

X-RAY CRYSTALLOGRAPHIC STUDIES OF
SULFOLOBUS TURETTED ICOSAHERAL VIRUS (STIV); A
HYPERTHERMOPHILIC VIRUS FROM YELLOWSTONE NATIONAL PARK

by

Eric Thomas Larson

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

of

Doctor of Philosophy

in

Biochemistry

MONTANA STATE UNIVERSITY
Bozeman, Montana

November 2006

©COPYRIGHT

by

Eric Thomas Larson

2006

All Rights Reserved

APPROVAL

of a dissertation submitted by

Eric Thomas Larson

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."

Eric Thomas Larson

November 2006

ACKNOWLEDGEMENTS

I would like to thank my teachers, colleagues, family, and friends for inspiring and encouraging me in my education and research for all these years.

TABLE OF CONTENTS

1. INTRODUCTION	1
Archaea, the Third Domain of Life.....	1
Archaeal Viruses.....	5
<i>Sulfolobus</i> Turreted Icosahedral Virus (STIV).....	8
Research Goals and Overview	14
2. CLONING, EXPRESSION, PURIFICATION, AND CRYSTALLIZATION OF THE STIV GENOME.....	17
Introduction.....	17
Background.....	17
The Overall Scheme to Rapidly Identify Viral Proteins Most Suitable for X-ray Structure Determination	23
Cloning.....	24
Choice of Cloning System	24
Primer Design and PCR Amplification.....	25
Creation of the Entry and Expression Clones	28
Expression and Purification	32
Expression in <i>E. coli</i>	38
Purification.....	40
Crystallization.....	43
Materials and Methods for the Structure Solution of A197.....	45
Cloning the A197 Gene	45
Expression, Purification, and Characterization of A197	46
Crystallization of A197 and Collection of X-Ray Diffraction Data	48
X-ray Crystal Structure Determination and Model Refinement of A197.....	50
Atomic Coordinates of A197	52
Materials and Methods for the Structure Solution of B116.....	52
Cloning the B116 Gene.....	52
Expression, Purification, and Characterization of B116.....	53
Crystallization of B116 and Collection of X-Ray Diffraction Data	55
X-ray Crystal Structure Determination and Model Refinement of B116	55
Atomic Coordinates of B116	57
Materials and Methods for the Structure Solution of F93	58
Cloning the F93 Gene	58
Expression, Purification, and Characterization of F93	58
Crystallization of F93 and Collection of X-Ray Diffraction Data.....	60
X-ray Crystal Structure Determination and Model Refinement of F93	61
Atomic Coordinates of F93.....	63
Subcellular Localization of STIV Proteins	63
Summary	64

TABLE OF CONTENTS - CONTINUED

3. STRUCTURAL CHARACTERIZATION OF A197; A GLYCOSYTRANSFERASE DISPLAYING THE GT-A FOLD	71
Introduction	71
Results	72
Structure of A197	73
Structural Homology with Glycosyltransferases	75
Donor Substrate-binding Site	81
Acceptor Substrate-binding Site	85
Discussion	86
Conclusions	92
4. STRUCTURAL CHARACTERIZATION OF B116; A NEW FOLD IN A HIGHLY CONSERVED CRENARCHAEAL VIRAL PROTEIN	95
Introduction	95
Results	97
Structure of B116	97
Sequence Homology in B116-like Proteins	100
Discussion	107
Conclusions	112
5. STRUCTURAL CHARACTERIZATION OF F93; A WINGED HELIX PROTEIN THAT POINTS TOWARD STABILIZING DISULFIDE BONDS IN THE INTRACELLULAR PROTEINS OF A HYPERTHERMOPHILIC VIRUS	115
Introduction	115
Results	117
Structure of F93	117
Putative Interactions with DNA	123
The Intermolecular Disulfide Bond	124
Genomic Analysis of Cysteine Content	125
Discussion	128
wHTH DNA Binding Protein	128
Intracellular Disulfide Bonds	130
Conclusions	137
6. CONCLUDING REMARKS	141

TABLE OF CONTENTS - CONTINUED

REFERENCES CITED.....	146
-----------------------	-----

LIST OF TABLES

Table	Page
2-1. General STIV ORF Information Relevant for Cloning through Structure Determination and Virus Characterization.....	19
2-2. Primer Sequences Used for Cloning the STIV Genome	29
2-3. The Progress towards the Complete Structural Characterization of STIV	33
2-4. X-ray Diffraction Data Collection for A197	49
2-5. Crystal Structure Model Refinement for A197	51
2-6. X-ray Diffraction Data Collection for B116.....	56
2-7. Crystal Structure Model Refinement for B116.....	57
2-8. X-ray Diffraction Data Collection for F93	61
2-9. Crystal Structure Model Refinement for F93	63
2-10. STIV ORF Information Relevant to Disulfide Bonds in the Intracellular Proteins of the Viral Proteome.	65

LIST OF FIGURES

Figure	Page
1-1. The Universal Tree of Life	2
1-2. Rabbit Creek Thermal Area, Yellowstone National Park	3
1-3. The Crenarchaeote, <i>Sulfolobus solfataricus</i>	4
1-4. The Genome Map of STIV	9
1-5. STIV	10
1-6. The Structure of the MCP from STIV and Related Viruses.....	12
2-1. Gateway Cloning.....	25
2-2. Desired PCR Product and Universal Primer Sequences.....	26
2-3. The BP Reaction.....	30
2-4. The LR Reaction.....	31
2-5. STIV Vector Naming Conventions.	32
2-6. Set-up for Hanging Drop Vapor Diffusion Crystallization	44
3-1. Location of A197 Gene in STIV Genome	72
3-2. Structure of A197	74
3-3. A197 and the Equivalent Domains of Structural Neighbors	78
3-4. A197 and the GT-A Structural Core.....	79
3-5. Structure-based Sequence Alignment of A197 Structural Neighbors.....	80
3-6. Electrostatic Surface of A197 Homodimer with Superposed Substrates	82
3-7. Predicted Disorder and Flexibility of A197	84
3-8. Glycosylation of the Major Capsid Protein (MCP) of STIV.....	90

LIST OF FIGURES - CONTINUED

4-1. Location of the B116 Gene in STIV Genome	96
4-2. Structure of B116.....	99
4-3. Multiple Sequence Alignment of the B116 Family.....	101
4-4. Conserved Surface of the Viral B116-like Proteins.	102
4-5. Surface Features of the B116 Family.	104
4-6. Predicted Disorder and Flexibility of B116.....	106
4-7. Spacing Between the $\beta 4/\alpha 3$ Loops and the Complementarity to dsDNA	108
4-8. Nonspecific Interaction between B116 and DNA	109
5-1. Location of the B116 Gene in STIV Genome	116
5-2. Ribbon Diagrams of the F93 Homodimer	119
5-3. Predicted Disorder and Flexibility of F93	121
5-4. F93 and Representative Structural Neighbors in Complex with Target DNA.....	122
5-5. Analysis of F93 under Reducing and Non-reducing Conditions.....	125
5-6. Analysis of the Genomic Cysteine Content.....	127
5-7. Cysteine, Serine, and Leucine Content as a Function of Protein Size.....	134
6-1. The Progress of the Structural Characterization of STIV.....	143

ABSTRACT

Sulfolobus turreted icosahedral virus (STIV) was isolated from acidic hot springs of Yellowstone National Park and was the first hyperthermophilic virus described with icosahedral capsid architecture. Structural analysis of the STIV particle and its major capsid protein suggests that it belongs to a lineage of viruses that predates the division of the three domains of life. Functional predictions of the viral proteins are hindered because they lack similarity to sequences of known function. Protein structure, however, may suggest functional relationships that are not apparent from the sequence. Thus, we have initiated crystallographic studies of STIV and expect to gain functional insight into its proteins while illuminating the viral life cycle. These studies may also provide genetic, biochemical, and evolutionary insight into its thermoacidophilic host and the requirements for life in these harsh environments.

The first three proteins studied in structural detail are A197, B116, and F93. As anticipated, these structures suggest possible functions. The structure of A197 reveals a glycosyltransferase GT-A fold. Within the context of the GT-A fold, are the canonical DXD motif and a putative catalytic base, hallmarks of this family of enzymes, strongly suggesting glycosyltransferase activity for A197. B116 is a unique structure that lacks significant homology to known protein structures. However, sequence similarity to proteins from other hyperthermophilic viruses reveals conserved surface features suggesting interaction with a host macromolecule, likely DNA. The F93 structure reveals a winged-helix fold common among DNA-binding proteins, in particular, the MarR-like family of transcriptional regulators. The most likely role for F93 is thus regulation of viral transcription.

Interestingly, B116 contains an intramolecular disulfide bond while F93 contains an intermolecular disulfide bond. The presence of these disulfide bonds was not anticipated because these proteins are expected to be localized within the host cell. This prompted analysis of the cysteine distribution in the STIV genome, which suggests that disulfide bonds are common in intracellular (cytoplasmic) proteins encoded by STIV. This work is in accordance with accumulating evidence that disulfide bonds are common stabilizing elements in the intracellular proteins of thermophilic organisms in general, and extends the observation to genomes of hyperthermophilic viruses.

CHAPTER 1

INTRODUCTION AND BACKGROUND

The research detailed in this dissertation is the beginning of a journey toward the complete structural characterization of a hyperthermophilic virus that infects a crenarchaeal host. This chapter will provide some brief background into the Archaea and the viruses that infect them, and then introduce the particular virus that is the focus of this study, *Sulfolobus* turreted icosahedral virus (STIV), and the goals of this research project.

Archaea, the Third Domain of Life

The Archaea were originally “discovered” in 1977 when Carl Woese and George Fox were attempting to phylogenetically classify the Bacteria by comparing the sequences of their ribosomal RNAs (rRNA). Through the course of this analysis, they realized that the methanogens were fundamentally different from typical prokaryotes, the bacteria (42, 141). Further study of these unique organisms led to the proposal of a universal phylogenetic tree that was divided into three domains, the Bacteria, the Eukarya, and the Archaea, based primarily on 16S rRNA sequence comparisons (Figure 1-1) (142). The root of the tree points toward the last universal common ancestor (LUCA) of all present-day organisms and its placement is a topic of debate. Most evidence points toward the root being somewhere between the primary bacterial and archaeal branches and the rRNA-based tree has generally been accepted.

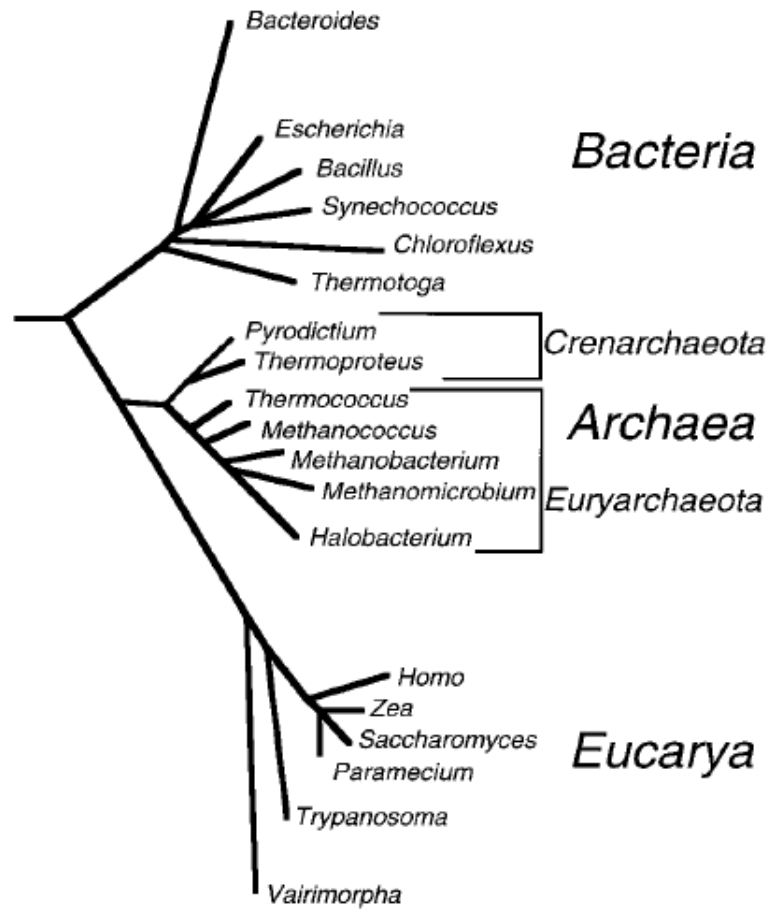


Figure 1-1. The Universal Tree of Life.

The universal tree is organized into three domains, with branching orders based upon rRNA sequence comparisons and the root placed by duplications in certain protein sequences. The two kingdoms (or phyla) of the Archaea are the Crenarchaeota and the Euryarchaeota. Figure from (89).

The domain Archaea is divided into two kingdoms (or more recently, phyla), the Euryarchaeota and the Crenarchaeota (142). The Euryarchaeota are primarily methanogens, methane-producing organisms, and halophiles, organisms that thrive in hypersaline environments. The range of environments inhabited by the methanogens is quite large and the majority of these organisms are strict anaerobes. Most live at

mesophilic temperatures though hyperthermophilic species do exist. The halophiles are found in almost all hypersaline environments and most are strict aerobes.



Figure 1-2. Rabbit Creek Thermal Area, Yellowstone National Park.

Viruses have been isolated from hyperthermophilic organisms that thrive in thermal features from around the world. STIV was isolated from *Sulfolobus* species that inhabit the acidic hot springs in the Rabbit Creek Thermal Area of Yellowstone National Park, USA. Dr. Mark Young took this photo.

The crenarchaeotes are primarily thermophilic organisms that thrive in temperatures above 80 °C. Thermophilic Crenarchaea are found in thermal features within Yellowstone National Park (Figure 1-2) and elsewhere around the world in both terrestrial and marine environments. However, Crenarchaeal organisms have been identified in mesophilic temperatures and even extremely low temperatures (12, 36). Though the majority is anaerobic, many of the best-characterized are aerobic, namely members of the order Sulfolobales. *Sulfolobus solfataricus* (Figure 1-3) has emerged as the model crenarchaeal organism, primarily due to the ease of culturing it in the

laboratory. It is prokaryotic, hyperthermophilic (optimal growth at $T > 80\text{ }^{\circ}\text{C}$), acidophilic (optimal growth at $\text{pH} < 4$), strictly aerobic, chemolithotrophic (oxidizes elemental sulfur, S^0 to H_2SO_4), and its genome has been completely sequenced (121). Furthermore, efforts have been made toward the production of genetic systems within *Sulfolobus* for microbiology applications (60, 126).

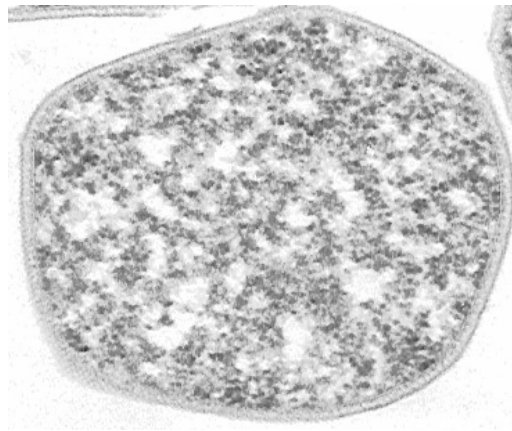


Figure 1-3. The Crenarchaeote, *Sulfolobus solfataricus*.

S. solfataricus has emerged as the model crenarchaeal organism. The majority of crenarchaeal (hyperthermophilic) viruses isolated to date have come from hosts who are close relatives of this species, likely reflecting a culture bias rather than the true extent of the diversity of these viruses.

