end of strand β2 and at the C-terminal end of strand β3, which form the core of the central β-sheet. Importantly, the two active sites in the D212 dimer are separated by about 25 Å. While this separation is similar to that seen in Holliday junction cleavage enzymes, this is significantly greater than that seen in type II restriction endonucleases, where two active sites come together to cleave opposing strands in a single piece of double stranded DNA.

While the structures of the archaeal Hjc enzymes lack DNA, the structure of the homing endonuclease and HincII in complex with DNA has been determined. This allows the DNA of homing endonuclease and HincII DNA complexes to be docked on D212 by structural superposition. Despite greater overall similarity to the homing endonuclease, the fit of the HincII DNA on D212 shows fewer steric clashes, and appears to be a superior model. We attribute this to greater structural similarity between the elements involved in DNA recognition in these two proteins. Indeed, while a pair-wise DALI alignment between the second subdomain of D212 and HE yields a Z score of 2.1 over 41 aligned residues, the superposition with HincII gives a Z-score of 3.5 over 47 aligned residues and an RMSD of 2.7 Å. The docking by superposition demonstrates how substrate DNA may be accommodated within the rather prominent clefts on the “sides”

of the D212 dimer. DNA recognition is likely to be mediated by residues at the N-terminus of helix α1, the α2/β2 loop, residues lying within the extended connection between β3 and β4, the N-terminus of α4, β4, and the β7-β8 loop. Many of these residues lie along the periphery of the predicted DNA binding site, in disordered or poorly ordered areas of the model, and are likely to become ordered upon interacting with DNA. While the majority of these interactions are likely to be with the ribose-phosphate backbone, the disordered β7- β 8 loop is predicted to lie within the DNA major groove. The equivalent loop in HincII is known as the recognition loop, or R-loop, and participates in base specific interactions. This suggests that the β7-β8 loop in D212 may play an equivalent role [145], raising the possibility that D212 may be involved in sequence specific interactions.

Although there is significant structural similarity between D212 and the Hjc, HincII and HE subunits, the location of the subunit interface in these structural homologs are substantially different from that seen in D212. For HincII, the subunit interface is such that two active sites are brought to bear on a single piece of double stranded DNA, which recognizes a degenerate 8 base-pair sequence to make a blunt ended cut. Similarly, the HE also utilizes two symmetrically arranged active sites to cleave a 23 base-pair target sequence, yielding 3 base, 3' overhangs. In contrast, the active sites in D212 and the Hjc are much further apart, ~ 25 Å, which in the case of D212, allows double stranded DNA to be docked into each of the active sites (Figure 5-5). These differences may suggest that D212 is unlikely to function as a restriction enzyme or homing endonucleases. While the structures of two Holliday junction cleavage enzymes in complex with DNA have been reported, these proteins lack significant structural

similarity to D212. Nevertheless, we attempted to dock these junctional complexes to D212, but were unable to arrive at a satisfying model using either manual or automated methods.

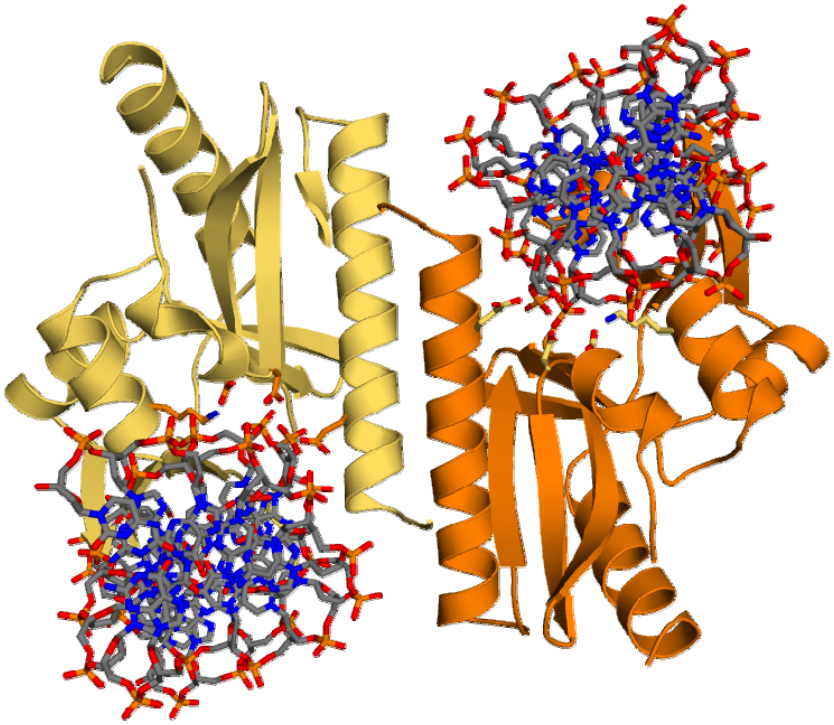


Figure 5-5. The interaction of D212 with dsDNA modeled using structural superposition. A model for the possible interaction of dsDNA with the two catalytic centers of D212 dimer is represented in this figure. One DNA bound HincII subunit was superimposed on each subunit of D212. The rotated DNA from HincII [141] shown as sticks was docked onto D212 (shown as cartoon) by structural superposition.

Surface Characteristics

The electrostatic surface of D212 reveals a bipolar charge distribution. The surface of Hjc from *S.solfataricus* is largely flat with an 'S' shaped positive electrostatic potential that snakes its way around the active site metal binding residues involved in Holliday junction binding [146]. Similar to Sso Hjc other nucleases such as *Pyrococcus furiosus* Hjc, T4 endonuclease VII and EcoRV also exhibit a significant bipolar

distribution of surface electrostatic charges (Figure 5-6). However, the surface of D212 is not flat and extensive like other nuclease. Instead it appears to be rather convoluted with many protrusions and clefts spread over the entire surface. In addition the position of the active site residues, which are embedded within the negatively charged surface patches, are found on the opposite side of the dimer as opposed to related residues in the other nucleases where they are on the positively charged surface.

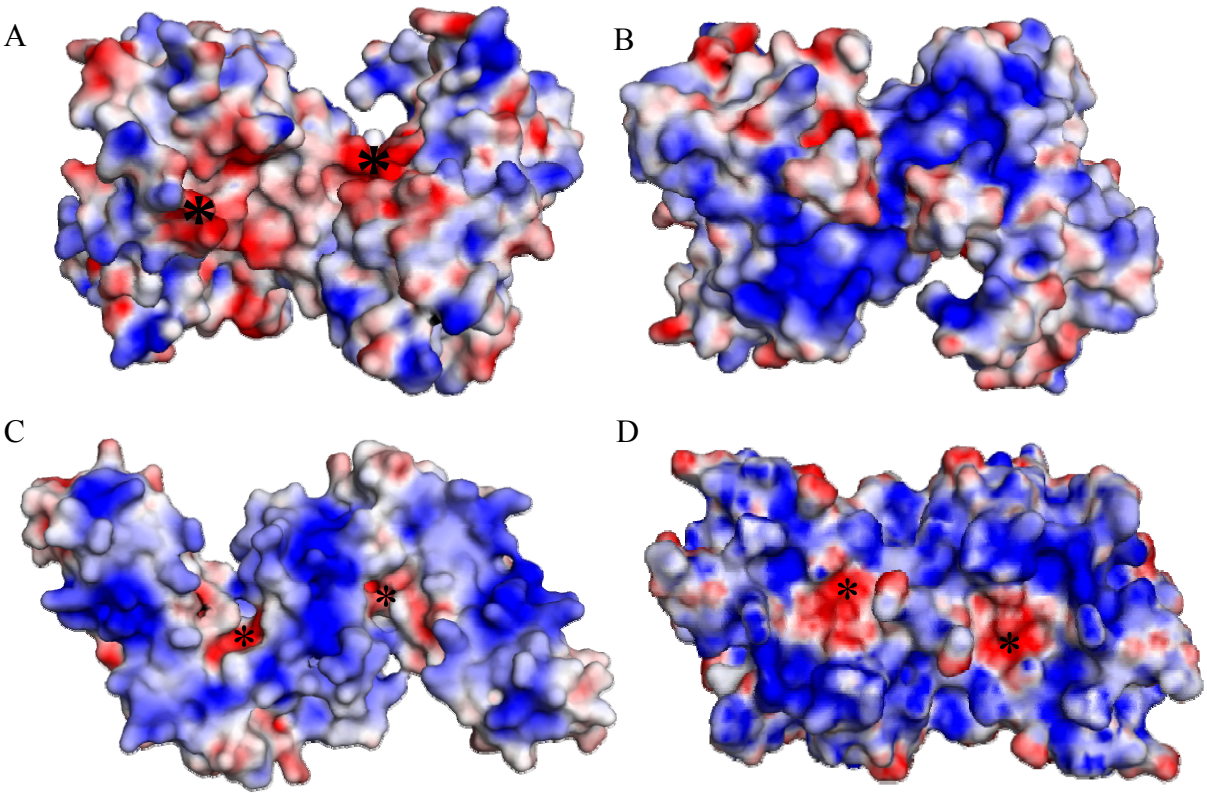


Figure 5-6. Electrostatic charge distribution on the surface of the D212 dimer.
(A) The electrostatic charge distribution mapped onto the surface of D212 reveals a characteristic bipolar distribution of charges similar to that seen in other endonucleases such as Sso Hjc. The orientation of D212 is the same as in Figure 1. The color ramps from red (-25 kT/e, acidic) to blue (+25 kT/e, basic). The active site residues, marked by an *, are embedded within the acidic (red) patches on the sides of the dimer (note that the divalent cations are not included in the calculation of surface electrostatic potentials) away from the basic (blue) face of the protein which may also be involved in DNA interactions. (B) Relative to panel A, D212 has been rotated by 180º about the depicted axis. Images were prepared by using APBS [147]. (C) electrostatic charge distribution mapped onto the surface of T4 endonuclease VII [148]. (D) Electrostatic charge distribution mapped onto the surface of SsoHjc [115].

Dimer Interface

The position of the dimer interface of D212 is significantly different than that seen in the archaeal Hjc resolvases, the homing endonucleases and the HincII restriction enzymes. The D212 dimer interface is composed of the N-terminal α-helix (α1) and the first β-strand (β1) of the central β-sheet. In contrast, the dimer for Sso Hjc is mainly formed through the interaction of the hydrophobic residues on one face of the central β-sheet. A second possible weaker dimer formation is also reported for Sso Hjc through the interaction of the β4-β5 and the β8-β9 loops between the peripheral β-sheet of the two Hjc subunits [115]. While for HincII the long C-terminal helix of HincII forms the major part of the dimer interface. Upon binding DNA, endonucleases undergo a major conformational change within their subunits while bending the DNA itself [145]. Major rotation and in some cases translation distortions are seen between the endonuclease dimers upon DNA binding. Thus features such as the orientation of the dimer interface and how tightly the dimer is held together are also important to determine the activity of these endonucleases [145].

Nuclease Activity:

The structural similarity to Holliday junction cleavage enzymes suggests that D212 may serve a similar function. Thus, we next tested whether D212 might exhibit this activity. We first looked for binding of D212 to a synthetic Holliday junction (Hj) using the electrophoretic mobility shift assays (EMSA) and ^{32}P-labelled fixed or mobile Hj as substrate. Constant amounts of the DNA junction Hj12 (10nM) were incubated with increasing concentrations of D212 in the presence of 5mM EDTA for 1 hour at room

temperature. The reactions were analyzed on a native 6%TBE gel. Various incubation temperatures 37^{o}C and 60^{o}C were also tried. The initial binding assays were done at pH 6.5, a pH range of 6.0 - 9.0 was also tried. No retardation in the mobility of the DNA due to binding of D212 was seen in any screened condition (no data shown).

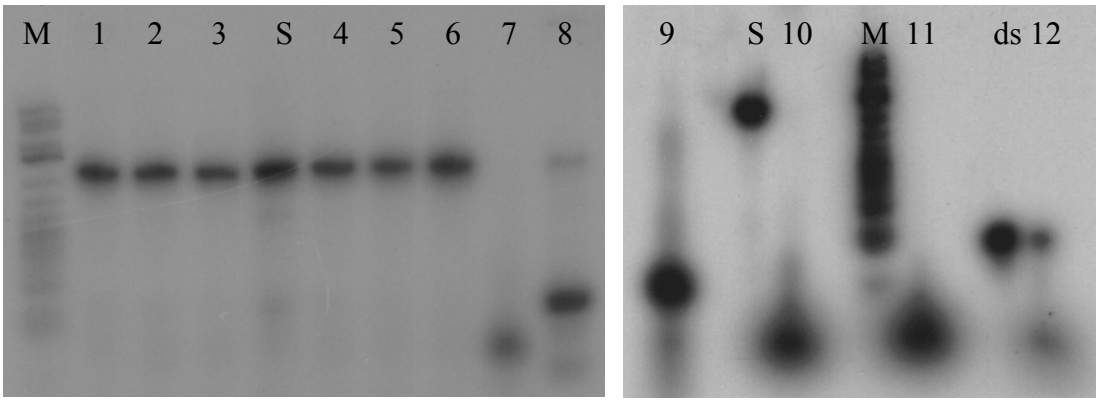


Figure 5-7. Biochemical characterization for D212.
Cleavage of the 10nM fixed four-way junction Hj 12 with lane1: 10nM D212, lane 2: 20 nM D212, lane 3: 30 nM D212, lane 4: 50 nM D212, lane 5: 100 nM D212, lane 6: 200 nM D212, was incubated as described in Materials and Methods and the products separated on a 8% native gel. Lane10 and lane 11 contained 10nM Hj 13 with 1μM D212. Lane 8 was the positive control which contained 10nM Hj 12 with 30nM T4 endonuclease VII. Lanes indicated with S contained substrate without protein and those with M are ^{32}P labeled ΦX174 DNA ladder. The lane marked 'ds' is 10nM double stranded DNA made as described in materials and methods section "synthetic DNA substrates" incubated with 1μM D212.

To test for Holliday junction cleavage (Hjc) activity, D212 was incubated with various substrates, using various reaction temperatures and buffer conditions as described in chapter 2. T4 endonuclease VII, was used as a positive control for the cleavage reaction (Figure 5-7, lane 8). After incubation of the reaction mixture at room temperature for 2 hours, the products were separated out on an 8% native acrylamide gel. No apparent cleavage of the Hj was observed by D212 even with protein concentrations at almost 20 times the DNA concentration (Figure 5-7 lanes 1-6). Cleavage was seen only when the concentration of the protein was increased to almost 1μM, a 100:1 protein:DNA

ratio. In this case the size of the cleaved product obtained was smaller than the expected size for the product (Figure 5-7 lanes 10 and 11). Although this difference in size could not be explained, the mobility of the product as judged from the native gel was slower than inorganic phosphate and hence a phosphatase type activity for D212 was rejected (data not shown). Cleavage activity of D212 for other substrates, like the dsDNA and ssDNA were also examined where some cleavage activity was seen in case of the dsDNA substrate.

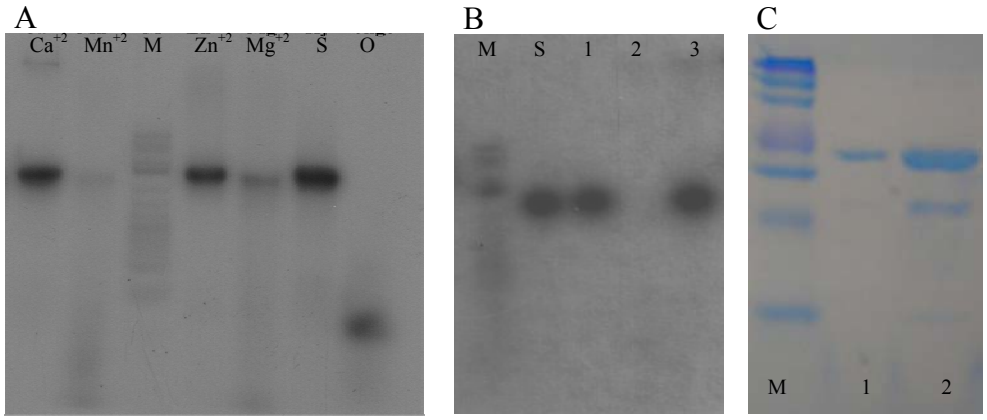


Figure 5-8. Effect of different divalent metal ion and proteolysis on D212 cleavage activity. (A) Cleavage of the 10nM fixed four-way junction Hj 12 with Ca^{+2}, Mn^{+2} and Zn^{+2} substituted in the original condition which contained 10mM Mg^{+2}. Lanes indicated with S contained substrate without protein and those with M are ^{32}P labeled ϕX174 DNA ladder. Lane marked as O contained 10nM oligonucleotide 12 used for making the Hj. (B) Hj Cleavage activity of D212 was only observed when the protein was allowed to age. 10nM Hj12 was incubated with 1μM D212, lane1 contains freshly purified D212, lane 2 has D212 which has been allowed to undergo proteolysis and lane 3 contains reaction carried out with freshly purified protein in elution buffer. (C) Proteolysis of D212. Lane1: freshly purified D212, lane2: D212 which has been stored for 2 weeks and lane M: precision plus protein marker.

Holliday junction cleavage enzymes generally bind and cleave substrate DNA based on its structure rather than a specific sequence. The presence of a divalent cation is one of the most critical factors to influence the structure which a Holliday junction assumes and is also important for the activity of the enzyme. Hence, the activity of D212 was also assayed in the presence of different cations, the Mg^{+2} present in the initial buffer

condition was substituted by other metal ions and the reaction was carried out keeping all other reaction parameters the same. Cleavage activity was seen to be most efficient in the presence of 10mM Mn^{+2} (Figure 5-8 A).