Fairly recently, two more archaeal phyla, the Korarchaeota (12) and the Nanoarchaeota (55), have been proposed. Both appear to branch close to root of the archaeal branch based on 16S rRNA sequences and thrive in extremely hot environments but little else is known about them. The korarchaeotes have been identified in Yellowstone and elsewhere but no species have been cultured. The nanoarchaeotes are extremely small symbiotic organisms, barely larger than some of the largest viruses. They were originally isolated from a deep-sea vent but have also been found in

Yellowstone National Park. At least one species has been cultured. It is interesting to note that thermophily is common among the oldest branches of each of the primary phyla, however, so it is assumed that the ancestor common to all Archaea possessed this phenotype (142).

Currently, 32 archaeal genomes have been completely sequenced (119) (<http://archaea.ucsc.edu/cgi-bin/hgGateway>) and several more are forthcoming. Among many things, these genome sequences are highlighting commonalities between the Archaea and the other domains of life but are also emphasizing the extreme differences and validating their placement within their own domain of life. Despite the growing knowledgebase of the Archaea, however, relatively little is known about this domain of life when compared to the Bacteria and Eukarya. This is partly because they are relatively young as far as their recognition as a distinct life form is concerned, but also because the extreme environments many of these organisms inhabit make them difficult to study in the lab.

Archaeal Viruses

Viruses have been found to infect both the Euryarchaea and Crenarchaea. In sharp contrast to the large number of viruses identified from bacterial or eukaryotic hosts, less than 50 have been described from archaeal hosts, with roughly equal numbers from the Euryarchaea and the Crenarchaea (90). In addition to the limited effort that has been directed toward the discovery of archaeal viruses, this discrepancy in numbers is likely due to the difficulties in culturing most archaeal organisms and in isolating the viruses directly from the often harsh environments rather than a true reflection of their abundance

(123, 124). Nonetheless, our understanding of the viruses that replicate within archaeal hosts is clearly very limited and warrants further study.

The majority of viruses from the Euryarchaeota resemble typical bacterial phages with head-and-tail morphologies and many even belong to these established viral families (37). Of these, most have been isolated from halophiles but a growing number are being isolated from the methanogens. Though in the minority, it is interesting that several morphologically unique viruses have been isolated from the hyperthermophilic euryarchaeal order Thermococcales found in deep sea vents (107). Some of these morphologies resemble those found among the crenarchaeal viruses, such as PAV1 that resembles the spindle-shaped viruses from *Sulfolobus* (SSV1) (107). All known euryarchaeal viruses carry a double-stranded (ds) DNA genome that contains ORFs whose predicted products generally lack significant sequence similarity to those in the public databases (37).

In contrast to the euryarchaeal viruses, the vast majority of known viruses infecting the hyperthermophilic Crenarchaea display remarkably diverse and unique morphologies (103, 104, 110, 111, 124, 145). Linear, either stiff or flexible, and spindle-shaped morphologies predominate but many others are commonly observed (103, 104, 110). Similar to the euryarchaeal viruses, all known crenarchaeal viruses possess dsDNA genomes with predicted gene products that display limited sequence similarity to other known sequences. Sequence similarity that is present is usually confined between uncharacterized homologues within the genomes of other crenarchaeal viruses, typically within the same viral family (139). The GC content of the genomes is typically less than 40%, mirroring that of the hosts. Seven new viral families have been created to reflect

the unique morphological and genomic characteristics of these hyperthermophilic viruses. They are: *Fuselloviridae* (spindle-shaped particles such as SSV1) (95), *Rudiviridae* (stiff rod-shaped particles such as SIRV1) (102), *Lipothrixviridae* (flexible filamentous particles such as SIFV) (8), *Guttaviridae* (droplet-shaped particle such as SNDV) (7), *Globuloviridae* (spherical particles such as PSV) (49), *Ampullaviridae* (bottle-shaped particles such as ABV) (50), and *Bicaudaviridae* (two-tailed, spindle-shaped particles such as ATV) (52). In addition, many more distinct morphologies have been observed in enrichment cultures (104, 106, 110, 111), and some have even been characterized but do not fit into these families (112, 143). Thus, the number of unique viral families among the Crenarchaea is sure to grow.

Crenarchaeal viruses have been isolated from thermal features around the world, including those in Yellowstone National Park, and are thus considered thermophilic organisms like their hosts. The majority has been isolated from acidophilic hosts of the genera *Sulfolobus* and *Acidianus*, but many have also come from the neutrophilic hosts of the genera *Thermoproteus* and *Pyrobaculum*. Generally speaking, hyperthermophilic viruses do not lyse their hosts and live in equilibrium with their growth and replication cycles in a persistent state. This may be a survival strategy of these viruses serving to limit their exposure to harsh environmental conditions (103, 104).

Sulfolobus Turreted Icosahedral Virus (STIV)

The particular crenarchaeal virus that is the focus of the structural studies presented in this dissertation is *Sulfolobus* turreted icosahedral virus (STIV). It is an icosahedral virus distinct from all the established families of crenarchaeal viruses and so is likely the founding member of a new family. It infects *Sulfolobus* species (Figure 1-3) resident in the acidic hot springs (pH 2.9-3.9 and 72-92 °C) of Yellowstone National Park (Figure 1-2) and was first isolated from enrichment cultures by members of Dr. Mark Young's laboratory at Montana State University (111, 112). It encapsidates a circular, dsDNA genome of 17,663-bp containing 37 open reading frames (ORFs) (Figure 1-4), with 36 identified in the original annotation (112) and an additional ORF, A78 in a subsequent proteomics study of the purified virus (78). The major capsid protein, B345, was identified due to its abundance from purified viral particles. As is typical of crenarchaeal viruses, the predicted gene products generally exhibit a lack of similarity to sequences in the publicly available databases. This largely precludes functional assignment of the viral proteins and is a driving force behind the structural studies we have undertaken.

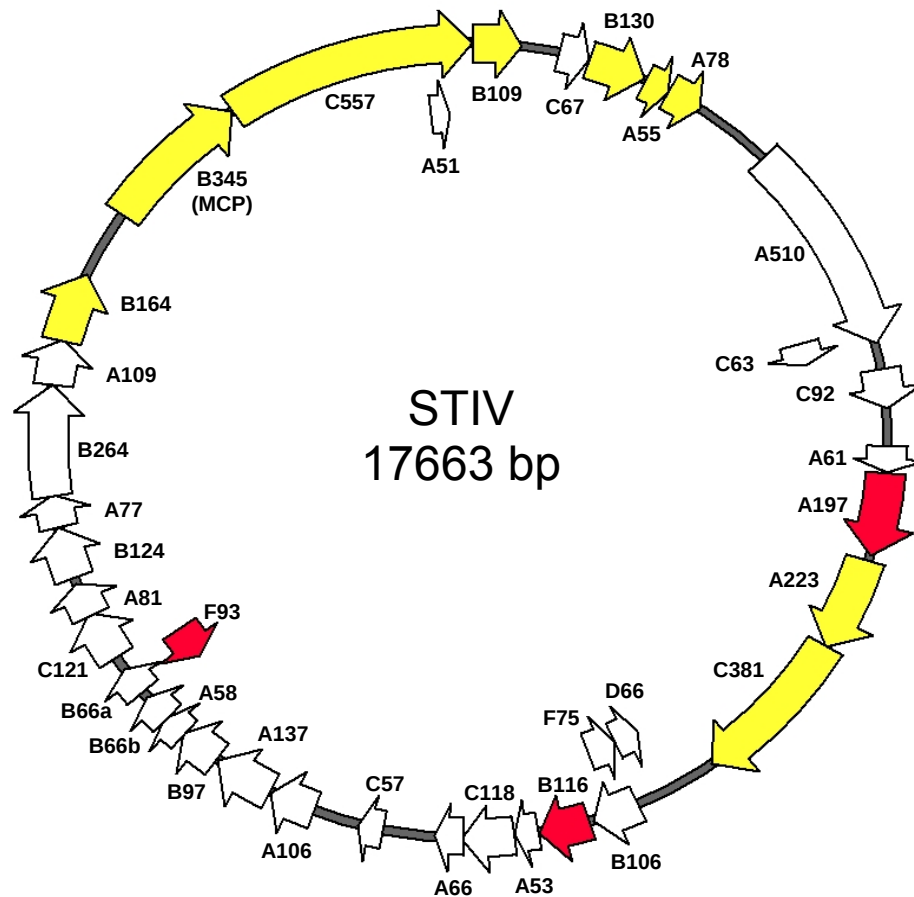


Figure 1-4. The Genome Map of STIV.

Predicted open reading frames (ORFs) are named according to the reading frame they are in (A-F) and the number of amino acids encoded. B345 was identified as the major capsid protein based on its abundance and has since had its structure determined. Eight additional ORF products have also been identified to co-purify with the virion particle (all the capsid-associated proteins are shaded in yellow). The three ORF products that are the focus of this research are shaded in pink. Figure was created with VectorNTI 10.

Though viruses displaying icosahedral capsid architecture are common among the Bacteria and Eukarya, STIV was the first to be described from an archaeal host. The cryo-electron microscopy (cryo-EM) reconstruction of the STIV particle reveals a pseudo T=31 icosahedral capsid with large turret-like projections extending from the 5-fold vertices and an apparent internal membrane (112) (Figure 1-5). The capsid architecture of STIV generated much interest because it is common to viruses from all three domains

of life, suggesting a viral lineage that shares a common ancestor that predates the fundamental evolutionary events that gave rise to the Archaea, Bacteria and Eukarya (10, 11, 15, 16, 112). Indeed the crystal structure of the STIV major capsid protein (MCP) reveals that it adopts the double-jellyroll fold (65) with stunning structural homology to the major capsid proteins of bacteriophage PRD1 (16), and the eukaryotic viruses PBCV-1 (87) and adenovirus (117), further supporting this evolutionary lineage (Figure 1-6) (65).

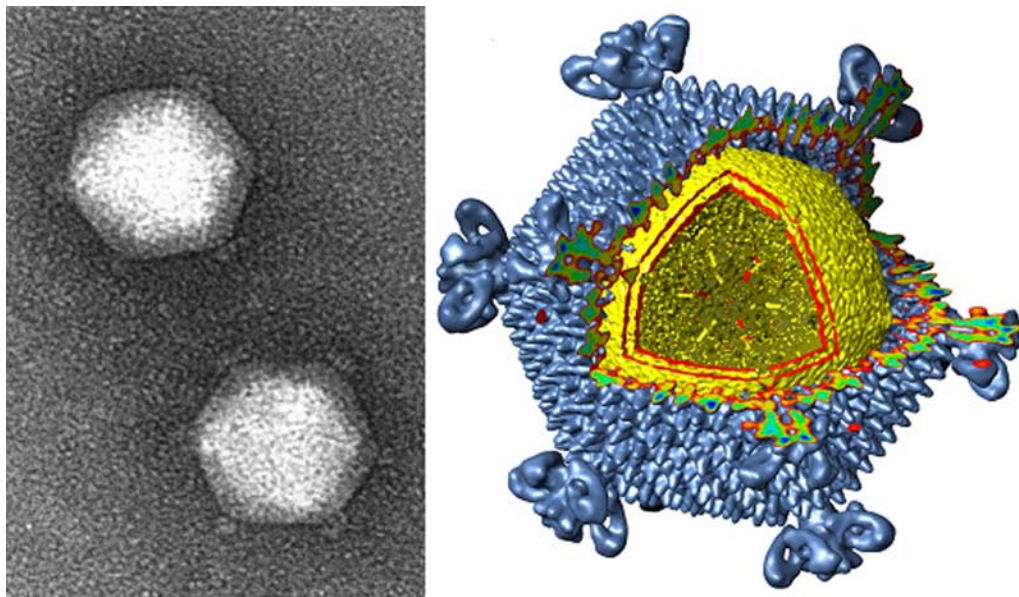


Figure 1-5. STIV.

The electron micrograph of STIV (left) and the cryo-EM reconstruction of the viral particle (right). STIV displays an icosahedral, pseudo T=31 architecture with prominent turret-like appendages extending from the 5-fold vertices. A quarter of the capsid shell has been removed from the cryo-EM image to show the internal lipid bilayer. Figure is a composite from (78, 111).

A lineage of evolutionarily related dsDNA virus that contain large T-number icosahedral capsids and spanning multiple domains had been suggested (10, 16), but prior

to the discovery of STIV, a member from the archaeal domain had not been seen. As the genomic similarities between these viruses [bacteriophage PRD1 (7), and the eukaryotic viruses PBCV-1 (41) and adenovirus (55), STIV (112), and perhaps the Euryarchaeal SH1 (101)] are practically nonexistent, this lineage is defined on the structure and assembly principles of the viral capsid itself, which must be evolutionarily maintained in the sequences. An ancestor virus likely existed before the divergence of the domains of life and through the course of evolution over billions of years, the sequences of the progeny have been altered due to specific pressures exerted by their respective hosts and environments and masked their common ancestry (10, 11, 15, 65, 112).

The crystal structure of the MCP also provided insight into how this protein achieves thermostability in the acidic and very hot extracellular environment in which it must spend at least a portion of its life (65). In comparison to the structures of the major capsid proteins of other viruses in the domain-spanning lineage, adenovirus hexon, PRD1 P3, and PBCV-1 vp54, STIV MCP (Figure 1-6) is much more tightly packed with much less total cavity volume. MCP also has extraordinary proline content and many of the prolines are in surface-exposed loops. Proline residues are believed to contribute to thermostability by reducing flexibility and decreasing the entropy of the unfolded state. It appears that improved hydrophobic packing and structural restraint are important for the thermostability of MCP while other common factors such as hydrogen bonding, electrostatic interactions, and disulfide bonds may not contribute as indeed these interactions will likely be disrupted in the harsh conditions of an acidic hot spring (65).

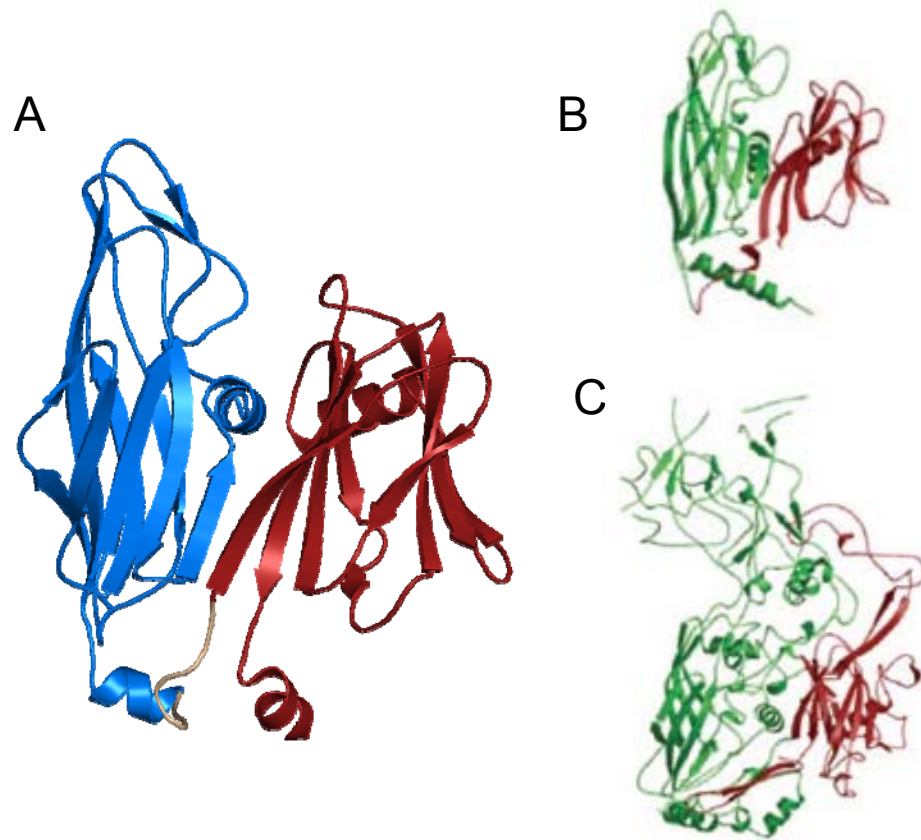


Figure 1-6. The Structure of the MCP from STIV and Related Viruses.