To reproduce the results from the activity assay, D212 was freshly purified and the assays mentioned above were repeated. Surprisingly, the cleavage reaction could not be reproduced. Comparison of freshly purified D212 to the protein, for which the activity was observed, revealed a striking difference between the two. The older active protein had clearly undergone proteolysis (Figure 5-8 C). Interestingly, the freshly purified D212 also showed the same proteolytic pattern upon storage at $4^{o}C$ for 2 weeks, resulting in the formation of two additional protein species as identified initially by SDS-PAGE and Western blots. Using MALDI and N-terminal Edman sequencing, the three protein constituents were identified as full length protein, and two cleaved products lacking 5 and 64 N-terminal residues respectively. Further storage resulted in complete proteolysis of the full length protein and only the two bands corresponding to the smaller cleaved products were visible. Importantly, the appearance of these proteolytic products was accompanied by return of the Hjc activity described above.

The nature of the identified cleavage site suggested that proteolysis might disrupt the dimer interface and that this might be detected using size exclusion chromatography to separate monomer from dimer. However, we found that the cleaved products show the same elution volume as intact protein. This suggests that even after proteolysis the cleaved products are still associated with each other.

However, recently I have not been able to reproduce the *in vitro* Hjc activity for D212, seen earlier. The fact that I have not been able to reproduce similar activity even

for the positive control (T4 endonuclease VII) also suggests a need for additional experiments. Different buffers, salt concentrations, divalent cations, temperatures and DNA targets have all been tried. This could mean that either the purified form of D212 used for the assays is probably non-functional or the reaction condition that worked before has somehow changed. Regardless, the extreme concentrations of D212 needed to see Hjc activity suggest this activity bring into question the physiological relevance of the observed activity, or suggest that it is an artifact of some kind.

Nucleases are also involved in replication initiation through nicking; this activity was also tested for D212. The SSV1 shuttle vector pKMSD48 [13] was used as the substrate to test for the nicking activity, no nickase like activity could be established for D212 under the tested condition. Specificity for other substrates such as Y-loops and D-loops also need to be tested.

Discussion

D212 and its homologs are well conserved in all *fuselloviridae* except SSVK. Hence an important role for D212 and its homologs in the viral lifecycle could be envisioned. Sequence comparisons revealed that D244, the SSV1 homolog of D212 is much larger than its predicted homologues from the other SSV genomes. D244 is longer by ~33 amino acids compared to its homologs SSVRH D212, SSV4 209, SSV2 D211 and SSV5 ORF D204. The sequence conservation in the homologs suggests a possible misannotation of ORF D244. On further sequence analysis an AUG codon is found to be present within the annotated gene sequence. If this was considered to be the true start codon for D244 it would reannotate D244 to D212.

The structure of D212 shows very close similarity to members of this family of enzymes that function in the cells to catalyze cleavage of the phosphodiester bonds in nucleic acids. Nucleases are ubiquitous enzymes that have been isolated from bacteria, archaea, eukaryotes and their viruses. They perform diverse array of functions that include restriction-modification, resolving intermediates formed during DNA repair, recombination and replication and other events processes requiring junctional cleavage. Structural examples include enzymes like MutH [149], Vsr [150], RecB and λ-exonulcease [151], which are involved in DNA repair, archaeal Holliday junction cleavage enzymes [115, 146, 152], TnsA which is a component of the TN7 transposase [153], and the homing endonuclease I-SspI [154]. In addition, bioinformatics, biochemical and/or genetic approaches have identified the PD-(D/E)-XK motif in non-LTR retrotransposases [155], in XisH which is required for excision of the fdxN element in Anabena [156], and a variety of predicted nucleases whose specific activities remain to be determined [157]. Similar to most of the gene encoded products of these unique viruses, analysis of the gene sequence for D212 did not suggest a possible function for the protein. In contrast, our structural work suggests that D212 functions as an endonuclease. As the activity assay fails to provide a definite answer for the functional annotation of D212 as an Hjc, other possible roles for D212 should also be considered and are discussed below.

D212 appears to show greatest similarity to Holliday junction resolvases, which are nucleases that introduce paired incisions at symmetrically located regions across from the branch point of a Holliday junction to generate double stranded products. Hjcs have been identified and characterized from most known hyperthermophilic archaea [158] and

also from two crenarchaeal viruses [88]. Most of these enzymes display extreme diversity in sequence and substrate specificity but are highly selective for the structure of the junction they resolve. Similar to Holliday junction resolvases and other endonucleases, D212 also has a characteristic basic pI (~8.98) and exhibits very distinct bipolar electrostatic potential distribution. The primarily acidic E-PD-(D/E)XK catalytic motif seen in several endonucleases is also spatially well conserved in D212 and its orthologs. The distance between the two active sites in D212 dimer is 24 Å, close to that seen for other endonucleases involved in recombination, including Sso Hjc, Sso Hje, other archaeal Hjcs, and T4 endonuclease VII (~25 Å, [159]). The phosphodiester bond cleavage catalyzed by the canonical active site is brought about in three steps; i) the hydroxide ion nucleophile is generated by the deprotonation of water by a general base [154] or the phosphate group from the substrate that is 3' to the scissile bond [160], ii) the attack on the scissile phosphorus group by the nucleophile leads to a pentacoordinated transition state, which is stabilized by the metal ion and iii) the protonation of the 3'-OH leaving group by a general acid [161]. As this motif is also well conserved in D212 and its orthologs, a catalytically similar endonuclease activity can be visualized for D212. However, the nature of the substrate remains an open question.

The structural similarity of D212 to the archaeal Holliday junction cleavage enzymes, including the distance between the two active sites, suggests that D212 is most likely to function as a junctional resolvase. However, the D212 dimer differs substantially from the archaeal enzymes in at least one respect. Specifically, the dimeric archaeal enzymes position their two active site centers on a single, relatively flat surface [115, 146, 152], and are predicted to recognize a planar, stacked-X Holliday junction [152,

162], similar to that seen in the structure of T4 endonuclesase VII [163]. In contrast, the predicted DNA binding channels in D212 do not reside on a single surface, instead, they run at oblique angles down the sides of the dimer. This is reminiscent of the structure of the T7 endonuclease I complex, which is distinctly different than those of other junction-enzyme complexes, and suggests that D212 may recognize a junctional complex whose geometry is substantially different than that recognized by the characterized archaeal enzymes. This may be one explanation for the lack of cleavage activity in the D212 Holliday junction cleavage assay. Like T7 endonuclease I, D212 may show strong selectivity for a DNA branch point with a specific structure [164], and might exhibit significant discrimination for particular base sequences within the junctional arms [165]. Similar to phage T4 endonuclease VII, which resolves branched DNA structures in newly synthesized DNA, D212 might utilize junctional cleavage or resolvase activity to resolve intermediates that arise during replication of the fuselloviral genome.

Alternativley, recombinational exchanges are involved in DNA repair pathways, and expression of D212 might play a role in the correcting replication defects, or other damage to the viral genome. And while viruses frequently employ host machinery for their own benefit, there are numerous instances where viruses are known to carry their own set of DNA modifying proteins to carry out specific reactions beneficial to the virus. Expression of D212 might play a role in the correcting replication defects, or other damage to the viral genome. In this light, it is interesting that SSV1 D244, a D212 ortholog, is thought to be packaged within the SSV1 viral particle [50]. If D212 were packaged in the released viral particle, the repair enzyme could be available for

immediate use upon infection of a new host cell, helping to maintain the integrity of the viral genome in the acidic, hyperthermal environments in which these viruses reside.

Other roles for a nuclease such as nucleases from viruses such a chlorella virus [166] and Epstein Barr virus [167] are involved in the breakdown of the host DNA and subsequent recycling of the nucleotides for the production of viral DNA. On infection viruses are known to inhibit the synthesis of host cell proteins either through transcriptional or post-transcriptional mechanism [168]. RNases such as such as the virion host shutoff (Vhs) proteins from herpes simplex virus [169] and from vaccinia virus [168] are known to bring about such inhibition of host cell protein synthesis by degradation of the host mRNA. Nucleases are also actively involved in the initiation of DNA replication through nicking, removal of the RNA primers in Okazaki fragments and assisting the polymerase enzyme by proof reading and repair of misincorporated nucleotides in the newly replicated DNA strand.

Viruses are known for their capacity to evolve and diversify. Recombinational exchange of genetic material serves as a primary method to maximize variability in viral genomes. Homing endonucleases are selfish genetic elements known to drive ectopic horizontal transfer of introns and inteins in addition to vertical transmission, and encourage their persistence in prokaryotic and phage genomes [154]. The homing endonuclease that D212 shows closest similarity is an atypical member of the homing endonuclease family. The E-PD-(D/E)XK active site of the type II restriction endonuclease is conserved in this homing endonuclease and hence is believed to employ a similar mechanism for double stranded DNA cleavage.