(A) The structure of the MCP from STIV adopts the double-jellyroll fold commonly used for viral capsid proteins. MCP looks remarkably similar to the major capsid proteins from related viruses in the domains: **(B)** Bacteria (P3 from PRD1); and **(C)** Eukarya (hexon from adenovirus). Panel A was made from the pdb (2BBD) with PyMol and B and C are from (112).

Proteomic studies of purified STIV virions have defined the subset of proteins that make up the virion and have helped to further characterize the virus and likely functions for a number of its predicted ORF products (78). In addition to B345, eight viral proteins (B109, B130, A55, A78, A223, C381, B164, C557; shaded yellow in Figure 1-4) and two host proteins (a small DNA-binding protein and a hypothetical protein thought to be a VPS24 vacuolar sorting protein) were identified. Based on the

proteomics/mass spectrometry results and a critical analysis of the cryo-EM density, it is suggested that A223, C381, and C557 likely interact and together comprise the pentamer subunits, the unique turret-like appendages, of the icosahedron. Sequence-based (58, 78, 105) and predictive structure-based analysis (78) suggest that B164 is a pumping ATPase similar to those found in the pox viruses. Interestingly, molecular modeling of this protein and its subsequent fitting into the cryo-EM density suggest that it makes up the base of the turret structure, filling density at the bottom of an open channel that is large enough to accommodate DNA (78, 112). It is thus proposed that this protein is involved in either the packaging of the viral genome or its delivery into the host cell.

The proteomic studies also confirmed the existence of a lipid bilayer within the STIV capsid and found that it is primarily composed of a specific subset of host-derived lipids (78). Further, this study established that the MCP is glycosylated in light of the results of our structural studies of A197 (to be discussed in Chapter 3). Surprisingly, the predicted glycosylation sites on the MCP correspond very well to the known glycosylation sites of the MCP (vp54) from PBCV-1 (87), a eukaryotic virus that is a member of the domain-spanning viral lineage. This proteomic analysis has clearly been a great contribution to the overall characterization of STIV and has extended the similarities observed between it and viruses in the other domains of life beyond the capsid architectures and MCP structures. However, despite this and all the other work on this interesting virus, the functions of the majority of viral proteins remain unknown.

Research Goals and Overview

The characterization of crenarchaeal viruses inhabiting the hot springs of Yellowstone National Park and other thermal features around the world is revealing fascinating diversity, apparent in both viral morphologies and genomic sequences, unseen in the viruses of the other domains of life. The uniqueness of these viruses is also a likely reflection of the unique organisms they infect and the unique environments they inhabit. Despite the growth in recent years of the number of viruses known to infect hyperthermophilic Crenarchaea, our knowledge of these viruses is still in its infancy when compared to similar studies from eukaryotic and bacterial viruses. This is particularly true at the molecular level, as we know very little in regards to their viral life cycles, virus-host relationships, genetics, or biochemistry. Thus, a significant understanding of even a single crenarchaeal virus would be a great advance and further study is clearly warranted.

This research project seeks to contribute to the molecular and structural characterization of STIV, and in turn a detailed understanding of its viral lifecycle. In particular, we would like to determine the functions of the products of all of its ORFs and to try to relate these functions to the viral lifecycle and the virus-host relationship. The functions of the majority of the predicted proteins coded by the STIV genome remain unknown. The identities of the nine proteins that comprise the virion are known but the functions of many of these are still undetermined. The deficiency of known functions for the STIV proteins is primarily due to their lack of obvious sequence similarity to proteins of known function. However, sequence analysis by our laboratory and others suggests

the presence of several classes of transcriptional regulators [M. Dlakic, personal communication and (105)], an ATPase (58, 78, 105), and two small C2H2 Zn-finger proteins (105). Functions for the remaining hypothetical proteins largely remain a mystery.

It is highly unlikely that most of the STIV proteome consists of unique protein folds serving novel functions. Rather, the general lack of sequence similarity in the STIV ORFs is probably due to the unique environment in which the virus lives and to the evolutionary distance between them and their bacterial and eukaryotic homologues. With respect to protein evolution, it is well established that similarity between related protein structures will persist longer than similarities in their corresponding amino acid sequences (85). Further, there exists an intimate relationship between protein structure and function with each exerting evolutionary constraints upon a protein. Thus, knowledge of the three-dimensional structure of a protein may allow for the inference of functional and evolutionary insight that is not readily apparent in the primary sequence (6, 15, 16, 23, 65, 67, 68, 72, 85, 116, 135, 136).

To this end, we have initiated x-ray crystallographic studies of the STIV proteome. It is expected that a thorough structural analysis will provide specific suggestions regarding the functions of these proteins and the roles they play in the STIV lifecycle. Furthermore, since a small subset of the STIV proteins have homologues within other crenarchaeal viruses of unrelated viral families, the knowledge afforded by these structures will extend beyond this single virus and provide insight into factors that are important in the lifecycles of the crenarchaeal viruses in general. Importantly, while such studies will lead to a greater understanding of the viruses themselves, they are also

expected to provide genetic, biochemical, and evolutionary insight into their crenarchaeal hosts and the requirements for life in the harsh environments in which these organisms often thrive.

CHAPTER 2

CLONING, EXPRESSION, PURIFICATION, AND CRYSTALLIZATION
OF THE STIV GENOMEIntroductionBackground

The full 17,663 base pair genome sequence of STIV was obtained from George Rice in Mark Young's laboratory in the Department of Plant Sciences and Plant Pathology at Montana State University while the initial characterization of the virus was ongoing. The major capsid protein was identified from its abundance in the purified viral particle and determined to be the product of ORF B345 but the functions of the remaining ORFs remained a mystery, primarily due to the lack of homology to sequences in the publicly available databases (112). Knowing that protein function is largely determined by its three-dimensional shape, we set out to determine the x-ray crystallographic structures of the STIV proteome because it may be possible to gain insight into the proteins' functions and their roles in the viral lifecycle based on structural homology to proteins of known function. If structurally similar proteins are not available, we still may gain insight into the functions based on the overall shape of the protein, its surface features, etc. that are apparent in the 3-D model of the protein. The field of archaeal virology is relatively young so the knowledgebase about these interesting and unique viruses is severely lacking compared to that of viruses in the domains Eukarya and

Bacteria. No virus in the domain Archaea has been studied in structural detail so the envisioned studies of STIV would represent an important advancement in the field.

The STIV genome was originally annotated with fifty putative ORFs coding for proteins with greater than fifty amino acids and varying in length from 51 to 557 amino acids. Since the annotation was not finalized, we created an internal numbering system for each gene starting with B109 being called #1 and each consecutive ORF on that strand numbered sequentially in a clockwise fashion. When all ORFs on that strand were numbered, the numbering continued for the ORFs on the opposite strand, again going clockwise from the same starting position. Due to the facts that the final ORF annotations were still in progress and many of the putative ORFs, in particular several of the smaller ones, overlapped with other ORFs, we decided to concentrate our cloning efforts on the 38 ORFs coding for proteins greater than sixty amino acids. General information about each of the STIV ORFs is presented in Table 2-1.

Table 2-1. General STIV ORF Information Relevant for Cloning through Structure Determination and Virus Characterization.

Internal ORF ID	ORF #	ORF name	NT seq. in genome	MW (Da) ^a	Theoretical pI ^a	MW+his tag (Da) ^a	Theoretical pI +His tag ^a	# Met (minus N- term. M)	% Met (with His-tag, minus N-term. M)	Predicted function ^c
ORFs of nonstructural proteins^b that have been cloned.										
	2	C67	588-791	7,159.44	9.74	7,982.28	9.74	1	1.37	unknown
	5	A510	2167-3699	60,217.92	9.29	61,040.76	9.29	7	1.36	tyrosine recombinase/integrase (CD hits; Rice, et. al. 2004)
	6	C63	3552-3743	7,223.43	11.49	8,046.27	11.49	0	0.00	unknown
	7	C92	3888-4166	9,834.24	5.11	10,657.09	6.69	2	2.04	unknown (high seq. conservation with SIRV1 and SIRV2)
	8	A61	4432-4617	7,378.90	8.48	8,201.74	8.52	3	4.48	NikR/CopG-like RHH transcriptional regulator (CD hits; high seq. similarity to ARV1, <i>S. solfataricus</i> P2, SIRV1, <i>S.</i> <i>tokodaii</i> , <i>S. acidocaldarius</i> , SIRV2, ATV, AFV1, <i>S. tengchongensis</i> , <i>Sulfolobus</i> sp. NOB8H2, AFV2, <i>S. islandicus</i> , SSVRH (ORF B64), STIV(A137), SIFV (96), SSV2 (ORF 82a), SSV1 (ORF C-80), SSV-K (ORF C81), <i>S. tokodaii</i> , and many other Archaea)
	9	A197	4624-5217	23,266.95	6.51	24,089.80	6.93	3	1.48	glycosyltransferase (Larson, et. al. 2006)
	12	B106	7646-7966	12,553.71	8.79	13,376.56	8.79	1	0.89	unknown
	13	B116	7994-8344	13,330.43	4.69	14,153.28	5.76	4	3.28	DNA-binding protein? (structural insight and seq. conservation; high seq. conservation among SIRV1, SIRV2, AFV1, ARV1, SIFV, ATV, unchar. YddF of <i>B.</i> <i>subtilis</i>)
	15	C118	8508-8864	13,429.69	4.89	14,252.54	6.13	5	4.03	unknown
	16	A66	8848-9048	8,012.22	7.19	8,835.06	7.51	0	0.00	single C2H2 Zn-finger protein, paralog of C57 (Prangishvili, et al. 2006.)
	18	A106	9877-10197	12,568.81	5.69	13,391.66	6.97	6	5.36	unknown

Table 2-1 (cont.). General STIV ORF Information Relevant for Cloning through Structure Determination and Virus Characterization.

Internal ORF ID	ORF name	NT seq. in genome	MW (Da) ^a	Theoretical pI ^a	MW+his tag (Da) ^a	Theoretical pI +His tag ^a	# Met (minus N- term. M)	% Met His-tag, minus N-term. M)	Predicted function ^c
ORFs of nonstructural proteins^b that have been cloned (cont.).									
19	A137	10201-10614	15,873.45	8.76	16,696.30	8.76	4	2.80	NikR/CopG-like RHH transcriptional regulator (CD hit, seq. sim. to STIV A61, <i>S. tokodaii</i> STSV1, <i>S. acidocaldarius</i> , several other Archaea)
20	B97	10625-10918	11,391.35	9.77	12,214.19	9.77	0	0.00	HTH transcription regulator (Prangishvili, et al. 2006)
22	B66b	11069-11269	7,815.36	8.50	8,638.21	8.52	0	0.00	unknown
23	B66a	11315-11515	7,326.46	10.31	8,149.31	10.31	2	2.78	unknown
24	C121	11616-11981	14,286.06	4.54	15,108.91	5.11	2	1.57	unknown
25	A81	11962-12207	9,412.64	4.44	10,235.48	5.45	0	0.00	unknown
26	B124	12251-12625	15,006.32	9.53	15,829.17	9.53	1	0.77	unknown
27	A77	12622-12855	8,682.00	4.37	9,504.84	5.45	1	1.20	unknown
28	B264	12848-13642	30,117.22	4.71	30,940.07	4.99	3	1.11	unknown
29	A109	13639-13968	12,544.44	4.84	13,367.29	6.09	2	1.74	unknown
41	D66	7598-7398	7,785.20	9.77	8,608.05	9.77	0	0.00	unknown
42	F75	7833-7606	8,955.59	7.20	9,778.43	7.47	2	2.47	unknown
ORFs of structural proteins^b that have been cloned (these clones have been supplied to Jack Johnson's lab at Scripps).									
46	F93	11547-11266	11,150.26	8.63	11,973.11	8.64	2	2.02	winged helix DNA-binding protein/transcription regulator (Larson, et. al. 2006; MarR-like, BLAST/CD hits - PadR-like)
1	B109	5-334	11,847.39	4.83	12,670.24	5.86	0	0.00	particle component (Maaty, et. al. 2006)
3	B130	797-1189	13,775.98	4.89	14,598.83	6.44	0	0.00	particle component (Maaty, et. al. 2006)
4	A55	1201-1368	6,340.67	4.49	7,163.51	6.60	0	0.00	particle component (Maaty, et. al. 2006); (seq. similar to STIV C67; seq. sim. to small part of spike protein from Porcine endemic diarrhea virus and env proteins of Simian foamy virus and HIV);

Table 2-1 (cont.). General STIV ORF Information Relevant for Cloning through Structure Determination and Virus Characterization.

Internal ORF ID	ORF #	ORF name	NT seq. in genome	MW (Da) ^a	Theoretical pI ^a	MW+his tag (Da) ^a	Theoretical pI +His tag ^a	# Met (minus N- term. M)	% Met (with His-tag, minus N-term. M)	Predicted function ^c
ORFs of structural proteins ^b that have been cloned (these clones have been supplied to Jack Johnson's lab at Scripps) (cont.).										
	10	A223	5260-5931	24,423.78	4.65	25,246.63	5.37	1	0.44	possible turret component, vertex (Maaty, et. al. 2006); (potential lectin-binding domain)
	11	C381	5931-7076	41,680.15	5.56	42,503.00	6.54	2	0.52	possible turret component, vertex (Maaty, et. al. 2006)
	30	B164	13958-14452	19,036.15	9.29	19,858.99	9.29	3	1.76	possible base of turret, ATPase (Maaty, et. al. 2006); FtsK-HerA pumping ATPase (Lakshminarayan, et al. Nucleic Acids Research. 2004. 32(17): 5260-79.; very high scoring CD hit to Archaeal ATPase family), may be involved in DNA packaging (Prangishvili, et al. 2006, Maaty, et. al. 2006)
	31	B345	14942-15979	37,832.11	6.41	38,654.96	7.09	4	1.14	major capsid protein (Rice, et. al. 2004, Khayat, et. al. 2005, Maaty, et. al. 2006)
	32	C557	15990-17662	58,609.35	5.71	59,432.19	6.06	2	0.36	particle component - petal, protein interaction (Maaty, et. al. 2006); long, helical protein (Prangishvili, et al. 2006)
ORFs of nonstructural proteins ^b that have not been cloned.										
	14	A53	8344-8505	6,214.82	3.23	7,037.67	3.94	1	1.69	unknown
	17	C57	9393-9566	7,042.13	7.97	7,864.98	8.03	0	0.00	single C2H2 Zn-finger protein, paralog of A66 (Prangishvili, et al. 2006.)
	21	A58	10903-11079	6,714.88	8.98	7,537.73	8.98	1	1.56	unknown
	33	A51	17359-17514	5,880.01	9.64	6,702.86	9.64	0	0.00	unknown
ORFs of structural proteins ^b that have not been cloned.										
	X	A78	1378-1615 (gtg start)	9,609.36	10.82	10,432.21	10.82	2	2.38	particle component (Maaty, et. al. 2006)
ORFs that have been cloned but are no longer considered probable ORFs.										
	34	E93	751-470	9,320.86	5.79	10,143.71	7.03	1	1.01	

Table 2-1 (cont.). General STIV ORF Information Relevant for Cloning through Structure Determination and Virus Characterization.