Another possibility, though remote, is that the type II restriction endonuclease motif found in D212 is inactive and in fact the main function of the protein is somewhat different as is seen for an intron encoded homing endonuclease, I-*Ppo*I. This endonuclease exhibits the signature PD-(D/E)XK type II restriction enzyme catalytic center [170]. Yet the crystal structure of I-*Ppo*I demonstrates that the type II endonuclease catalytic site identified in this protein is only coincidental and the endonuclease activity of I-*Ppo*I is in fact the result of a His-Cys box motif, another catalytic center involved in nuclease activity.

The SSV1 integration into the host chromosome can be compared to that of phage λ. The SSV1 integrase has been shown to promote both integrative and excisive recombination of linear DNA substrates [171]. However, the integration reaction was found to be more efficient (~50%) compared to its excision activity (~10%). Thus a possibility for the involvement of accessory proteins was suggested by Muskhelishvili [171]. In bacteriophage the integrase catalyzes the excision reaction in the presence of accessory proteins. However, the integration of the SSV1 viral genome into the host chromosome results in disruption of the intact integrase proteins. This further warrants a need for another nuclease like protein to bring about the excision of the prophage genome before initiation of the virus DNA replication. Excisionase proteins with a PD-(D/E)XK motif have been identified from other viruses [172] and such an excisionase type activity for D212 could be conjectured and needs to be tested.

Conclusions

Structural analysis strongly suggests a possible nuclease like activity for D212 although additional experiments to confirm such an activity are required. Nucleases are ubiquitous enzymes that catalyze the cleavage of phosphodiester bonds in nucleic acids, with possible roles in DNA recombination, repair, replication, host DNA and mRNA degradation, restriction endonuclease or excision in a virus. Once a functional activity for D212 is established, further characterization of the active site residues as identified by structural annotation studies could be explored. Site directed mutagenesis of the active site residues might help us elucidate the role of each residue in the enzyme activity. D244 the homolog of D212 from SSV1, was found to co purify with the SSV1 viral particles [50]. The presence of the protein in the virus particle, suggests an involvement of the protein in the early or late stages of virus infection.

CHAPTER 6

CONCLUDING REMARKS

In comparison to the 5000 viruses known to infect bacterial and eukaryotic hosts, only 50 such viruses have been described that infect the archaea. In spite of a rapid increase in our knowledge of these viruses over the last few years, our understanding of these unique archaeal viruses is still very limited, particularly for the viruses infecting the Crenarchaeota. Most of the proteins encoded by these genomes lack any recognizable similarity to genes of known function due to the extreme genetic and morphological diversity seen among the archaeal viruses. This further hinders our progress towards characterizing these unique viruses. Structural studies can often reveal distant evolutionary relationships and provide functional insights that are not apparent from the primary amino acid sequence alone. This work details such structural annotation based functional predictions for two different proteins SSV1 F112 and SSVRH D212 and provides additional insights into the increasingly recognizable proteomescapes of the *fuselloviridae*, particularly the SSV1 and SSVRH viruses.

The structure of F112 revealed a monomeric winged helix-turn-helix DNA binding protein. Structural homology suggests close similarity to several eukaryotic and prokaryotic transcription factors. Thus, F112 is also likely to act as a transcriptional regulator. According to the microarray analysis, F112 which is encoded by the early T5 transcript [43] is among the first proteins to be expressed by virus infected UV induced host cells [44]. Hence, it may be involved in either the activation of the transcription of early gene products or may act as a repressor of the UV induced T_{ind} transcript.

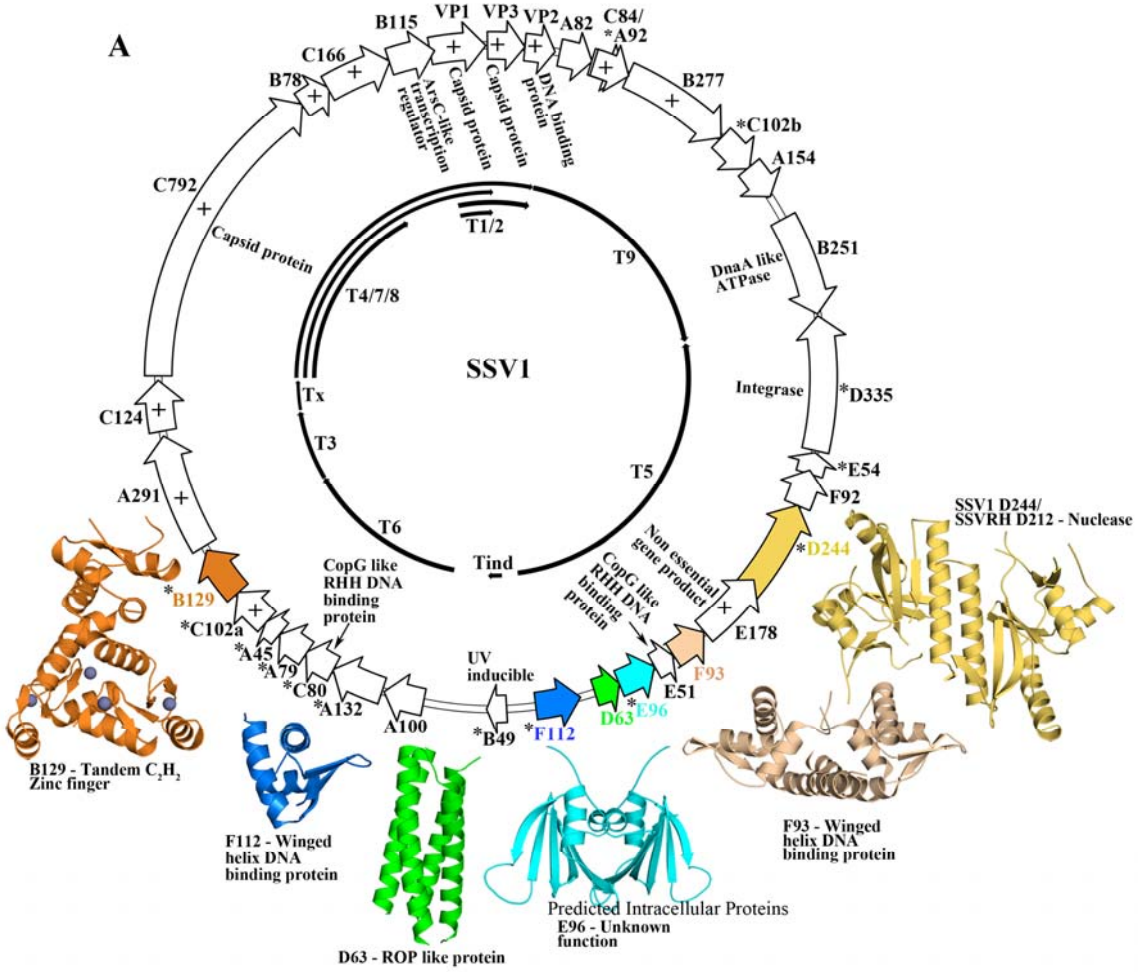


Figure 6-1. The Progress of the Structural Characterization of SSVs.
(A) The SSV1 ORFs are shown as open arrows, while transcript identities are shown in the interior (black lines). Predicted membrane proteins are denoted by a "+", proteins containing cysteine by an "*". The proteins whose structures have been determined are shown Eilers et.al. unpublished, Menon et.al., unpublished, [49-51].

The structure of SSVRH D212 shows close structural similarity to members of the nuclease family of enzymes particularly the Holliday Junction Cleavage enzyme. Our structural studies show that the PD-(D/E)XK motif commonly found in many type II restriction endonuclease is also well conserved in D212. Nucleases are enzymes involved in a wide range of functions from resolving DNA recombination, repair and replication intermediates to providing a competitive edge to the virus during multiple infections

through a restriction endonuclease type function to integration or excision by recombination. SSVs have a circular genome which integrates into the host chromosome upon infection. Akin to lambda phages, the SSVs might also require an excisionase to carry out the excision of the prophage efficiently. Integrases encoded by the viruses themselves are known to carry out such excision reactions with the help of ancillary proteins. However the SSV1 integrase (D335) is shown to be very ineffective in carrying out this reaction *in-vitro* [171] and may be completely inactive as the gene integrase gene gets disrupted on integration of the viral genome into the host chromosome [173]. Thus the integrase might require another protein to efficiently carry out the excision reaction or maybe another protein is entirely responsible for such an activity. Members of the PD-(D/E)XK containing motif are also implicated in viral genome excision [172] and such a role for D212 could be projected.