Internal		ORF ID	ORF # name	NT seq. in genome	MW (Da) ^a	Theoretical		MW+his tag (Da) ^a	Theoretical pI +His tag ^a	# Met (minus N-term. M)		% Met (with His-tag, minus N-term. M)	Predicted function ^c
						pI ^a							
ORFs that have been cloned but are no longer considered probable ORFs (cont.)													
36	D150			2696-2244	16,762.67	7.59		17,585.51	7.67	1		0.64	
37	E65			3247-3050	7,685.89	7.91		8,508.74	7.97	0		0.00	
44	D72			10244-10026	8,149.26	8.18		8,972.11	8.23	0		0.00	
49	E61			14080-13895	7,200.43	10.04		8,023.27	10.04	1		1.49	
50	F65			14259-14062	7,635.80	7.19		8,458.64	7.52	2		2.82	
ORFs not cloned and no longer considered probable ORFs.													
35	D50			1040-888	5,979.12	11.46		6,801.96	11.46	0		0.00	
38	E55			3553-3386	6,726.94	9.91		7,549.78	9.91	1		1.64	
39	E52			3952-3794	6,071.18	9.45		6,894.03	9.45	0		0.00	
40	F55			6132-5965	6,332.69	9.68		7,155.53	9.68	0		0.00	
43	D56			9476-9306	6,804.16	8.93		7,627.01	8.94	0		0.00	
45	F54			10626-10462	6,649.78	9.45		7,472.62	9.45	1		1.67	
47	F59			12825-12646	7,457.56	6.29		8,280.41	7.04	0		0.00	
48	D54			13673-13509	6,091.35	7.15		6,914.19	7.64	1		1.67	

^a Calculated by DNASTar (Lasergene, Madison, WI).^b Proteins identified in Maaty, et. al. 2006. Journal of Virology. 80:7625-7635. are considered structural proteins, all others are considered nonstructural.^c Predictions based on results from BLAST (<http://www.ncbi.nlm.nih.gov/BLAST/>), CD Search (<http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi>), and the primary literature.

The Overall Scheme to Rapidly Identify Viral Proteins
Most Suitable for X-ray Structure Determination

All ORFs greater than 60 amino acids in length (from the original annotation of the STIV genome) were cloned into *E. coli* expression vectors. Small-scale expression and purification trials were performed to rapidly identify the subset of ORFs that produced detectable levels of soluble protein under “standard” conditions. Expressions of these soluble proteins were then scaled-up to optimize the purification procedure with subsequent chromatographic steps to try to increase the yield and purity. The subset of these that produced significant amounts of relatively pure protein were put through concentration experiments to see how concentrated the protein can get and remain soluble and then into crystallization experiments. Crystallization conditions were optimized for those that gave promising results and then x-ray structure determination was pursued for those that produced diffraction-quality crystals. This chapter describes the methods that were carried out in this journey and also serves to highlight what was not done. In other words, issues that may be addressed in future studies of the STIV proteome to try to obtain soluble, well-behaved protein from those ORFs that did not result in structures this time around. The first half of this chapter is devoted primarily to the methods that were used for all the ORFs greater than 60 amino acids in length and the second half describes the methods that were specific for each of the three proteins that have had their structures solved.

Cloning

Choice of Cloning System

The need for fairly high throughput and the desire to easily modify expression strategies down the road for those proteins that did not initially produce soluble proteins led us to use the Gateway® Technology cloning system (Invitrogen Corporation). The Gateway cloning system is recombination-based and allows for expression in various expression systems (i.e. bacteria, insect) and with various affinity tags, if desired, after a single PCR event (Figure 2-1). The gene of interest is amplified with the addition of specific recombination sites attB1 and attB2 to facilitate its incorporation into an entry clone, essentially a gene storage vector, where the gene sequence is verified. The gene of interest need only be sequenced after creation of the entry clone, because all subsequent steps are recombination-based so there little chance of mutations being introduced. For protein expression, the gene must be transferred into an expression clone that is specific for the desired expression system (i.e. *E. coli*) and incorporates the desired features (i.e. affinity tag) by a subsequent recombination reaction. Directionality of the ORF is ensured because there are slight differences in the recombination sequences at the 5' and 3' ends of the cloned gene.

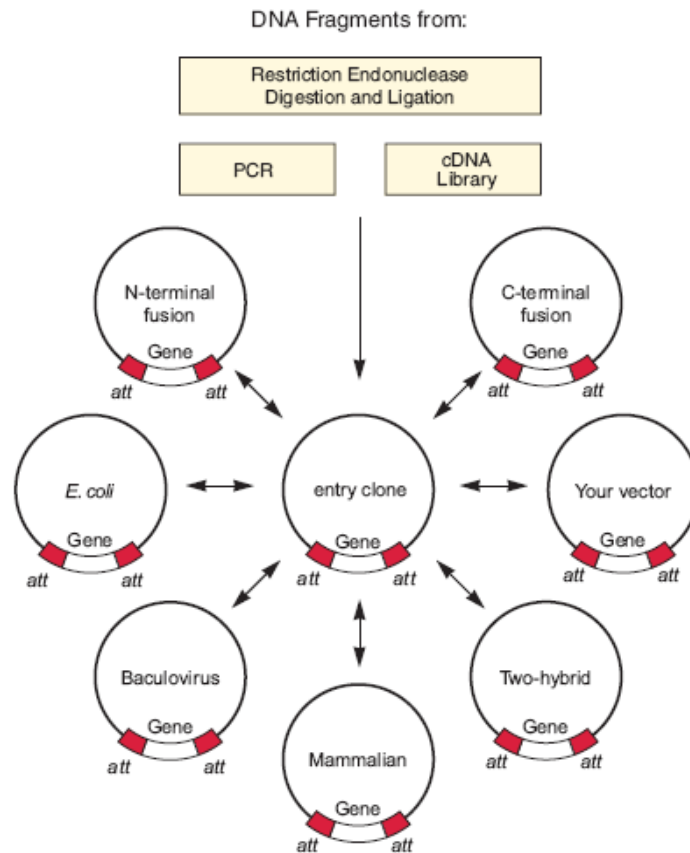


Figure 2-1. Gateway Cloning.

Gateway cloning allows for ease of moving cloned gene to alternate expression systems or to incorporate alternate features with only a single PCR event. Figure from (57).

Primer Design and PCR Amplification

PCR was used to amplify the STIV ORFs directly from purified viral particles (112) that were kindly provided by Debbie Willits from Mark Young's laboratory in the Department of Plant Sciences and Plant Pathology at Montana State University. PCR also allowed for the addition of sequences necessary for Gateway system cloning and incorporation of desired features into the final gene construct. In order to insert the gene of interest into the Gateway DONR vectors, it must be flanked by specific recombination sequences, attB sites. To increase the efficiency of expression, the ribosome signal

sequences, Shine-Dalgarno for Bacteria and Kozak for Eukaryotes, are also added. To facilitate rapid purification of the expressed proteins via immobilized metal-affinity chromatography (IMAC), we chose to add an uncleavable His tag directly to the C-terminus for simplicity. The desired PCR-amplified gene product is schematically depicted in Figure 2-2A.

- A) 
- B) 5'-GGGGACAAGTTTGTACAAAAAAGCAGGCTTCGAAGGAGATAGAACCC-3'
- C) 5'-GGGGACCACTTTGTACAAGAAAGCTGGGTCCTAGTGATGGTGATGGTGATG-3'

Figure 2-2. Desired PCR Product and Universal Primer Sequences.

(A) Schematic of desired PCR product for creation of STIV ORF expression constructs. Each specific feature is color-coded. (B) Sequence of 5' Universal Primer. (C) Sequence of 3' Universal Primer. Color-coding in B and C match the features colored in panel A.

In order to make the cloning procedure broadly applicable to many gene constructs and to keep primers lengths a reasonable size, a nested PCR scheme was used. In this scheme, there are “universal” primers that can be used with every gene and gene-specific primers that are designed to overlap with both the gene of interest and the universal primers. The universal (outside) primers (Figures 2-2B and 2-2C) add the necessary recombination sequences (attB sites) as well as the Shine-Dalgarno and Kozak sequences at the 5' end and the His tag at the 3' end. Gene-specific (inside) primers were designed to have an 18 nucleotide overlap with the universal primers at the 5' ends followed by enough gene-specific overlap to have a melting temperature (T_m) of at least 54 °C, roughly calculated using the formula $T_m \approx (2 \times (A+T)) + (4 \times (G+C))$, and to contain at

least 16 nucleotides of overlap with the gene. All gene-specific primer sequences are tabulated along with their approximate T_m s in Table 2-2. All primers were ordered from Sigma-Genosys.

The nested PCR scheme was performed essentially as described in the “Preparing attB-PCR Products Using attB Adapter PCR” section of the Gateway manual (57) with a few notable exceptions. To limit the chances of PCR-induced mutations and to decrease the time necessary for PCR, KOD HiFi polymerase (Novagen), a proofreading enzyme with very high fidelity and very high processivity, was used. This enzyme has all the desirable traits of the recommended polymerases but is much faster and significantly reduced the time needed to complete the PCR steps. Purified STIV viral particles were used for the starting template because the viral DNA could not be sufficiently purified, so the initial denaturation step was set to seven-minutes and that of each subsequent cycle was set at 30 seconds in the first, “inside”, 10-cycle PCR reaction with the gene-specific (inside) primers. Times for the elongation steps in both PCR reactions were shortened according to the recommendations for the KOD polymerase for the length of desired PCR product. 10 μ l of the “inside” reaction was used as the template for the 25 cycle “outside” reaction, which now uses the universal (outside) primers. The remainder of the inside reaction was also put through this subsequent 25 cycle PCR reaction to aid in trouble-shooting if necessary; i.e. to help isolate the problem to either the “inside” reaction or the “outside” reaction. Following the second, “outside”, PCR reaction of 25 cycles, the quality and yield of the PCR products were checked on an agarose gel. If there was only a single band of the expected size on the gel, the recommended gel

purification step was skipped, and the product used directly in the BP reaction, otherwise the band was gel purified before use in the BP reaction.

Creation of the Entry and Expression Clones

After creating PCR products containing the STIV gene flanked by attB recombination sites, the first recombination reaction, “BP reaction”, is performed to create an entry clone (Figure 2-3), essentially a gene storage vector. The donor vector pDONR201 was used and BP reactions were performed as described in the Gateway manual (57) with two exceptions. Only one-quarter of the recommended volumes were used and the reactions were incubated with proteinase K for two to four hours before the products were transformed into *E. coli* strain DH5 α competent cells (Invitrogen) and plated on LB-agar supplemented with 50 μ g/ml kanamycin. We found that the much longer incubation with proteinase K increased the transformation efficiency. Five colonies were then selected for each construct and PCR reactions were performed using the gene-specific primers to confirm that the BP reaction was successful and the entry clone contained an insert of the expected size. BP clones that tested positive for the gene insert were then sequenced to verify that the gene is mutation-free. Sequencing of all entry clones by ABI BigDye Terminator Cycle methods was performed at the DNA Facility, Iowa State University (Ames, Iowa) and the data were analyzed using DNASTar (Lasergene Inc., Madison, WI).

Table 2-2. Primer Sequences Used for Cloning the STIV Genome^a.

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	Approx. T _m of Gene Overlap (°C) ^b
ORFs that have been cloned.					
1	B109	5-334	5'SDK1 3'His1	TTCGAAGGAGATAGAACC ATGGGTTTAGGTTTCAGAATT GTGATGGTGATGGTGATG ACGCTGCTTTTCAAGCC	54 54
2	C67	588-791	5'SDK2 3'His2	TTCGAAGGAGATAGAACC ATGGCTATGGCAGGCGTA GTGATGGTGATGGTGATG CCCGTGCTTGTGCTTCTT	56 56
3	B130	797-1189	5'SDK3 3'His3	TTCGAAGGAGATAGAACC ATGGCATTAGGAGATGCAT GTGATGGTGATGGTGATG CTTCGCGTAGCTATTAG	54 54
4	A55	1201-1368	5'SDK4 3'His4	TTCGAAGGAGATAGAACC ATGAACATTGAAGAAATTCAG GTGATGGTGATGGTGATG TTTTATTATCAAACTAACAGAA	54 54
5	A510	2167-3699	5'SDK5 3'His5	TTCGAAGGAGATAGAACC ATGCCCGGCATGTAGTA GTGATGGTGATGGTGATG ACTTAGAGCGTTTAAACGCTT	56 56
6	C63	3552-3743	5'SDK6 3'His6	TTCGAAGGAGATAGAACC ATGGGAAGACCCGAACAC GTGATGGTGATGGTGATG AATAGAAAATAGAAGTATAAAAATT	56 56
7	C92	3888-4166	5'SDK7 3'His7	TTCGAAGGAGATAGAACC ATGCTAGCAGAAATAGCATT GTGATGGTGATGGTGATG CTTTGTAGCGGGCTGATT	54 54
8	A61	4432-4617	5'SDK8 3'His8	TTCGAAGGAGATAGAACC ATGACAATGAGAATAATTACAT GTGATGGTGATGGTGATG CAATCTCATTTATCTTTTCTAC	54 54
9	A197	4624-5217	5'SDK9 3'His9	TTCGAAGGAGATAGAACC ATGAGAACGCTTTTTTTTATAC GTGATGGTGATGGTGATG CTTTATTCTAGCTACGTGAT	56 54
10	A223	5260-5931	5'SDK10 3'His10	TTCGAAGGAGATAGAACC ATGGGAGAAGTTTTTAAAGAA GTGATGGTGATGGTGATG TATCACCTCTTCATATATTAAAT	54 54
11	C381	5931-7076	5'SDK11 3'His11	TTCGAAGGAGATAGAACC ATGAGTGTTACTACATTGGG GTGATGGTGATGGTGATG GCTTAGTCTCCCCGTATT	56 56
12	B106	7646-7966	5'SDK12 3'His12	TTCGAAGGAGATAGAACC ATGTTACACGGTAAAAATTGTA GTGATGGTGATGGTGATG GACTCTATTTAAATGTTTTTGC	56 56
13	B116	7994-8344	5'SDK13 3'His13	TTCGAAGGAGATAGAACC ATGGGTAAGGTATTCCTCA GTGATGGTGATGGTGATG CCACACCTCATAAAATGAG	54 54
15	C118	8508-8864	5'SDK15 3'His15	TTCGAAGGAGATAGAACC ATGAGTGAAGCTGAAGTTC GTGATGGTGATGGTGATG ATCTTTTTTCTCATTGTTCTCA	54 54
16	A66	8848-9048	5'SDK16 3'His16	TTCGAAGGAGATAGAACC ATGAGAAAAAAGATTAATTTGTAT GTGATGGTGATGGTGATG GGCAAGTCCTAACGCCTC	56 58
18	A106	9877-10197	5'SDK18 3'His18	TTCGAAGGAGATAGAACC ATGACAGAAAGCGATAAA GTGATGGTGATGGTGATG CCCCTTCTTCAACAATGC	54 54
19	A137	10201-10614	5'SDK19 3'His19	TTCGAAGGAGATAGAACC ATGGAAAGAGTAATACTTTTAT GTGATGGTGATGGTGATG CCCAATCATTATCCCCTC	54 54
20	B97	10625-10918	5'SDK20 3'His20	TTCGAAGGAGATAGAACC ATGCCAGCAGAAAGTGATAG GTGATGGTGATGGTGATG CTTCACCCCATGCCAA	56 58
22	B66b	11069-11269	5'SDK22 3'His22	TTCGAAGGAGATAGAACC ATGAGAATTGATATAGAAGGA GTGATGGTGATGGTGATG ACGAAGTTTGCAAAAAAGAC	54 54
23	B66a	11315-11515	5'SDK23 3'His23	TTCGAAGGAGATAGAACC ATGCCCTTCTCTGTAAGAA GTGATGGTGATGGTGATG CATTATTCTGAAAGTATTGGT	54 54
24	C121	11616-11981	5'SDK24 3'His24	TTCGAAGGAGATAGAACC ATGAACGATAATACTTATATAC GTGATGGTGATGGTGATG ATCGTTATCTTGCTCATATAA	54 54
25	A81	11962-12207	5'SDK25 3'His25	TTCGAAGGAGATAGAACC ATGAGCAAGATAACGATTAAT GTGATGGTGATGGTGATG CTTATTTTCAGAATCAATAAAG	54 54
26	B124	12251-12625	5'SDK26 3'His26	TTCGAAGGAGATAGAACC ATGAGTCAAGCTTCTGAAAA GTGATGGTGATGGTGATG TTTCTCCACCTTTTCCAAA	54 54
27	A77	12622-12855	5'SDK27 3'His27	TTCGAAGGAGATAGAACC ATGAGGAAAGAGGCAATTG GTGATGGTGATGGTGATG CTCATTTTCTTCCACCTTTG	54 56
28	B264	12848-13642	5'SDK28 3'His28	TTCGAAGGAGATAGAACC ATGAGTGAGATATGACCC GTGATGGTGATGGTGATG TATCTCCATTTCTCCATTA	56 54
29	A109	13639-13968	5'SDK29 3'His29	TTCGAAGGAGATAGAACC ATGAAAGTTCCTAGCAGCA GTGATGGTGATGGTGATG GGATTCATTTTCTATCCCC	54 54
30	B164	13958-14452	5'SDK30 3'His30	TTCGAAGGAGATAGAACC ATGAATCCTGACGATATTG GTGATGGTGATGGTGATG GTTCATTAGGCTCTCTGG	54 54
31	B345	14942-15979	5'SDK31 3'His31	TTCGAAGGAGATAGAACC ATGGGAGAAATCTATACAGA GTGATGGTGATGGTGATG TCTCTTTATCCTTCTTTTTTG	54 54
32	C557	15990-17662	5'SDK32 3'His32	TTCGAAGGAGATAGAACC ATGTCTTTTTTTGGAGGATTG GTGATGGTGATGGTGATG CCTCATGACTATTATTCCTAT	56 56

Table 2-2 (cont.). Primer Sequences Used for Cloning the STIV Genome^a.

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	Approx. T _m of Gene Overlap (°C) ^b
ORFs that have been cloned (cont.).					
41	D66	7598-7398	5'SDK41 3'His41	TTCGAAGGAGATAGAACC ATGACGTGCTTTTACGTGA GTGATGGTGATGGTGATG GCCAGCCGAAACCAGC	54 54
42	F75	7833-7606	5'SDK42 3'His42	TTCGAAGGAGATAGAACC ATGCATTTCGCTTTTGGATTGA GTGATGGTGATGGTGATG CAAATCGATTTTGGCCAAG	54 54
46	F93	11547-11266	5'SDK46 3'His46	TTCGAAGGAGATAGAACC ATGAAAATCAGAAAGTATATGA GTGATGGTGATGGTGATG GCATACTATTTTTTGAAAAAATT	54 54
ORFs that have not been cloned.					
X	A78	1378-1615 (gtg start)		**New ORF identified by proteomic studies in Brian Bothner's lab.**	
14	A53	8344-8505			
17	C57	9393-9566			
21	A58	10903-11079			
33	A51	17359-17514			
ORFs that were originally annotated and cloned but are no longer considered probable ORFs.					
34	E93	751-470	5'SDK34 3'His34	TTCGAAGGAGATAGAACC ATGACTAAAGCTATGATCACT GTGATGGTGATGGTGATG CAACCTTCCTCCAACCTAAC	56 56
36	D150	2696-2244	5'SDK36 3'His36	TTCGAAGGAGATAGAACC ATGACCGTTTAAACCCATC GTGATGGTGATGGTGATG GTTAAATAGACGGGTAGAA	54 54
37	E65	3247-3050	5'SDK37 3'His37	TTCGAAGGAGATAGAACC ATGTCTTTCCCGTCTTTGA GTGATGGTGATGGTGATG AAGGAGGAAAAAAGAGAGA	54 54
44	D72	10244-10026	5'SDK44 3'His44	TTCGAAGGAGATAGAACC ATGTTAGACTCTGAGTTAGA GTGATGGTGATGGTGATG AAAAAATAGACGAAATGAAAAAAT	54 54
49	E61	14080-13895	5'SDK49 3'His49	TTCGAAGGAGATAGAACC ATGGTCATCTATTATGTAAGA GTGATGGTGATGGTGATG GAAGAGAGTTAATAGAAATGT	54 54
50	F65	14259-14062	5'SDK50 3'His50	TTCGAAGGAGATAGAACC ATGAGGACACTAGTTCCC GTGATGGTGATGGTGATG CATAATAGATGACCATAATCT	54 54
ORFs that were originally annotated but not cloned and are no longer considered probable ORFs.					
35	D50	1040-888			
38	E55	3553-3386			
39	E52	3952-3794			
40	F55	6132-5965			
43	D56	9476-9306			
45	F54	10626-10462			
47	F59	12825-12646			
48	D54	13673-13509			
Universal Primers - used to add attB sites, Shine-Dalgarno sequence, Kozak sequence, and C-terminal His tag					
	5' Universal		PK1-3	GGGGACAAGTTTGTACAAAAAAGCAGGCTTCGAAGGAGATAGAACC	
	3' Universal		PK1-4	GGGGACCACTTTGTACAAGAAAGCTGGGTCTAGTGATGGTGATGGTGATG	

^a all primers were ordered from Sigma Genosys.^b approximate melting temperature calculated by using the formula $T_m = 2 \times (A+T) + 4 \times (G+C)$.

Figure 2-3. The BP Reaction.

A PCR product that contains the gene of interest flanked by specific recombination sites (attB) sites is created. This PCR product is combined with a donor vector (pDONR201) and the site-specific recombinase (BP Clonase) to yield the entry clone, essentially a gene storage vector. Figure from (57).

After an entry clone is created and the gene sequence verified, it must undergo a subsequent recombination reaction, the “LR reaction”, into a destination vector that will allow for protein expression (Figure 2-4), the expression clone. Several destination vectors are available that allow for expression in various organisms and incorporate various features. We chose to use pDEST14 for expression in *E. coli*. This destination vector does not add any additional features, such as fusion tags, and is intended for native protein expression. We did add C-terminal His tags to all the genes in the PCR step, however, to facilitate purification via IMAC. LR reactions were performed as described (57) except, again, quarter volumes were used for the recombination reaction and the proteinase K incubation was lengthened. Reaction products were transformed into *E. coli* strain DH5 α competent cells (Invitrogen) and plated on LB-agar supplemented with 100 μ g/ml ampicillin. Colonies were selected and the success of the LR reaction was verified by PCR using the gene-specific primers. PCR verification of the expression clones is sufficient because they are produced by recombination from an entry clone containing a verified gene sequence.



Figure 2-4. The LR Reaction.

The entry clone, which contains the verified gene sequence, is combined with a destination vector (pDEST14 in our case) and the site-specific recombinase (LR Clonase) to yield the expression clone, a vector suitable for protein expression. Figure from (57).

A) BP reaction:

attB-flanked STIV ORF #1 PCR product + pDONR201 → pENTR201-1
 5 colonies chosen from transformations and colony 2 had the correct
 gene sequence so the resultant entry clone is pENTR201-1.2

B) LR reaction:

pENTR201-1.2 + pDEST14 → pEXP14-1.2

Figure 2-5. STIV Vector Naming Conventions.

(A) BP reactions yield entry clones with vector names containing ENTR plus the number of the donor vector used followed by the gene identifier after a dash and the number of the verified sequence after the period. **(B)** LR reactions yield expression clones with vector names containing EXP plus the number of the destination vector used followed by the gene identifier from the entry clone after a dash.

Vector naming conventions were essentially as established in the Gateway manual for all vectors created. An example for STIV ORF B109 (#1 in our ORF numbering scheme) is shown in Figure 2-5. The names of all entry clones and expression clones produced as well as other statistics relevant to the cloning of the STIV ORFs are presented in Table 2-3.

Expression and Purification

Statistics relevant to the expression and purification of the STIV proteome are presented in Table 2-3. Specific information about the expression and purification of the three proteins whose structures were determined are presented at the end of this chapter after the general methods.

Table 2-3. The Progress Towards the Complete Structural Characterization of STIV.

Internal ORF ID	ORF name	Cloned and Sequenced ^b ?	Vector names (entry, expression)	Soluble Expression?	Crystallization trials?	Crystals?	Diffraction?	Structure?	Polyclonal Antibodies ^c ?	What next?
ORFs of nonstructural proteins ^d that have been cloned.										
2	C67	yes	pENTR201-2.4, pEXP14-2.4	no	no	no	no	no	no	determine suitable expression/purification conditions
5	A510	yes	pENTR201-5.1, pEXP14-5.1, pENTR201-5.1y(east), pEXP52-5.1y(east)	in yeast and low pH elution	Hampton 1 & 2, Cryo 1 & 2, HTS	no	no	no	yes	retry bacterial purification at low pH, optimize purification, play with starting crystallization conditions (protein conc., buffer, pH, temp, etc.)
6	C63	yes	pENTR201-6.4, pEXP14-6.4	no	no	no	no	no	no	determine suitable expression/purification conditions
7	C92	yes	pENTR201-7.1, pEXP14-7.1	no	no	no	no	no	no	determine suitable expression/purification conditions
8	A61	yes	pENTR201-8.4, pEXP14-8.4	no	no	no	no	no	no	determine suitable expression/purification conditions
9	A197	yes	pENTR201-9.2, pEXP14-9.2	yes, dimer	Hampton 1 & 2, Cryo 1 & 2, HTS	yes	yes	2CON	yes	functional studies
12	B106	yes	pENTR201-12.2, pEXP14-12.2	no	no	no	no	no	no	determine suitable expression/purification conditions
13	B116	yes	pENTR201-13.1, pEXP14-13.1	yes, dimer	Hampton 1 & 2, Cryo 1 & 2, HTS	yes	yes	2BLK, 2J85	yes	functional studies
15	C118	yes	pENTR201-15.3, pEXP14-15.3	no	no	no	no	no	no	determine suitable expression/purification conditions

Table 2-3 (cont.). The Progress Towards the Complete Structural Characterization of STIV.

Internal ORF ID #	ORF name	Cloned and Sequenced ^b ?	Vector names (entry, expression)	Soluble Expression?	Crystallization trials?	Crystals?	Diffraction?	Structure?	Polyclonal Antibodies ^c ?	What next?
ORFs of nonstructural proteins ^d that have been cloned (cont.).										
16	A66	yes	pENTR201-16.1, pEXP14-16.1	no	no	no	no	no	no	determine suitable expression/purification conditions
18	A106	yes	pENTR201-18.1, pEXP14-18.1	no	no	no	no	no	no	determine suitable expression/purification conditions
19	A137	yes	pENTR201-19.1, pEXP14-19.1	no	no	no	no	no	no	determine suitable expression/purification conditions
20	B97	yes	pENTR201-20.2, pEXP14-20.2	no	no	no	no	no	no	determine suitable expression/purification conditions
22	B66b	yes	pENTR201-22.2, pEXP14-22.2	no	no	no	no	no	no	determine suitable expression/purification conditions
23	B66a	yes	pENTR201-23.2, pEXP14-23.2	no	no	no	no	no	no	determine suitable expression/purification conditions
24	C121	yes	pENTR201-24.3, pEXP14-24.3	no	no	no	no	no	no	determine suitable expression/purification conditions
25	A81	yes	pENTR201-25.3, pEXP14-25.3	yes, tetramer	Hampton 1 & 2, Cryo 1 & 2, HTS, in-house robot	yes	yes	no	no	solve the structure, make polyclonal antibodies
26	B124	yes	pENTR201-26.1, pEXP14-26.1	no	no	no	no	no	no	determine suitable expression/purification conditions

Table 2-3 (cont.). The Progress Towards the Complete Structural Characterization of STIV.

Internal ORF ID	ORF name	Cloned and Sequenced ^b ?	Vector names (entry, expression)	Soluble Expression?	Crystallization trials?	Crystals?	Diffraction?	Structure?	Polyclonal Antibodies ^c ?	What next?
ORFs of nonstructural proteins^d that have been cloned (cont.).										
27	A77	yes, 8Xhis	pENTR201-27.1, pEXP14-27.1	yes, only in shake flasks, dirty	no	no	no	no	no	optimize expression/purification conditions, consider different constructs
28	B264	yes	pENTR201-28.1, pEXP14-28.1, pENTR201-28.2y(east), pEXP52-28.2y(east)	yes, dimer?, 2 major bands; pure protein from alternate start codon construct	Hampton 1 & 2, Cryo 1 & 2 of double band material	no	no	no	no	use the alternate start codon construct and optimize purification conditions and protein conditions to use in crystallization experiments (protein conc., buffer, pH, etc.)
29	A109	yes	pENTR201-29.2, pEXP14-29.2	yes, dimer?	HTS	no	no	no	no	play with starting crystallization conditions (protein conc., buffer, pH, etc.)
41	D66	yes	pENTR201-41.3, pEXP14-41.3	no	no	no	no	no	no	determine suitable expression/purification conditions
42	F75	yes	pENTR201-42.1, pEXP14-42.1	no	no	no	no	no	no	determine suitable expression/purification conditions
46	F93	yes	pENTR201-46.2, pEXP14-46.2	yes, dimer	Hampton 1 & 2, Cryo 1 & 2, HTS	yes	yes	2CO5	yes	functional studies
ORFs of structural proteins^d that have been cloned (these clones have been supplied to Jack Johnson's lab at Scripps).										
1	B109	yes	pENTR201-1.2, pEXP14-1.2	yes, 2 major bands	Hampton 1 & 2, Cryo 1 & 2 of double band material	no	no	no	no	determine suitable expression/purification conditions, consider different constructs
3	B130	yes	pENTR201-3.1, pEXP14-3.1	no	no	no	no	no	no	determine suitable expression/purification conditions

Table 2-3 (cont.). The Progress Towards the Complete Structural Characterization of STIV.

Internal ORF ID	ORF name	Cloned and Sequenced ^b ?	Vector names (entry, expression)	Soluble Expression?	Crystallization trials?	Crystals?	Diffraction?	Structure?	Polyclonal Antibodies ^c ?	What next?
ORFs of structural proteins ^d that have been cloned (these clones have been supplied to Jack Johnson's lab at Scripps) (cont.).										
4	A55	yes	pENTR201-4.5, pEXP14-4.5	no	no	no	no	no	no	determine suitable expression/purification conditions
10	A223	yes	pENTR201-10.2, pEXP14-10.2	yes, band may be bit too small	HTS, in-house robot	yes	yes (~4 Å)	no	yes	optimize crystallization conditions, try Se-Met, identify heavy atom derivatives, add more Met by site-directed mutagenesis
11	C381	yes	pENTR201-11.1, pEXP14-11.1	yes, band may be bit too small, dirty	in-house robot	no	no	no	yes	MS/sequence band to see if it is a fragment or just runs funny on gel, determine suitable expression/purification conditions, consider alternate constructs
30	B164	yes	pENTR201-30.2, pEXP14-30.2	no	no	no	no	no	no	determine suitable expression/purification conditions
31	B345	yes	pENTR201-31.3, pEXP14-31.3	yes	yes	yes	yes	2BBD	yes	
32	C557	not fully sequenced, couple hundred NTs from 5' and 3' OK and insert is expected size.	pENTR201-32.3, pEXP14-32.3, pENTR201-32.4y(east), pEXP52-32.4y(east)	no	no	no	no	no	no	determine suitable expression/purification conditions
ORFs of nonstructural proteins ^d that have not been cloned.										
14	A53	no								clone
17	C57	no								clone

Table 2-3 (cont.). The Progress Towards the Complete Structural Characterization of STIV.