The structure of F112 also revealed an unexpected intramolecular disulfide bond. Several other protein structures from SSV1 and STIV namely, SSV1 E96, B129 and STIV B116 and F93 have also revealed the presence of disulfide bonds. This prompted us to investigate the distribution of cysteines in the SSV1 genome and expand the analysis to all available crenarchaeal viral genomes. Strong evidence for an abundance of disulfide bonds in the intracellular proteins of these viruses was obtained as suggested by a preference for proteins with an even numbers of cysteines compared to those with odd numbers of cysteine residues. Thermal melt studies on SSV1 F112, STIV B116 and F93 in the presence or absence of a reducing agent confirms the importance of an intact disulfide bond in the stability of these proteins at high temperatures. Elimination of the disulfides by mutational studies on F112 further validates the importance of a disulfide

bond in protein thermostability. Thus our structural work [50, 98, 107], mutational studies on F112, the bioinformatics analysis of the crenarchaeal viral genomes, and the combined metagenome, provide significant evidence for the common occurrence of disulfides in thermophilic viruses and points to a role in protein thermostability.

Structural characterization of F112 and D212 proposes a likely interaction of these proteins with nucleic acids. Initial studies to identify the respective binding partners have characterized a non-specific DNA binding activity for F112 while similar activity for D212 is still under investigation. Winged helix transcription factors often act as homo- or hetero dimers and the missing 40 residues on the C-terminal of F112 suggest a possible protein interaction for F112 as well. Although the structures of the proteins have suggested a possible function further biochemical characterization of these proteins is necessary to confirm these functions.

REFERENCES CITED

1. Wilhelm, S.W., and Suttle, C.A., *Viruses and nutrient cycles in the sea.* Bioscience, 1999. 49: p. 781-788.

2. Woese, C.R., O. Kandler, and M.L. Wheelis, *Towards a Natural System of Organisms: Proposal for the Domains Archaea, Bacteria, and Eucarya.* PNAS, 1990. 87(12): p. 4576-4579.

3. Barns, S.M., et al., *Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences.* Proc Natl Acad Sci U S A, 1996. 93(17): p. 9188-93.

4. Huber, H., et al., *A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont.* Nature, 2002. 417(6884): p. 63-7.

5. Allers, T. and M. Mevarech, *Archaeal genetics - the third way.* Nat Rev Genet, 2005. 6(1): p. 58-73.

6. Preston, C.M., et al., *A psychrophilic crenarchaeon inhabits a marine sponge: Cenarchaeum symbiosum gen. nov., sp. nov.* Proc Natl Acad Sci U S A, 1996. 93(13): p. 6241-6.

7. Baker, B.J., et al., *Lineages of acidophilic archaea revealed by community genomic analysis.* Science, 2006. 314(5807): p. 1933-5.

8. Wolfram Zillig , K.O.S., Simon Wunderl , Wolfgang Schulz, Harro Priess , and Ingrid Scholz *The Sulfolobus-"Caldariella" Group: Taxonomy on the Basis of the Structure of DNA-Dependent RNA Polymersaes.* Archives of Microbiology, 1980. 125: p. 259-269.

9. Chen, L., et al., *The genome of Sulfolobus acidocaldarius, a model organism of the Crenarchaeota.* J Bacteriol, 2005. 187(14): p. 4992-9.

10. She, Q., et al., *The complete genome of the crenarchaeon Sulfolobus solfataricus P2.* Proc Natl Acad Sci U S A, 2001. 98(14): p. 7835-40.

11. Suzuki, T., et al., *Sulfolobus tokodaii sp. nov. (f. Sulfolobus sp. strain 7), a new member of the genus Sulfolobus isolated from Beppu Hot Springs, Japan.* Extremophiles, 2002. 6(1): p. 39-44.

12. She, Q., et al., *The complete genome of the crenarchaeon Sulfolobus solfataricus P2.* PNAS, 2001. 98(14): p. 7835-7840.

13. Stedman, K.M., et al., *Genetic requirements for the function of the archaeal virus SSV1 in Sulfolobus solfataricus: construction and testing of viral shuttle vectors.* Genetics, 1999. 152(4): p. 1397-1405.

14. She, Q., et al., *Genetic profile of pNOB8 from Sulfolobus: the first conjugative plasmid from an archaeon.* Extremophiles, 1998. 2(4): p. 417-25.

15. Prangishvili, D., P. Forterre, and R.A. Garrett, *Viruses of the Archaea: a unifying view.* Nat Rev Microbiol, 2006. 4(11): p. 837-48.

16. Prangishvili, D., R.A. Garrett, and E.V. Koonin, *Evolutionary genomics of archaeal viruses: Unique viral genomes in the third domain of life.* Virus Research, 2006. 117(1): p. 52-67.

17. Pietila, M.K., et al., *An ssDNA virus infecting archaea; A new lineage of viruses with a membrane envelope.* Mol Microbiol, 2009.

18. Forterre, P., *New viruses for the new millennium.* Trends Microbiol, 2001. 9(3): p. 114.

19. Martin, A., et al., *SAV 1, a temperate u.v.-inducible DNA virus-like particle from the archaebacterium Sulfolobus acidocaldarius isolate B12.* Embo J, 1984. 3(9): p. 2165-2168.

20. Palm, P., et al., *Complete nucleotide sequence of the virus SSV1 of the archaebacterium Sulfolobus shibatae.* Virology, 1991. 185(1): p. 242-50.

21. Schleper, C., K. Kubo, and W. Zillig, *The particle SSV1 from the extremely thermophilic archaeon Sulfolobus is a virus: demonstration of infectivity and of transfection with viral DNA.* Proc Natl Acad Sci U S A, 1992. 89(16): p. 7645-9.

22. Wiedenheft, B., et al., *Comparative genomic analysis of hyperthermophilic archaeal Fuselloviridae viruses.* J. Virol., 2004. 78(4): p. 1954-1961.

23. Stedman, K.M., et al., *Relationships between fuselloviruses infecting the extremely thermophilic archaeon Sulfolobus: SSV1 and SSV2.* Res Microbiol, 2003. 154(4): p. 295-302.

24. Peng, X., *Evidence for the horizontal transfer of an integrase gene from a fusellovirus to a pRN-like plasmid within a single strain of Sulfolobus and the implications for plasmid survival.* Microbiology, 2008. 154(Pt 2): p. 383-91.

25. Prangishvili, D., et al., *A novel virus family, the Rudiviridae: structure, virus-vost interactions and genome variability of the Sulfolobus viruses SIRV1 and SIRV2.* Genetics, 1999. 152(4): p. 1387-1396.

26. Vestergaard, G., et al., *A novel rudivirus, ARV1, of the hyperthermophilic archaeal genus Acidianus.* Virology, 2005. 336(1): p. 83-92.

27. Janekovic D, W.S., Holz I, Zillig W, Gierl A and Neumann H, *TTV1, TTV2 and TTV3, a family of viruses of the extremely thermophilic, anaerobic, sulfur reducing archaebacterium Thermoproteus tenax.* Molecular and General Genetics, 1983. 192(1-2): p. 39-45.

28. Arnold, H.P., et al., *A novel Lipothrixvirus, SIFV, of the extremely thermophilic Crenarchaeon Sulfolobus.* Virology, 2000. 267(2): p. 252-266.

29. Neumann, H., et al., *Identification and characterization of the genes encoding three structural proteins of the Thermoproteus tenax virus TTV1.* Mol Gen Genet, 1989. 217(1): p. 105-10.

30. Haring, M., et al., *Structure and genome organization of AFV2, a novel Archaeal Lipothrixvirus with unusual terminal and core structures.* J. Bacteriol., 2005. 187(11): p. 3855-3858.

31. Bettstetter, M., et al., *AFV1, a novel virus infecting hyperthermophilic Archaea of the genus Acidianus.* Virology, 2003. 315(1): p. 68-79.

32. Haring, M., et al., *Independent virus development outside a host.* Nature, 2005. 436(7054): p. 1101-1102.

33. Arnold, H.P., U. Ziese, and W. Zillig, *SNDV, a novel virus of the extremely thermophilic and acidophilic archaeon Sulfolobus.* Virology, 2000. 272(2): p. 409-16.

34. Ahn, D.G., et al., *TTSV1, a new virus-like particle isolated from the hyperthermophilic crenarchaeote Thermoproteus tenax.* Virology, 2006. 351(2): p. 280-90.

35. Haring, M., et al., *Morphology and genome organization of the virus PSV of the hyperthermophilic archaeal genera Pyrobaculum and Thermoproteus: a novel virus family, the Globuloviridae.* Virology, 2004. 323(2): p. 233-42.

36. Haring, M., et al., *Viral diversity in hot springs of Pozzuoli, Italy, and characterization of a unique archaeal virus, Acidianus bottle-shaped virus, from a new family, the Ampullaviridae.* J Virol, 2005. 79(15): p. 9904-11.

37. Rice, G., et al., *The structure of a thermophilic archaeal virus shows a double-stranded DNA viral capsid type that spans all domains of life.* PNAS, 2004. 101(20): p. 7716-7720.

38. Xiang, X., et al., *Sulfolobus tengchongensis Spindle-Shaped Virus STSV1: virus-host interactions and genomic features.* J. Virol., 2005. 79(14): p. 8677-8686.

39. Arnold, H.P., et al., *A novel lipothrixvirus, SIFV, of the extremely thermophilic crenarchaeon Sulfolobus.* Virology, 2000. 267(2): p. 252-66.

40. Ortmann, A.C., et al., *Transcriptome analysis of infection of the archaeon Sulfolobus solfataricus with Sulfolobus turreted icosahedral virus.* J Virol, 2008. 82(10): p. 4874-83.

41. Prangishvili, D., et al., *Structural and genomic properties of the hyperthermophilic archaeal virus ATV with an extracellular stage of the reproductive cycle.* J Mol Biol, 2006. 359(5): p. 1203-16.

42. Nadal M, M.G., Forterre P, Reiter WD, Duguet M, *Positively supercoiled DNA in a virus-like particle of an archaebacterium.* Nature, 1986. 321: p. 256-258.

43. Reiter, W.D., et al., *Gene expression in archaebacteria: physical mapping of constitutive and UV-inducible transcripts from the Sulfolobus virus-like particle SSV1.* Mol Gen Genet, 1987. 209(2): p. 270-5.

44. Frols, S., et al., *Elucidating the transcription cycle of the UV-inducible hyperthermophilic archaeal virus SSV1 by DNA microarrays.* Virology, 2007. 365(1): p. 48-59.

45. Wiedenheft, B., et al., *Comparative genomic analysis of hyperthermophilic archaeal Fuselloviridae viruses.* J Virol, 2004. 78(4): p. 1954-61.

46. Koonin, E.V., *Archaebacterial virus SSV1 encodes a putative DnaA-like protein.* Nucleic Acids Res, 1992. 20(5): p. 1143.

47. Muskhelishvili, G., P. Palm, and W. Zillig, *SSV1-encoded site-specific recombination system in Sulfolobus shibatae.* Mol Gen Genet, 1993. 237(3): p. 334-42.

48. Clore, A.J. and K.M. Stedman, *The SSV1 viral integrase is not essential.* Virology, 2007. 361(1): p. 103-11.

49. Kraft, P., et al., *Structure of D-63 from Sulfolobus Spindle-Shaped Virus 1: surface properties of the dimeric four-helix bundle suggest an adaptor protein function.* J. Virol., 2004. 78(14): p. 7438-7442.

50. Menon, S.K., et al., *Cysteine usage in Sulfolobus spindle-shaped virus 1 and extension to hyperthermophilic viruses in general.* Virology, 2008. 376(2): p. 270-8.

51. Kraft, P., et al., *Crystal structure of F-93 from Sulfolobus Spindle-Shaped Virus 1, a winged-helix DNA binding protein.* J. Virol., 2004. 78(21): p. 11544-11550.

52. Goulet, A., et al., *Crystallization and preliminary X-ray diffraction analysis of protein 14 from Sulfolobus islandicus filamentous virus (SIFV).* Acta Crystallogr Sect F Struct Biol Cryst Commun, 2006. 62(Pt 9): p. 884-6.

53. Larson, E.T., et al., *Structure of A197 from Sulfolobus turreted icosahedral virus: a crenarchaeal viral glycosyltransferase exhibiting the GT-A fold.* J. Virol., 2006. 80(15): p. 7636-7644.

54. Larson, E.T., et al., *A winged-helix protein from Sulfolobus Turreted Icosahedral Virus points toward stabilizing disulfide bonds in the intracellular proteins of a hyperthermophilic virus.* Submitted.

55. Zarembinski, T.I., et al., *Structure-based assignment of the biochemical function of a hypothetical protein: a test case of structural genomics.* Proc Natl Acad Sci U S A, 1998. 95(26): p. 15189-93.

56. Zhang, C. and S.H. Kim, *Overview of structural genomics: from structure to function.* Curr Opin Chem Biol, 2003. 7(1): p. 28-32.

57. Invitrogen Corporation, *Gateway Technology: a universal technology to clone DNA sequences for functional analysis and expression in multiple systems*. 2003.

58. Larson, E., *X-ray crystallographic studies of Sulfolobus Turetted Icosahedral Virus (STIV); a hyperthermophilic virus from Yellowstone National park.*, in *Department of Chemistry and Biochemistry*. 2006, Montana State University: Bozeman.

59. Bradford, M.M., *A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding.* Anal Biochem., 1976. 72: p. 248-54.

60. Gauss, G.H., et al., *Structure of the DPS-like protein from Sulfolobus solfataricus reveals a bacterioferritin-like dimetal binding site within a DPS-like dodecameric assembly.* Biochemistry, 2006. 45(36): p. 10815-27.

61. Ramakrishnan, V., et al., *Crystal structure of globular domain of histone H5 and its implications for nucleosome binding.* Nature, 1993. 362(6417): p. 219-23.

62. Leslie, A.G., *The integration of macromolecular diffraction data.* Acta Crystallogr D Biol Crystallogr, 2006. 62(Pt 1): p. 48-57.

63. Evans, P., *Scaling and assessment of data quality.* Acta Crystallogr D Biol Crystallogr, 2006. 62(Pt 1): p. 72-82.

64. Terwilliger, T.C. and J. Berendzen, *Automated MAD and MIR structure solution.* Acta Crystallogr., 1999. D55: p. 849-861.

65. Terwilliger, T.C., *Maximum likelihood density modification.* Acta Cryst, 2000. D56: p. 965-972.

66. Jones, T.A., et al., *Improved methods for building protein models in electron density maps and the location of errors in these models.* Acta Crystallogr., 1991. A47: p. 110-119.

67. Bailey, S., *The CCP4 suite - programs for protein crystallography.* Acta Cryst., 1994. D50: p. 760-763.

68. Murshudov, G.N., A.A. Vagin, and E.J. Dodson, *Refinement of macromolecular structures by the maximum-likelihood method.* Acta Cryst., 1997. D53: p. 240-255.

69. Painter, J. and E.A. Merritt, *TLSMD web server for the generation of multi-group TLS models.* Journal of Applied Crystallography, 2006. 39(1): p. 109-111.

70. Painter, J. and E.A. Merritt, *Optimal description of a protein structure in terms of multiple groups undergoing TLS motion.* Acta Crystallogr D Biol Crystallogr, 2006. 62(Pt 4): p. 439-50.

71. Holm, L. and C. Sander, *Protein structure comparison by alignment of distance matrices.* J Mol Biol, 1993. 233(1): p. 123-38.

72. Gibrat, J.-F., T. Madej, and S.H. Bryant, *Surprising similarities in structure comparison.* Current Opinion in Structural Biology, 1996. 6(3): p. 377-385.

73. Kabsch, W., *A solution for the best rotation to relate two sets of vectors.* Acta Cryst., 1976. A32: p. 922-923.

74. DeLano, W.L., *The PyMOL Molecular Graphics System.* http://www.pymol.org. 2002.

75. Peng, X., et al., *Genome of the Acidianus bottle-shaped virus and insights into the replication and packaging mechanisms.* Virology, 2007. 364(1): p. 237-43.

76. Bendtsen, J.D., et al., *Improved prediction of signal peptides: SignalP 3.0.* J Mol Biol, 2004. 340(4): p. 783-95.

77. Krogh, A., et al., *Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes.* J Mol Biol, 2001. 305(3): p. 567-80.

78. Sonnhammer, E.L., G. von Heijne, and A. Krogh, *A hidden Markov model for predicting transmembrane helices in protein sequences.* Proc Int Conf Intell Syst Mol Biol, 1998. 6: p. 175-82.

79. Gattiker, A., E. Gasteiger, and A. Bairoch, *ScanProsite: a reference implementation of a PROSITE scanning tool.* Appl Bioinformatics, 2002. 1(2): p. 107-8.

80. Hulo, N., et al., *The PROSITE database.* Nucleic Acids Res, 2006. 34(Database issue): p. D227-30.

81. Sigrist, C.J., et al., *PROSITE: a documented database using patterns and profiles as motif descriptors.* Brief Bioinform, 2002. 3(3): p. 265-74.

82. Vedadi, M., et al., *Chemical screening methods to identify ligands that promote protein stability, protein crystallization, and structure determination.* Proc Natl Acad Sci U S A, 2006. 103(43): p. 15835-40.

83. Yeh, A.P., A. McMillan, and M.H. Stowell, *Rapid and simple protein-stability screens: application to membrane proteins.* Acta Crystallogr D Biol Crystallogr, 2006. 62(Pt 4): p. 451-7.

84. Otwinowski, Z. and W. Minor, *Processing of X-ray diffraction data collected in oscillation mode*, in *Macromolecular Crystallography, Part A*, C. Carter and R. Sweet, Editors. 1997, Academic Press. p. 307-326.

85. Emsley, P. and K. Cowtan, *Coot: model-building tools for molecular graphics.* Acta Crystallographica Section D, 2004. 60(12 Part 1): p. 2126-2132.