Internal ORF ID	ORF name	Cloned and Sequenced ^b ?	Vector names (entry, expression)	Soluble Expression?	Crystallization trials?	Crystals?	Diffraction?	Structure?	Antibodies ^c ?	Polyclonal Antibodies ^c ?	What next?
ORFs of nonstructural proteins^d that have not been cloned (cont.).											
21	A58	no									clone
33	A51	no									clone
ORFs of structural proteins^d that have not been cloned.											
X	A78	no									clone
ORFs that have been cloned but are no longer considered probable ORFs.											
34	E93	yes	pENTR201-34.1, pEXP14-34.1	no							
36	D150	yes	pENTR201-36.3, pEXP14-36.3	no							
37	E65	yes	pENTR201-37.1, pEXP14-37.1	no							
44	D72	yes	pENTR201-44.2, pEXP14-44.2	no							
49	E61	yes	pENTR201-49.4, pEXP14-49.4	no							
50	F65	yes	pENTR201-50.3, pEXP14-50.3	no							
ORFs that have not been cloned and no longer considered probable ORFs.											
35	D50										
38	E55										
39	E52										
40	F55										
43	D56										
45	F54										
47	F59										
48	D54										

^a Calculated by DNASTar (Lasergene, Madison, WI).^b Sequencing was done at the DNA Facility, Iowa State University (Ames, IA).^c Rabbit polyclonal antibodies were produced by Ingenex Corp. (San Diego, CA).^d Proteins identified in Maaty, *et al.* 2006. Journal of Virology. 80:7625-7635. are considered structural proteins, all others are considered nonstructural.

Expression in *E. coli*

Bacterial expression is rather simple and so *E. coli* was chosen as the first expression system. Since we were dealing with a large number of ORFs and knew that only a limited number of them would give us soluble proteins with minimal effort, we sought to identify this subset as quickly as possible using established protocols. Those that did not produce soluble proteins under these conditions would require further troubleshooting by altering the expression conditions, or even the expression system or expression construct. Only minimal efforts were made to address these issues because several proteins were found to behave nicely under these initial conditions. Common tricks employed to increase protein expression, such as lower temperature cell growth, growth in richer medium, or co-expression with the chaperones GroEL/GroES, were not done. Furthermore, we did not look at the insoluble fractions for protein expression. As the majority of the proteins in the STIV proteome are fairly small, purification from the insoluble fraction followed by refolding should be investigated, however. It is common for the expression construct itself to cause problems with protein expression. The construct may be redesigned to move, change, or remove the affinity tag or to allow expression in a different system such as yeast or insect cells. Furthermore, proteins that have transmembrane helices are not expected to give soluble protein. In addition, some of the ORFs may be mispredicted. Modification of the gene-construct to remove potential transmembrane regions or to utilize downstream start codons may be warranted.

Expression clones were transformed into *E. coli* strain BL21-CodonPlus(DE3)-RIL (Stratagene) and a single colony was used to inoculate 5 ml of LB medium containing 100 µg/ml ampicillin and grown overnight at 37 °C with vigorous shaking.

The overnight culture was used to inoculate a larger culture of the same medium and it was incubated under the same conditions. Cells were grown to an optical density at 600 nm (OD_{600}) of 0.6 to 0.8 and protein expression was then induced by addition of 1 mM isopropyl- β -D-thiogalactopyranoside (IPTG). After an additional 4 to 6 hours of growth, cells were harvested and pellets were stored at -20 to -80 °C until needed for purification. For the initial small-scale expression trials, 100 ml cultures in shaker flasks were used. Those ORFs that appeared to produce soluble protein were then scaled up to 1 L cultures in shaker flasks to further optimize purification conditions and those that then produced well-behaving protein were scaled up to 6×1 L shaker flask cultures and/or eventually to 10-L fermentor batch cultures for large-scale protein production. All shaker flask cultures were treated essentially the same, but, necessarily, a slightly different procedure was followed for fermentor batch cultures and a different medium is used.

For fermentor batch cultures, a single colony from a transformation was used to inoculate 5 ml of LB medium and grown overnight, with subsequent serial expansion to a 10-liter fermentor (New Brunswick; BIOFLO 2000) batch culture, all at 37 °C. Medium for batch fermentation was as recommended by the manufacturer (109) [3.5 g/l KH_2PO_4 , 5.0 g/l K_2HPO_4 , 3.5 g/l $(NH_4)_2HPO_4$, 0.5g/l $MgSO_4 \cdot 7H_2O$, 30 g/l glucose, 5.0 g/l yeast extract, 1 ml/l trace metals solution (1.6 g/l $FeCl_3 \cdot 6H_2O$, 0.2g/l $CoCl_2$, 0.1g/l $CuCl_2$, 0.2 g/l $ZnCl_2 \cdot 4H_2O$, 0.2 g/l $Na_2MoO_4 \cdot 2H_2O$, 0.05g/l H_3BO_3 , 10 ml/l HCl)] and all media contained 100 μ g/ml ampicillin. Fermentor batch cultures were supplemented with filtered air at a flow rate of 10 liters/minute and stirred at 300 to 500 rpm. The pH and dissolved oxygen levels are not monitored, however. Cells were grown to an OD_{600} of 0.8 to 1.5, and protein expression was then induced by addition of 1 mM IPTG. After an

additional 4 to 6 hours of growth, cells were harvested by centrifugation (IEC PR-7000 M) at 6,000- $\times$ g for 20 minutes and pellets were stored at -20 to -80 °C until purification.

Purification

All STIV ORFs were cloned with a His tag to facilitate purification of the protein products via IMAC. During the initial small-scale purifications, IMAC was the only chromatographic step. Scale-up, however, involved the optimization of the purification scheme with the addition of further chromatographic steps to increase the purity of the proteins. For the most part, purification of all proteins used the same set of “standard” purification protocols and only those that produced well-behaved soluble protein were further optimized. As the goal is to determine as many crystal structures as quickly as possible, minimal effort was placed on trouble-shooting purification conditions for those proteins that did not appear to produce soluble proteins in the initial small-scale experiments. For instance, expression in inclusion bodies and purification from the insoluble fraction followed by refolding was not attempted.

The “standard” IMAC purification scheme for a large-scale prep follows. The scheme was essentially the same as the initial small-scale purifications but the volumes were adjusted accordingly. Cell pellets were thawed and resuspended in lysis buffer (10 mM Tris-HCl, pH 8.0, 400 mM NaCl) supplemented with the protease inhibitor, phenylmethylsulfonyl fluoride (PMSF; 0.1 mM), at approximately 10 ml/g of wet cell pellet mass. Cells were lysed by passage through a microfluidizer (Microfluidics Corporation, Newton, MA) or a power laboratory press (American Instrument Co., Inc., Silver Spring, MD), depending on the volume of the cell suspension. Since the expressed

proteins came from a thermophilic organism, we assumed they were heat stable, so the cell lysate was incubated at 65 °C for 20 minutes to denature many of the contaminating *E. coli* proteins. (Though all proteins were put through this heat denaturation step, this step maybe should be skipped for the proteins that did not yield soluble proteins. Not all thermophilic proteins are excessively thermostable and they may actually be regulated by having a short half-life at elevated temperatures.) The heat-denatured cell lysate was then clarified by centrifugation (Beckman; Avanti J-30I) at 22,000- $\times$ g for 20 minutes and all subsequent steps were carried out on ice. The resulting cleared lysate was applied to a column containing a 3- to 5-ml bed volume of HIS-Select nickel affinity gel (Sigma-Aldrich), washed with at least 10 column volumes of wash buffer (lysis buffer plus 5 mM imidazole), and eluted with elution buffer (10 mM Tris-HCl, pH 8.0, 50 mM NaCl, 200 mM imidazole). All steps of the purification from the cleared lysate to the elution were monitored with SDS-PAGE using gels of an appropriate percentage of acrylamide for the size of the protein. For proteins under 12 kDa in size, we routinely used a Tris-tricine buffer system for better resolution.

Efforts were made to optimize the purification conditions for the viral proteins that showed promise of producing soluble protein under the standard expression and purification conditions. Further chromatographic steps using an FPLC (Amersham-Pharmacia) were added to try to increase their purity. Typical chromatographic techniques that were tried included size exclusion with various pore sizes, ion exchange with both cation and anion, and hydrophobic interaction. Though many of these were used, typically very little was done to optimize them on a protein-by-protein basis as those that worked with minimal effort were carried on and those that did not were left

behind. For instance, buffer conditions such as pH, salt concentrations, and additives (i.e. other salts, detergents, reducing agents, etc.) were not investigated in detail. Gel filtration, or size exclusion chromatography, was always the final step because it is desirable to have protein in a minimal buffer, i.e. 10 mM buffer with little (50 mM) or no salt, prior to crystallization experiments so the experimental buffer conditions dominate the equilibrium system. SDS-PAGE was used to visualize peak fractions and assess molecular weight and purity. Protein concentrations were determined by Bradford assay (20) using “protein assay reagent” (Bio-Rad) and bovine serum albumin (BSA) as a standard. Occasionally, proteins were further characterized with isoelectric focusing gels, MALDI-TOF mass spectrometry, or dynamic light scattering.

Following buffer exchange by gel filtration chromatography, peak fractions were pooled and concentrated using a spin concentrator with a molecular mass cutoff appropriate for the size of the protein (Amicon ultracentrifugal filter devices; Millipore). The initial target protein concentration for use in crystallization experiments was at least 10 mg/ml. If protein precipitation was observed before that value was reached, then the protein was concentrated as high as possible for initial crystallization experiments. In addition, if protein concentration was difficult due to solubility limitations, solubility matrix screens that varied pH in one dimension and salt concentration in the other were performed to try to identify buffer conditions that maximized its solubility. These solubility screens were set up using hanging drop vapor diffusion just like described below for crystallization screens. If buffer conditions were identified that increased the solubility of the protein, adjustments were made upstream in the purification process.

Crystallization

Purified, concentrated proteins were then set up in hanging drop vapor diffusion crystallization experiments (Figure 2-6). Initial crystallization experiments cast a wide net to explore many different conditions. Those that are identified to yield crystals, or “crystal-like” objects, are subsequently optimized by exploring the specific conditions in finer detail around pH and precipitant concentration. Several commercially available screens from Hampton and Emerald Biosciences were used that employ a variety of strategies that may identify initial crystallization conditions that can be refined in a systematic manner to produce diffraction quality crystals. We also routinely use a more systematic approach to explore the crystallization space in order to identify the precipitation edges as a function of precipitant concentration and pH. Once the precipitation edge is determined, it is explored in finer detail to produce diffraction quality crystals.

To optimize crystallization conditions, fine grid screens were set up around the identified condition with small changes in the pH in one dimension of the screen and a slight range of precipitant concentrations in the other to hopefully produce crystals suitable for diffraction studies. The effects of changing the concentrations of other components within the buffer and of various additives on crystallization are also investigated during the optimization. Additives include cations and anions of different valence, volatile and nonvolatile organics, detergents, polyamines, chaotropes, reducing agents, chelators, and potential ligands such as nucleotides and nucleotide analogs. It was also common to explore crystallization optimization at various temperatures [4 °C,

17 °C, 22 °C, and room temperature (a nonconstant temperature so slight fluctuations will occur)] and, to a lesser extent, varying protein concentrations in the initial drop.

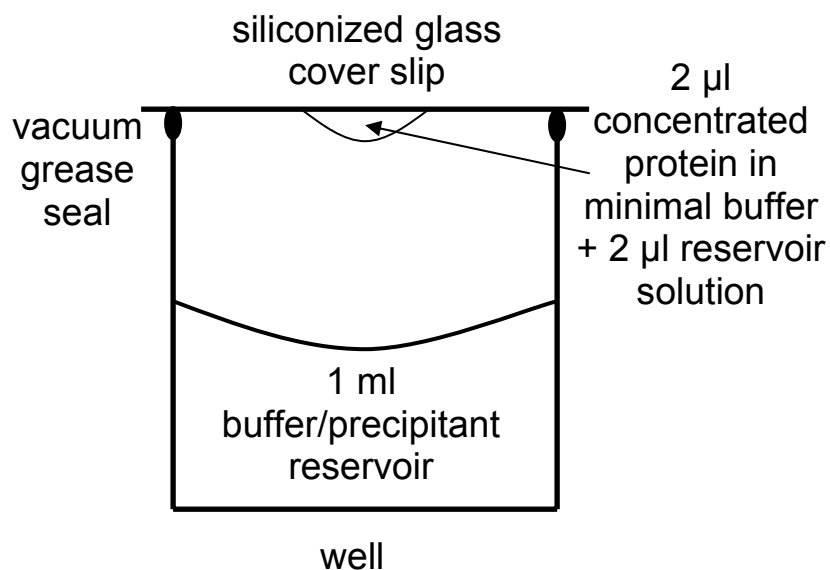


Figure 2-6. Set-up for Hanging Drop Vapor Diffusion Crystallization.

Trays contain 24 wells (a 6 x 4 grid). Each well is filled with 1 ml of a defined solution. 2 μ l of the concentrated protein (in as minimal a buffer as possible) is mixed with 2 μ l of the well solution on a siliconized glass cover slip. This cover slip is then inverted over the well and sealed with vacuum grease. The levels of buffer components in the drop slowly equilibrate with those in the well. As vapor leaves the drop, its volume decreases, and the protein concentration increases. If conditions are appropriate, protein crystals grow in the drop.

Protein samples were also submitted to the high throughput micro-batch crystallization screen offered by the Hauptman-Woodward Institute in Buffalo, New York (77). Using this service, it is possible to screen over 1500 conditions, many of which are the same commercially available conditions we were already using, with much less protein sample because the experiments use much smaller volumes. If conditions from this screen were identified, attempts were made to optimize them as described

above. This service utilizes the microbatch method of crystallization, however, so conditions may not always be directly transferable to vapor diffusion so it may be necessary to use the microbatch technique while optimizing conditions identified in this screen.

Our lab has recently purchased a Honeybee crystallization robot. This robot uses 96-well format trays and 0.4 μ l of protein per drop so it is possible to screen all the conditions with much less sample. It uses the same commercially available screens and we have created some custom screens. These robot screens use the sitting drop method of vapor diffusion so conditions are usually easily transferable back to the 24 well tray format for optimization in hanging drops. This is not always the case, however and sitting drops may have to be used. It has also been observed that the transfer from the small volumes to larger volumes sometimes requires fairly significant alterations in the crystallization condition in order to reproduce the crystals, as the equilibration in the large wells is not always scalable from the small volumes. For this reason, it is best to do a fairly coarse screen in the first round of optimization.

Specific crystallization conditions for each protein that yielded a structure are presented in their respective sections at the end of this chapter and the progress towards the crystallization of the rest of the STIV proteome is tabulated Table 2-3.

Materials and Methods for the Structure Solution of A197

Cloning the A197 Gene

The A197 open reading frame (ORF) was amplified by nested PCR directly from viral particles purified as previously described (112). The PCR primers added attB sites

to facilitate ligase-free cloning using the Gateway system (Invitrogen), a Shine-Dalgarno sequence to facilitate efficient translation, and a C-terminal His-tag to facilitate protein purification via nickel-affinity chromatography. The internal forward and reverse primers were 5'-TTCGAAGGAGATAGAACCATGAGAACGCTTTTTTTTATAC-3' and 5'-G TGATGGTGATGGTGATGCTTTATTCTAGCTACGTGAT-3', respectively, while the external forward and reverse primers were 5'-GGGGACAAGTTTGTACAAA AAAGCAGGCTTCGAAGGAGATAGAACC-3' and 5'-GGGGACCACTTTGTACAA GAAAGCTGGGTCCTAGTGATGGTGATGGTGATG-3', respectively. The resulting entry clone was sequenced and confirmed to be mutation-free using ABI BigDye terminator cycle sequencing. The His-tagged A197 construct was then inserted into destination vector pDEST14 (Invitrogen), yielding the expression vector pEXP14-A197, for protein expression in *Escherichia coli*.

Expression, Purification, and Characterization of A197

Typically, pEXP14-A197 was transformed into *E. coli* strain BL21-CodonPlus (DE3)-RIL (Stratagene), and a single colony was used to inoculate 5 ml of LB medium and grown overnight, with subsequent serial expansion to a 10-liter fermentor (New Brunswick; BIOFLO 2000) batch culture, all at 37 °C. Medium for batch fermentation was as recommended by the manufacturer (109)(above). All media contained 100 µg/ml ampicillin. Fermentor batch cultures were supplemented with filtered air at a flow rate of 10 liters/minute and stirred at 300 to 500 rpm. Cells were grown to an optical density at 600 nm (OD₆₀₀) of 0.8 to 1.5, and protein expression was then induced by addition of 1 mM isopropyl-β-D-thiogalactopyranoside (IPTG). After an additional 4 to 6 hours of

growth, cells were harvested by centrifugation (IEC PR-7000 M) at 6,000- $\times$ g for 20 minutes and pellets were stored at -80 °C until needed.

For purification, cell pellets were thawed and resuspended in lysis buffer (10 mM Tris-HCl, pH 8.0, 400 mM NaCl) at 10 ml/g of wet cell pellet mass.

Phenylmethylsulfonyl fluoride (PMSF; 0.1 mM) was added to the cell suspension, and cells were lysed by passage through a microfluidizer (Microfluidics Corporation, Newton, MA) or a Power laboratory press (American Instrument Co., Inc., Silver Spring, MD), depending on the volume of the cell suspension. The cell lysate was incubated at 65 °C for 20 minutes to denature many of the contaminating *E. coli* proteins and clarified by centrifugation (Beckman; Avanti J-30I) at 22,000- $\times$ g for 20 minutes. The resulting cleared lysate was applied to a column containing a 3- to 5-ml bed volume of HIS-Select nickel affinity gel (Sigma-Aldrich), washed with at least 10 column volumes of wash buffer (lysis buffer plus 5 mM imidazole), and eluted with elution buffer (10 mM Tris-HCl, pH 8.0, 50 mM NaCl, 200 mM imidazole). Fractions containing A197 were then pooled and loaded onto a Superdex 75 gel filtration column (Amersham Biosciences) for further purification and buffer exchange into a minimal buffer (10 mM Tris-HCl, pH 8.0, 50 mM NaCl) for crystallization. Peak fractions were pooled and concentrated using a 10-kDa molecular mass cutoff (Amicon ultracentrifugal filter devices; Millipore) to 10.5 mg/ml. Protein concentrations were determined by Bradford assay (20) using “protein assay reagent” (Bio-Rad) and bovine serum albumin (BSA) as a standard. Protein yield was typically 2 to 5 mg/g of cell pellet. Molecular weight and purity were assessed by 12% sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE). Dynamic light scattering (DLS) measurements were made on a Brookhaven Instruments ZetaPals

(phase analysis light scattering) particle size/ ζ potential analyzer. DLS was measured at 90° using a 661-nm diode laser, and the correlation functions were fitted using a constrained least-squares analysis as described previously (3).

For expression and purification of selenomethionine-incorporated A197, pEXP14-A197 was transformed into *E. coli* strain B834(DE3)pLysS (Novagen), a methionine auxotroph. Methionine auxotrophy was confirmed for a single colony, which was then used to inoculate 5 ml of medium, prepared essentially as described previously (130) [Vogel-Bonner minimal medium, supplemented per L with 1 mg thiamine, 2 g glucose, a heavy metal mix (10 mg MoNa_2O_4 , 2 mg CoCl_2 , 2 mg $\text{CuSO}_4 \cdot 5\text{H}_2\text{O}$, 10 mg $\text{MnCl}_2 \cdot 4\text{H}_2\text{O}$, 10 mg ZnCl_2 , 5 mg $\text{FeSO}_4 \cdot 7\text{H}_2\text{O}$, 50 mg $\text{CaCl}_2 \cdot 2\text{H}_2\text{O}$, 10 mg H_3BO_3), and an amino acid mix (50 mg of each tryptophan, tyrosine, valine, isoleucine, leucine, serine, cysteine, aspartate, glutamate, lysine, arginine, histidine, alanine, and threonine)] but also supplemented with 1 ng/ml biotin, 50 ng/ml L-(+)-selenomethionine, and 100 $\mu\text{g}/\text{ml}$ ampicillin, followed by serial expansion to a 10-liter batch culture (New Brunswick; BIOFLO 2000) in the same medium. After growth to an OD_{600} of 0.6 to 0.8, protein expression was induced and the protein purified as described above for native protein.

Crystallization of A197 and Collection of X-Ray Diffraction Data

Purified A197 was crystallized by hanging drop vapor diffusion at 22 °C. Drops were assembled with 2 μl of A197 (10.5 mg/ml in 10 mM Tris-HCl, pH 8.0, 50 mM NaCl) mixed with 2 μl of well solution (0.1 M morpholineethanesulfonic acid [MES], pH 6.5, 1.2 to 1.6 M ammonium sulfate, 10% 1,4-dioxane). Selenomethionine-incorporated A197 crystals were grown under identical conditions. Single rod-shaped or wedge-

shaped crystals were isolated, and cryoprotectant was introduced with a quick soak (10 to 60 seconds) in well solution supplemented with 22.5 to 25% glucose. Crystals were then flash frozen in liquid nitrogen.

Multiwavelength anomalous diffraction (MAD) was used to solve the structure. Data were collected to 2.28-Å resolution at the Se-K edge (peak, remote, and inflection wavelengths) on beamline 11-1 at the Stanford Synchrotron Radiation Laboratory (SSRL) (Table 2-4). An additional single-wavelength data set to 1.86-Å resolution was collected on beamline 1-5 at SSRL and used for refinement (Table 2-4). Data were integrated, scaled, and reduced in space group I222 (Table 2-4) using the HKL-2000 software package (91).

Table 2-4. X-ray Diffraction Data Collection for A197^a

Parameter	Result for Data Set ^b			
	Se-Edge	Se-Peak	Se-Remote	Refinement
Wavelength (Å)	0.97966	0.97935	0.94642	0.97845
Space Group	I222	I222	I222	I222
Cell Constants (a,b,c; Å; $\alpha = \beta = \gamma = 90^\circ$)	66.54, 70.96, 81.77	66.54, 70.96, 81.77	66.54, 70.96, 81.77	66.42, 70.56, 81.27
Resolution Range (Å)	30-2.28 (2.36-2.28)	30-2.28 (2.36-2.28)	30-2.28 (2.36-2.28)	29.1-1.86 (1.93-1.86)
Unique Reflections	8389 (893)	8403 (896)	8346 (890)	15646 (1575)
Average Redundancy	3.5 (3.6)	3.5 (3.6)	3.5 (3.6)	8.6 (8.5)
I/ σ	22.9 (11.2)	23.2 (11.7)	23.4 (11.9)	18.2 (4.1)
Completeness (%)	92.3 (99.9)	92.3 (99.9)	91.9 (99.9)	95.5 (99.1)
R _{sym} ^c (%)	5.0 (8.7)	5.3 (8.6)	4.7 (8.2)	5.1 (17.5)

^aData were integrated, scaled, and reduced using the HKL-2000 software package (91).

^bNumbers in parenthesis refer to the highest resolution shell.

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

In order to identify manganese-binding sites, crystals of native A197 were soaked in mother liquor supplemented with 10 mM $\text{MnSO}_4 \cdot \text{H}_2\text{O}$ for 4 hours. Crystals were flash frozen in liquid nitrogen as described above, and data to 3.0-Å resolution were collected in house using a Rigaku RUH3R X-ray generator producing Cu-K α X-rays and a MAR345 image plate detector. The data were integrated and scaled using HKL-2000 (91) and maps were prepared using various CCP4 programs (9) and the high resolution model.

X-ray Crystal Structure Determination and Model Refinement of A197

SOLVE (129) was used to determine the positions of the selenium substructure and to calculate initial phases. Four Se sites were identified, corresponding to one A197 molecule per asymmetric unit. The SOLVE output was then used in RESOLVE (127, 128) for density modification and initial model building. The resulting electron density map was of good quality, allowing the best parts of various models output by RESOLVE to be manually assembled into a composite model using O (59), followed by refinement with REFMAC5 (9, 86). The model was then further refined against the 1.86-Å data set. The refinement included the use of TLS (translation/libration/screw) parameters in which the model was broken into 12 TLS groups (1, 1 to 27; 2, 28 to 41; 3, 42 to 52; 4, 53 to 67; 5, 68 to 81; 6, 82 to 91; 7, 92 to 107; 8, 108 to 150; 9, 151 to 166; 10, 167 to 176; 11, 177 to 189; 12, 190 to 198) as suggested by the TLS motion determination home (<http://skuld.bmsc.washington.edu/~tlsmd/>) (92-94).

Iterative rounds of model building and refinement with O and REFMAC5 resulted in a final model with an R_{cryst} of 17.1% and an R_{free} of 20.4% (Table 2-5). The final

model has good stereochemistry, with all residues in allowed regions of the Ramachandran plot (74) (Table 2-5). Amino acids 139 to 147 were not modeled due to the lack of interpretable electron density, presumably due to disorder in this region of the protein. The protein quaternary structure server (<http://pqs.ebi.ac.uk/>) was used to calculate the surface area at the putative dimer interface. Structural comparisons were performed using the DALI (<http://www.ebi.ac.uk/dali>) (54) and VAST (<http://www.ncbi.nlm.nih.gov/Structure/VAST/>) (46) servers. Structural figures were generated with PYMOL (<http://www.pymol.org>) (35).

Table 2-5. Crystal Structure Model Refinement for A197^a

Parameter	Value
R_{cryst}^b (%)	17.1
R_{free}^b (%)	20.4
Real Space CC ^b (%)	95.5
Mean B Value (overall; Å ²)	18.3
Coordinate Error (based on maximum likelihood, Å)	0.082
RMSD from ideality:	
Bonds (Å)	0.015
Angles (°)	1.545
Ramachandran Plot ^c :	
Most Favored (%)	90.6
Additional Allowed (%)	9.4
PDB Accession Code	2C0N

^aModel refinement was carried out using Refmac5 (86).

^b $R_{\text{cryst}} = \sum ||F_o| - F_c| / \sum F_o$ 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%).

^bCorrelation coefficient (CC) is agreement between the model and $2mF_o - DF_c$ density map.

^cCalculated using Procheck (74).

Atomic Coordinates for A197

Atomic coordinates and structure factors for A197 are on deposit in the Protein Data Bank (www.pdb.org) under accession code 2C0N.

Materials and Methods for the Structure Solution of B116

Cloning the B116 Gene

The B116 open reading frame (ORF) was amplified by nested PCR directly from viral particles purified as previously described (112). The PCR primers added attB sites to facilitate ligase-free cloning using the Gateway system (Invitrogen), a Shine-Dalgarno sequence to facilitate efficient translation, and a C-terminal His-tag to facilitate protein purification via nickel-affinity chromatography. The internal forward and reverse primers were 5'-TTCGAAGGAGATAGAACCATGGGTAAGGTATTCCTCA-3' and 5'-GTGATGGTGATGGTGATG CCACACCTCATAAAATGAG-3', respectively, while the external forward and reverse primers were 5'-GGGGACAAGTTTGTACAAA AAAGCAGGCTTCGAAGGAGATAGAACC-3' and 5'-GGGGACCACTTTGTACAA GAAAGCTGGGTCCTAGTGATGGTGATGGTGATG-3', respectively. The sequence of the resulting entry clone was determined using ABI BigDye terminator cycle sequencing and is mutation free. The His-tagged B116 construct was then inserted into destination vector pDEST14 (Invitrogen), yielding the expression vector pEXP14-B116, for protein expression in *Escherichia coli*.

Expression, Purification, and Characterization of B116

Typically, pEXP14-B116 was transformed into *E. coli* strain BL21-CodonPlus (DE3)-RIL (Stratagene), and a single colony was used to inoculate 5 ml of LB medium and grown overnight, with subsequent serial expansion to a 10-liter fermentor (New Brunswick; BIOFLO 2000) batch culture, all at 37 °C. Medium for batch fermentation was as recommended by the manufacturer (109)(above). All media contained 100 µg/ml ampicillin. Fermentor batch cultures were supplemented with filtered air at a flow rate of 10 liters/minute and stirred at 300 to 500 rpm. Cells were grown to an optical density at 600 nm (OD₆₀₀) of 0.8 to 1.5, and protein expression was then induced by addition of 1 mM isopropyl-β-D-thiogalactopyranoside (IPTG). After an additional 4 to 6 hours of growth, cells were harvested by centrifugation (IEC PR-7000 M) at 6,000-×g for 20 minutes and pellets were stored at -80 °C until needed.

For purification, cell pellets were thawed and resuspended in lysis buffer (10 mM Tris-HCl, pH 8.0, 400 mM NaCl) at 10 ml/g of wet cell pellet mass.

Phenylmethylsulfonyl fluoride (PMSF; 0.1 mM) was added to the cell suspension, and cells were lysed by passage through a microfluidizer (Microfluidics Corporation, Newton, MA) or a Power laboratory press (American Instrument Co., Inc., Silver Spring, MD), depending on the volume of the cell suspension. The cell lysate was incubated at 65 °C for 20 minutes to denature many of the contaminating *E. coli* proteins and clarified by centrifugation (Beckman; Avanti J-30I) at 22,000-×g for 20 minutes. The resulting cleared lysate was applied to a column containing a 3- to 5-ml bed volume of HIS-Select nickel affinity gel (Sigma-Aldrich), washed with at least 10 column volumes of wash buffer (lysis buffer plus 5 mM imidazole), and eluted with elution buffer (10 mM Tris-

HCl, pH 8.0, 50 mM NaCl, 200 mM imidazole). Fractions containing B116 were then pooled and loaded onto a Superdex 75 gel filtration column (Amersham Biosciences) for further purification and buffer exchange into a minimal buffer (10 mM Tris-HCl, pH 8.0, 50 mM NaCl) for crystallization. Peak fractions were pooled and concentrated using a 10-kDa molecular mass cutoff (Amicon ultracentrifugal filter devices; Millipore) to 6.5 mg/ml. Protein concentrations were determined by Bradford assay (20) using “Protein Assay Reagent” (Bio-Rad) and bovine serum albumin (BSA) as a standard. Protein yield was typically 8 to 12 mg/g of cell pellet. Molecular weight and purity were assessed by 12% sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE).

For expression and purification of selenomethionine-incorporated B116, pEXP14-B116 was transformed into *E. coli* strain B834(DE3)pLysS (Novagen), a methionine auxotroph. Methionine auxotrophy was confirmed for a single colony, which was then used to inoculate 5 ml of medium, prepared essentially as described previously (130)(above), but supplemented with 1 ng/ml biotin, 50 ng/ml L-(+)-selenomethionine, and 100 µg/ml ampicillin, followed by serial expansion to a 10-liter batch culture (New Brunswick; BIOFLO 2000) in the same medium. After growth to an OD₆₀₀ of 0.6 to 0.8, protein expression was induced and the protein purified as described above for native protein. Selenium incorporation was monitored by matrix-assisted laser desorption ionization–time of flight (MALDI-TOF) mass spectrometry (MS) with a Bruker Biflex III in the matrix-cyano-4-hydroxycinnamic acid and found to be essentially 100%.