86. Davis, I.W., et al., *MolProbity: all-atom contacts and structure validation for proteins and nucleic acids.* Nucleic Acids Res, 2007. 35(Web Server issue): p. W375-83.

87. Krissinel, E. and K. Henrick, *Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions.* Acta Crystallogr D Biol Crystallogr, 2004. 60(Pt 12 Pt 1): p. 2256-68.

88. Birkenbihl, R.P., et al., *Holliday junction resolving enzymes of archaeal viruses SIRV1 and SIRV2.* J Mol Biol, 2001. 309(5): p. 1067-76.

89. Casey Schlenker, S.M., C. Martin Lawrence, and Valérie Copié *1H, 13C, 15N Backbone and Sidechain Resonance Assignments for E73 from Sulfolobus Spindle Shaped Virus - Ragged Hills, a Hyperthermophilic Crenarchaeal Virus from Yellowstone National Park.* JBNMR, 2009. Submitted.

90. Thomas, J.G., A. Ayling, and F. Baneyx, *Molecular chaperones, folding catalysts, and the recovery of active recombinant proteins from E. coli. To fold or to refold.* Appl Biochem Biotechnol, 1997. 66(3): p. 197-238.

91. Studier, F.W., *Protein production by auto-induction in high density shaking cultures.* Protein Expr Purif, 2005. 41(1): p. 207-34.

92. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic Acids Res., 1997. 25: p. 3389-3402.

93. Hendrickson, W.A., and Ogata, C. M. , *Phase determination from multiwavelength anomalous diffraction measurements.* In "Macromolecular Crystallography, Part A" (C. W. Carter, and R. M. Sweet, Eds.), Academic Press, San Diego., 1997. 276: p. 494-523.

94. Hendrickson, W.A., Horton, J. R., and LeMaster, D. M., *Selenomethionyl proteins produced for analysis by multiwavelength anomalous diffraction (MAD): a vehicle for direct determination of three-dimensional structure. .* Embo J 1990. 9(5): p. 1665-72.

95. Aravind, L., et al., *The many faces of the helix-turn-helix domain: Transcription regulation and beyond.* FEMS Microbiology Reviews, 2005. 29(2): p. 231-262.

96. Laskowski, R.A., et al., *PROCHECK: a program to check the stereochemical quality of protein structures.* J. Appl. Cryst., 1993. 26: p. 283-291.

97. Kunst, F., et al., *The complete genome sequence of the Gram-positive bacterium Bacillus subtilis.* Nature, 1997. 390(6657): p. 249-256.

98. Larson, E.T., et al., *A winged-helix protein from Sulfolobus turreted icosahedral virus points toward stabilizing disulfide bonds in the intracellular proteins of a hyperthermophilic virus.* Virology, 2007. 368(2): p. 249-61.

99. Kadokura, H., F. Katzen, and J. Beckwith, *Protein disulfide bond formation in prokaryotes.* Annu Rev Biochem, 2003. 72: p. 111-35.

100. Aravind, L. and E.V. Koonin, *DNA-binding proteins and evolution of transcription regulation in the archaea.* Nucleic Acids Res, 1999. 27(23): p. 4658-70.

101. Cole, C., J.D. Barber, and G.J. Barton, *The Jpred 3 secondary structure prediction server.* Nucleic Acids Res, 2008. 36(Web Server issue): p. W197-201.

102. Vogt, G. and P. Argos, *Protein thermal stability: hydrogen bonds or internal packing?* Fold Des, 1997. 2(4): p. S40-6.

103. Helmer-Citterich, M., et al., *Control of ColE1 replication: low affinity specific binding of Rop (Rom) to RNAI and RNAII.* Embo J, 1988. 7(2): p. 557-66.

104. Banner, D.W., M. Kokkinidis, and D. Tsernoglou, *Structure of the ColE1 rop protein at 1.7 A resolution.* J Mol Biol, 1987. 196(3): p. 657-75.

105. Lawrence C.M, M.S., Eilers B.J, Bothner B, Khayat Reza, Douglas T and Young M.J., *Structural and Functional Studies of Archaeal Viruses.* Journal of Biological Chemistry, 2009. 284(19): p. 12599-603.

106. Xue, H., et al., *An abundant DNA binding protein from the hyperthermophilic archaeon Sulfolobus shibatae affects DNA supercoiling in a temperature-dependent fashion.* J Bacteriol, 2000. 182(14): p. 3929-33.

107. Larson, E.T., et al., *A new DNA binding protein highly conserved in diverse crenarchaeal viruses.* Virology, 2007. 363(2): p. 387-96.

108. Beeby, M., et al., *The genomics of disulfide bonding and protein stabilization in thermophiles.* PLoS Biology, 2005. 3(9): p. 1549-1558.

109. Mallick, P., et al., *Genomic evidence that the intracellular proteins of archaeal microbes contain disulfide bonds.* PNAS, 2002. 99(15): p. 9679-9684.

110. Gilbert, H.F., *Thiol/disulfide exchange equilibria and disulfide bond stability.* Methods Enzymol, 1995. 251: p. 8-28.

111. Larson, E.T., et al., *A winged-helix protein from sulfolobus turreted icosahedral virus points toward stabilizing disulfide bonds in the intracellular proteins of a hyperthermophilic virus.* Virology, 2007.

112. Peng, X., et al., *Sequences and replication of genomes of the archaeal Rudiviruses SIRV1 and SIRV2: Relationships to the archaeal Lipothrixvirus SIFV and some eukaryal viruses.* Virology, 2001. 291(2): p. 226-234.

113. O'Connor, B.D. and T.O. Yeates, *GDAP: a web tool for genome-wide protein disulfide bond prediction.* Nucl. Acids Res., 2004. 32(suppl_2): p. W360-364.

114. Senkevich, T.G., et al., *A viral member of the ERV1/ALR protein family participates in a cytoplasmic pathway of disulfide bond formation.* Proc Natl Acad Sci U S A, 2000. 97(22): p. 12068-73.

115. Bond, C.S., et al., *Structure of Hjc, a Holliday junction resolvase, from Sulfolobus solfataricus.* Proc Natl Acad Sci U S A, 2001. 98(10): p. 5509-14.

116. Senkevich, T.G., A.S. Weisberg, and B. Moss, *Vaccinia virus E10R protein is associated with the membranes of intracellular mature virions and has a role in morphogenesis.* Virology, 2000. 278(1): p. 244-52.

117. Maaty, W.S.A., et al., *Characterization of the archaeal thermophile Sulfolobus turreted icosahedral virus validates an evolutionary link among double-stranded DNA viruses from all domains of life.* J. Virol., 2006. 80(15): p. 7625-7635.

118. Gross, E., et al., *Structure of Ero1p, source of disulfide bonds for oxidative protein folding in the cell.* Cell, 2004. 117(5): p. 601-10.

119. Senkevich, T.G., et al., *Expression of the vaccinia virus A2.5L redox protein is required for virion morphogenesis.* Virology, 2002. 300(2): p. 296-303.

120. Senkevich, T.G., et al., *Complete pathway for protein disulfide bond formation encoded by poxviruses.* Proc Natl Acad Sci U S A, 2002. 99(10): p. 6667-72.

121. Iyer, L.M., L. Aravind, and E.V. Koonin, *Common origin of four diverse families of large eukaryotic DNA viruses.* J Virol, 2001. 75(23): p. 11720-34.

122. Ladenstein, R. and B. Ren, *Reconsideration of an early dogma, saying "there is no evidence for disulfide bonds in proteins from archaea".* Extremophiles, 2008. 12(1): p. 29-38.

123. Pedone, E., et al., *Characterization of a multifunctional protein disulfide oxidoreductase from Sulfolobus solfataricus.* Febs J, 2006. 273(23): p. 5407-20.

124. Lingaraju, G.M., et al., *A DNA glycosylase from Pyrobaculum aerophilum with an 8-oxoguanine binding mode and a noncanonical helix-hairpin-helix structure.* Structure, 2005. 13(1): p. 87-98.

125. Toth, E.A., et al., *The crystal structure of adenylosuccinate lyase from Pyrobaculum aerophilum reveals an intracellular protein with three disulfide bonds.* J Mol Biol, 2000. 301(2): p. 433-50.

126. Yoneda, K., et al., *The first crystal structure of an archaeal helical repeat protein.* Acta Crystallogr Sect F Struct Biol Cryst Commun, 2005. 61(Pt 7): p. 636-9.

127. Appleby, T.C., et al., *Three-dimensional structure of a hyperthermophilic 5'-deoxy-5'-methylthioadenosine phosphorylase from Sulfolobus solfataricus.* J Biol Chem, 2001. 276(42): p. 39232-42.

128. Cacciapuoti, G., et al., *Purification and characterization of 5'-methylthioadenosine phosphorylase from the hyperthermophilic archaeon Pyrococcus furiosus: substrate specificity and primary structure analysis.* Extremophiles, 2003. 7(2): p. 159-68.

129. Cacciapuoti, G., et al., *Methylthioadenosine phosphorylase from the archaeon Pyrococcus furiosus. Mechanism of the reaction and assignment of disulfide bonds.* Eur J Biochem, 2004. 271(23-24): p. 4834-44.

130. Cacciapuoti, G., et al., *Purification and characterization of extremely thermophilic and thermostable 5'-methylthioadenosine phosphorylase from the archaeon Sulfolobus solfataricus. Purine nucleoside phosphorylase activity and evidence for intersubunit disulfide bonds.* J Biol Chem, 1994. 269(40): p. 24762-9.

131. DeDecker, B.S., et al., *The crystal structure of a hyperthermophilic archaeal TATA-box binding protein.* J Mol Biol, 1996. 264(5): p. 1072-84.

132. Feese, M.D., et al., *Crystal structure of glycosyltrehalose trehalohydrolase from the hyperthermophilic archaeum Sulfolobus solfataricus.* J Mol Biol, 2000. 301(2): p. 451-64.

133. Kumar, S. and R. Nussinov, *How do thermophilic proteins deal with heat?* Cell Mol Life Sci, 2001. 58(9): p. 1216-33.

134. Scandurra, R., et al., *Protein thermostability in extremophiles.* Biochimie, 1998. 80(11): p. 933-41.

135. Tigerstrom, A., *Thermostability of proteins.* BIOS, 2005. 76(1): p. 22-27.

136. Vetriani, C., et al., *Protein thermostability above 100 degreesC: a key role for ionic interactions.* Proc Natl Acad Sci U S A, 1998. 95(21): p. 12300-5.

137. Karshikoff, A. and R. Ladenstein, *Ion pairs and the thermotolerance of proteins from hyperthermophiles: a "traffic rule" for hot roads.* Trends Biochem Sci, 2001. 26(9): p. 550-6.

138. Higgins, D.G., *CLUSTAL V: multiple alignment of DNA and protein sequences.* Methods Mol Biol, 1994. 25: p. 307-18.

139. Laskowski, R.A., *PDBsum new things.* Nucleic Acids Res, 2009. 37(Database issue): p. D355-9.

140. Niv, M.Y., et al., *Topology of Type II REases revisited; structural classes and the common conserved core.* Nucleic Acids Res, 2007. 35(7): p. 2227-37.

141. Joshi, H.K., et al., *Alteration of sequence specificity of the type II restriction endonuclease HincII through an indirect readout mechanism.* J Biol Chem, 2006. 281(33): p. 23852-69.

142. Selent, U., et al., *A site-directed mutagenesis study to identify amino acid residues involved in the catalytic function of the restriction endonuclease EcoRV.* Biochemistry, 1992. 31(20): p. 4808-15.

143. Galburt, E.A. and B.L. Stoddard, *Catalytic mechanisms of restriction and homing endonucleases.* Biochemistry, 2002. 41(47): p. 13851-60.

144. Feder, M. and J.M. Bujnicki, *Identification of a new family of putative PD-(D/E)XK nucleases with unusual phylogenomic distribution and a new type of the active site.* BMC Genomics, 2005. 6(1): p. 21.

145. Little, E.J. and N.C. Horton, *DNA-induced conformational changes in type II restriction endonucleases: the structure of unliganded HincII.* J Mol Biol, 2005. 351(1): p. 76-88.

146. Nishino, T., et al., *Crystal structure of the archaeal holliday junction resolvase Hjc and implications for DNA recognition.* Structure, 2001. 9(3): p. 197-204.

147. Baker, N.A., et al., *Electrostatics of nanosystems: application to microtubules and the ribosome.* Proc Natl Acad Sci U S A, 2001. 98(18): p. 10037-41.

148. Raaijmakers, H., et al., *X-ray structure of T4 endonuclease VII: a DNA junction resolvase with a novel fold and unusual domain-swapped dimer architecture.* Embo J, 1999. 18(6): p. 1447-58.

149. Ban, C. and W. Yang, *Structural basis for MutH activation in E.coli mismatch repair and relationship of MutH to restriction endonucleases.* Embo J, 1998. 17(5): p. 1526-34.

150. Tsutakawa, S.E., et al., *Crystallographic and functional studies of very short patch repair endonuclease.* Mol Cell, 1999. 3(5): p. 621-8.

151. Kovall, R. and B.W. Matthews, *Toroidal structure of lambda-exonuclease.* Science, 1997. 277(5333): p. 1824-7.

152. Middleton, C.L., et al., *Substrate recognition and catalysis by the Holliday junction resolving enzyme Hje.* Nucleic Acids Res, 2004. 32(18): p. 5442-51.

153. Hickman, A.B., et al., *Unexpected structural diversity in DNA recombination: the restriction endonuclease connection.* Mol Cell, 2000. 5(6): p. 1025-34.

154. Zhao, L., et al., *The restriction fold turns to the dark side: a bacterial homing endonuclease with a PD-(D/E)-XK motif.* Embo J, 2007. 26(9): p. 2432-42.

155. Yang, J., H.S. Malik, and T.H. Eickbush, *Identification of the endonuclease domain encoded by R2 and other site-specific, non-long terminal repeat retrotransposable elements.* Proc Natl Acad Sci U S A, 1999. 96(14): p. 7847-52.

156. Ramaswamy, K.S., et al., *Cell-type specificity of the Anabaena fdxN-element rearrangement requires xisH and xisI.* Mol Microbiol, 1997. 23(6): p. 1241-9.

157. Aravind, L., K.S. Makarova, and E.V. Koonin, *SURVEY AND SUMMARY: holliday junction resolvases and related nucleases: identification of new families,*

phyletic distribution and evolutionary trajectories. Nucleic Acids Res, 2000. 28(18): p. 3417-32.

158. Neef, K., R.P. Birkenbihl, and B. Kemper, *Holliday junction-resolving enzymes from eight hyperthermophilic archaea differ in reactions with cruciform DNA.* Extremophiles, 2002. 6(5): p. 359-67.

159. Declais, A.C. and D.M. Lilley, *New insight into the recognition of branched DNA structure by junction-resolving enzymes.* Curr Opin Struct Biol, 2008. 18(1): p. 86-95.

160. Jeltsch, A., et al., *Substrate-assisted catalysis in the cleavage of DNA by the EcoRI and EcoRV restriction enzymes.* Proc Natl Acad Sci U S A, 1993. 90(18): p. 8499-503.

161. Pingoud, A., et al., *Type II restriction endonucleases: structure and mechanism.* Cell Mol Life Sci, 2005. 62(6): p. 685-707.

162. Khuu, P.A., et al., *The stacked-X DNA Holliday junction and protein recognition.* J Mol Recognit, 2006. 19(3): p. 234-42.

163. Biertumpfel, C., W. Yang, and D. Suck, *Crystal structure of T4 endonuclease VII resolving a Holliday junction.* Nature, 2007. 449(7162): p. 616-20.

164. Hadden, J.M., et al., *The structural basis of Holliday junction resolution by T7 endonuclease I.* Nature, 2007. 449(7162): p. 621-4.

165. Declais, A.C., et al., *Structural recognition between a four-way DNA junction and a resolving enzyme.* J Mol Biol, 2006. 359(5): p. 1261-76.

166. Li, L., S. Lin, and F. Yanga, *Functional identification of the non-specific nuclease from white spot syndrome virus.* Virology, 2005. 337(2): p. 399-406.

167. Feighny, R.J., B.E. Henry, 2nd, and J.S. Pagano, *Epstein-Barr virus-induced deoxynuclease and the reutilization of host-cell DNA degradation products in viral DNA replication.* Virology, 1981. 115(2): p. 395-400.

168. Rice, A.P. and B.E. Roberts, *Vaccinia virus induces cellular mRNA degradation.* J Virol, 1983. 47(3): p. 529-39.

169. Everly, D.N., Jr., et al., *mRNA degradation by the virion host shutoff (Vhs) protein of herpes simplex virus: genetic and biochemical evidence that Vhs is a nuclease.* J Virol, 2002. 76(17): p. 8560-71.

170. Flick, K.E., et al., *DNA binding and cleavage by the nuclear intron-encoded homing endonuclease I-PpoI.* Nature, 1998. 394(6688): p. 96-101.

171. Muskhelishvili, G., *The Archaeal SSV integrase promotes Intermolecular excisive recombination in vitro* System. Appl. Microbiol, 1994. 16: p. 605-608.

172. Bujnicki, J.M. and L. Rychlewski, *Grouping together highly diverged PD-(D/E)XK nucleases and identification of novel superfamily members using structure-guided alignment of sequence profiles.* J Mol Microbiol Biotechnol, 2001. 3(1): p. 69-72.

173. Reiter, W.D. and P. Palm, *Identification and characterization of a defective SSV1 genome integrated into a tRNA gene in the archaebacterium Sulfolobus sp. B12.* Mol Gen Genet, 1990. 221(1): p. 65-71.