Crystallization of B116 and Collection of X-Ray Diffraction Data

Purified B116 was crystallized by hanging drop vapor diffusion at 17 °C. Drops were assembled with 2 µl of B116 (6.5 mg/ml in 10 mM Tris-HCl, pH 8.0, 50 mM NaCl) mixed with 2 µl of well solution (0.1 M Tris-HCl [pH 8.0], 1.95 M $\text{NH}_4\text{H}_2\text{PO}_4$). Selenomethionine-incorporated B116 crystals were grown under identical conditions. Single rod-shaped or wedge-shaped crystals were isolated, and cryoprotectant was introduced with a quick soak (30 to 300 seconds) in well solution supplemented with 10% glycerol. Crystals were then flash frozen in liquid nitrogen.

Single-wavelength anomalous diffraction (SAD) was used to solve the structure. Data were collected to 2.4 Å resolution at the peak Se-K wavelength on beamline X9-B at the National Synchrotron Light Source (NSLS) during the RapiData 2004 course at Brookhaven National Labs (Table 2-6). Data were integrated, scaled, and reduced in space group $\text{P2}_1\text{2}_1\text{2}_1$ (Table 2-6) keeping the Bijvoet pairs separate to allow for anomalous signal detection, using the HKL-2000 software package (91).

X-ray Crystal Structure Determination and Model Refinement of B116

SOLVE (129) was used to determine the positions of the selenium substructure and to calculate initial phases. Eight out of 10 possible Se sites were identified, corresponding to two B116 molecules per asymmetric unit. The SOLVE output was then used in RESOLVE (127, 128) for density modification and initial model building. The resulting electron density map was of good quality, allowing the best parts of various models output by RESOLVE to be manually assembled into a composite model using O (59), followed by refinement with REFMAC5 (9, 86). Model building and refinement

cycles were continued using then same indexed data set in which the Bijvoet pairs were merged upon scaling. The refinement included the use of TLS (translation/libration/screw) parameters in which each chain of the model was broken into five TLS groups (1:A2-A40, 2:A41-A75, 3:A76-A86, 4:A87-A108, 5:A109-A119, 6:B2-B24, 7:B25-B46, 8:B47-B84, 9:B85-B89, 10:B90-B118), for a total of 10 groups, with the assistance of the TLS Motion Determination Home (92-94) (<http://skuld.bmsc.washington.edu/~tlsmd/>).

Table 2-6. X-ray Diffraction Data Collection for B116^a

Parameter	Result ^b for Data Set:
	Se-Peak
Wavelength (Å)	0.97946
Space Group	P2 ₁ 2 ₁ 2 ₁
Cell Constants (a,b,c; Å) $\alpha=\beta=\gamma=90^\circ$	48.6, 83.3, 89.5
Resolution Range (Å)	30-2.4 (2.49-2.4)
Unique Reflections	14,911 (1,466)
Average Redundancy	2.8 (2.8) ^c
	5.1 (5.3) ^d
I/ σ	22 (5)
Completeness (%)	99.9 (100)
R _{sym} ^e	0.07 (0.26)

^aData were integrated, scaled, and reduced using the HKL-2000 software package (27).

^bNumbers in parenthesis refer to the highest resolution shell.

^cBijvoet pairs separate for scaling.

^dBijvoet pairs merged for scaling.

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

Iterative rounds of model building and refinement with O and Refmac5 resulted in a final model with an R_{cryst} of 21% and an R_{free} of 24% (Table 2-7). The final model has good stereochemistry, with all residues in allowed regions of the Ramachandran plot (74) (Table 2-7). The protein quaternary structure server (<http://pqs.ebi.ac.uk/>) was used to

calculate the surface area at the putative dimer interface. Structural comparisons were performed using the DALI (54) (<http://www.ebi.ac.uk/dali>) and VAST (46) (<http://www.ncbi.nlm.nih.gov/Structure/VAST/>) servers. Structural figures were generated with PyMol (35) (<http://www.pymol.org>).

Table 2-7. Crystal Structure Model Refinement for B116^a

Parameter	Value
Resolution (Å)	29.81 - 2.39
R_{cryst}^b (%)	20.99
R_{free}^b (%)	24.05
Real Space CC ^c (%)	93
Mean B Value (overall; Å ²)	28.3
Coordinate Error (based on maximum likelihood, Å)	0.15
RMSD from ideality:	
Bonds (Å)	0.014
Angles (°)	1.35
Ramachandran Plot ^d :	
Most Favored (%)	96.3
Additional Allowed (%)	3.7
PDB Accession Code	2J85

^aModel was refined using refmac5 (86).

^b $R_{\text{cryst}} = \Sigma ||F_o| - F_c| / \Sigma F_o$ 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 (5.1%).

^cCorrelation coefficient (CC) is agreement between the model and $2mF_o - DF_c$ density map.

^dCalculated using Procheck (28).

Atomic Coordinates of B116

Atomic coordinates and structure factors for B116 are on deposit in the Protein Data Bank (www.pdb.org) under accession code 2J85.

Materials and Methods for the Structure Solution of F93

Cloning the F93 Gene

ORF F93 was amplified by nested PCR directly from viral particles purified as previously described (112). The PCR primers added attB sites to facilitate ligase-free cloning using the Gateway system (Invitrogen), a Shine-Dalgarno sequence to facilitate efficient translation, and a C-terminal His-tag to facilitate protein purification via immobilized metal affinity chromatography (IMAC). The internal forward and reverse primers were 5'- TTCGAAGGAGA TAGAACCATGAAAATCAGAAAGTATATGA-3' and 5'- GTGATGGTGATGGTGATGGCA TACTATTTTTTTTGAAAAATT-3', while the external forward and reverse primers were 5'-GGGGACAAGTTTGTACAAAAAAG CAGGCTTCGAAGGAGATAGAACC-3' and 5'-GGGGACCACTTTGTACAAGAAAG CTGGGTCCTAGTGATGGTGATGGTGATG-3', respectively. The resulting entry clone was sequenced using ABI BigDye Terminator Cycle sequencing and was confirmed to be mutation-free. The His-tagged F93 construct was then inserted into destination vector pDEST14 (Invitrogen), yielding the expression vector pEXP14-F93, for protein expression in *E. coli*.

Expression, Purification, and Characterization of F93

Typically, pEXP14-F93 was transformed into BL21-CodonPlus (DE3)-RIL *E. coli* (Stratagene) and a single colony used to inoculate 5 ml of Luria-Bertani medium and grown overnight, with subsequent serial expansion to a 10 L fermentor (New Brunswick BIOFLO 2000) batch culture, all at 37 °C. Medium for batch fermentation was as recommended by the manufacturer (109). All media contained 100 µg/ml ampicillin.

Fermentor batch cultures were supplemented with filtered air at a flow rate of 10 L/minute and stirred at 300-500 RPM. Cells were grown to an OD₆₀₀ of 0.8 to 1.5 and protein expression was then induced with addition of 1mM isopropyl-beta-D-thiogalactopyranoside (IPTG). After an additional four to six hours of growth, cells were harvested by centrifugation (IEC PR-7000M) at 6,000- $\times$ g for 20 minutes and pellets were stored at -80 °C until needed.

For purification, cell pellets were thawed and resuspended in lysis buffer (10 mM Tris-HCl, pH 8.0, 400 mM NaCl) at 10 mL/gram of wet cell pellet mass. Phenylmethylsulphonyl-fluoride (PMSF, 0.1 mM) was added to the cell suspension and cells were lysed by passage through a microfluidizer (Microfluidics Corporation, Newton, MA) or a Power Laboratory Press (American Instrument Co., Inc., Silver Springs, MD), depending on the volume of the cell suspension. The cell lysate was incubated at 65 °C for 20 minutes to denature many of the contaminating *E. coli* proteins and clarified by centrifugation (Beckman Avanti J-30I) at 22,000- $\times$ g for 20 minutes. The resulting cleared lysate was applied to a column containing a 3- to 5-mL bed volume of HIS-Select™ Nickel Affinity Gel (Sigma-Aldrich), washed with at least 10 column volumes of wash buffer (lysis buffer plus 5 mM imidazole), and eluted with elution buffer (10 mM Tris-HCl, pH 8.0, 50 mM NaCl, 200 mM imidazole). Elution fractions containing F93 were pooled and loaded onto a Superdex™ 75 gel filtration column (Amersham Biosciences) for further purification and buffer exchange into a minimal buffer (10 mM Tris-HCl, pH 8.0, 50 mM NaCl) for crystallization. Peak fractions were pooled and concentrated using 5-kDa molecular weight cutoff Amicon Ultra Centrifugal Filter Devices (Millipore) to 10.5 mg/mL. Protein concentrations were determined by

Bradford assay (20), using Protein Assay Reagent (Bio-Rad) and bovine serum albumin (BSA) as a standard. Protein yield was typically 3 to 7 mg/g of cell pellet. Molecular weight and purity were assessed with 14 % sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE) using a Tris-tricine buffer system.

For expression and purification of selenomethionine incorporated F93, pEXP14-F93 was transformed into B834 (DE3) pLysS *E. coli* (Novagen), a methionine auxotroph. Methionine auxotrophy was confirmed for a single colony, which was then used to inoculate 5 ml of medium, essentially as described (130), but supplemented with 1 ng/ml biotin, 50 ng/ml L(+)-selenomethionine, and 100 µg/ml ampicillin, followed by serial expansion to a 10 L fermentor batch culture (New Brunswick BIOFLO 2000) in the same medium. After growth to OD₆₀₀ of 0.6 to 0.8, protein was expressed and purified as described for native protein (above).

Crystallization of F93 and Collection of X-Ray Diffraction Data

Purified F93 was crystallized by hanging drop vapor diffusion at 22 °C. Drops were assembled with 2 µL of F93 (9.0-10.5 mg/ml in 10 mM Tris-HCl, pH 8.0, 50 mM NaCl, 0.5 mM Tris(2-carboxyethyl)phosphine hydrochloride (TCEP)) mixed with 2 µL of well solution (0.1 M sodium acetate, pH 4.75-5.0, 0.1 M magnesium nitrate hexahydrate, 16-20 % PEG 20,000). Selenomethionine-incorporated F93 crystals were grown under identical conditions. Single rod-shaped or plate-shaped crystals were isolated and cryoprotectant introduced with a quick soak (30-300 seconds) in well solution supplemented with 12.5-15 % glycerol. Crystals were then flash frozen in liquid nitrogen.

The structure was solved using single-wavelength anomalous diffraction (SAD). Data were collected from a crystal of selenomethionine-incorporated protein to 2.4 Å resolution at the Se-K edge (peak wavelength, 0.97908 Å) on beamline 9-2 at the Stanford Synchrotron Radiation Laboratory (SSRL) (Table 2-8). An additional dataset from a native crystal was collected to 2.2 Å resolution on beamline 9-1 at SSRL and was used for refinement (Table 2-8). Data were integrated, scaled, and reduced in space group C2221 (Table 2-8) using HKL-2000 (91).

Table 2-8. X-ray Diffraction Data Collection for F93^a.

Parameter	Result ^b for Data Set:	
	Se-Peak	Native
Wavelength (Å)	0.97908	0.98789
Space Group	C222 ₁	C222 ₁
Cell Constants (a,b,c; Å) $\alpha=\beta=\gamma=90^\circ$	41.99, 102.55, 92.76	42.05, 102.59, 92.40
Resolution Range (Å)	50-2.4 (2.49-2.40)	50-2.2 (2.28-2.20)
Unique Reflections	8000 (791)	10466 (1005)
Average Redundancy	6.3 (5.8)	7.1 (7.3)
I/ σ	28.8 (5.5)	33.0 (4.2)
Completeness (%)	97.8 (99.4)	99.8 (100)
R _{sym} ^c	0.068 (0.22)	0.058 (0.34)

^aData were integrated, scaled, and reduced using the HKL-2000 software package (91).

^bNumbers in parenthesis refer to the highest resolution shell.

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

X-ray Crystal Structure Determination and Model Refinement of F93

SOLVE (129) was used to determine the positions of the selenium-substructure and to calculate initial phases. Five of six possible Se sites were identified corresponding to two F93 molecules, one with two and one with three Se sites, per asymmetric unit.

The SOLVE output was then used in RESOLVE (127, 128) for density modification and initial model-building. The resulting electron density map was of good quality, allowing the best parts of various models output by RESOLVE to be manually assembled into a composite model using O (59). To aid initial modeling efforts, the electron density was further modified by noncrystallographic symmetry (NCS) averaging with DM (32), using the rotation matrix between individual chains in the asymmetric unit determined by Lsqkab (61). Building and refinement proceeded with iterative cycles of model building, using O (59) or Coot (39), and refinement with Refmac5 (9, 86).

As the model neared completion, we switched to the native 2.2 Å data set. Refinement included the use of TLS parameters in which the model was broken into seven TLS groups per monomer (1:A5-A12, 2:A13-A27, 3:A28-A39, 4:A40-A62, 5:A63-A78, 6:A79-A83, 7:A84-A96, 8:B1-B16, 9:B17-B36, 10:B37-B51, 11:B52-B62, 12:B63-B78, 13:B79-B87, 14:B88-B94) as suggested by the TLS Motion Determination Home (92-94) (<http://skuld.bmsc.washington.edu/~tlsmd/>). The final model results in an R_{cryst} of 19.0% and an R_{free} of 23.2% (Table 2-9). The model exhibits good geometry, with all residues falling in allowed regions of the Ramachandran plot (74) (Table 2-9). Residues A1-4, A97-99, and B95-99 were not modeled due to the lack of interpretable electron density, presumably due to disorder in the terminal regions of the protein chains. The Protein Quaternary Structure server (<http://pqs.ebi.ac.uk/>) was used to calculate the surface area at the putative dimer interface. Structural comparisons were performed using the DALI (54) (<http://www.ebi.ac.uk/dali>) and VAST (46) (<http://www.ncbi.nlm.nih.gov/Structure/VAST>) servers. Structural figures were generated with PyMOL (35) (<http://www.pymol.org>).

Atomic Coordinates of F93

Atomic coordinates and structure factors for F93 are on deposit in the Protein Data Bank (<http://www.pdb.org>) under accession code 2CO5.

Table 2-9. Crystal Structure Model Refinement for F93.

Parameter	Value
R_{cryst}^b (%)	19.0
R_{free}^b (%)	23.2
Real Space CC^c (%)	95.7
Mean B Value (overall; $\text{\AA}^2$)	42.8
Coordinate Error (based on maximum likelihood, $\text{\AA}$)	0.170
RMSD from ideality:	
Bonds ($\text{\AA}$)	0.007
Angles ($^\circ$)	0.991
Ramachandran Plot ^d :	
Most Favored (%)	90.6
Additional Allowed (%)	9.4
PDB Accession Code	2co5

^aRefinement was carried out with Refmac5 (9, 86).

^b $R_{\text{cryst}} = \sum ||F_o| - F_c| / \sum F_o$ 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%).

^cCorrelation coefficient (CC) is agreement between the model and $2mF_o - DF_c$ density map.

^dCalculated using Procheck (74).

Subcellular Localization of STIV Proteins

Proteins known to be present in the purified viral particle or exhibiting detectable bacterial or eukaryotic signal sequences as determined by the program SignalP (14) were considered to be extracellular. Putative integral membrane proteins were identified using the program TMHMM (69, 125). These programs are not designed specifically for the Archaea so to be conservative, a positive signal from any of the Eukarya or Bacteria tests

was considered positive, and the protein was removed. The remaining proteins represent a genomic pool enriched in intracellular proteins. Because even numbers of cysteine residues are frequently employed in metal binding, these motifs were identified using the program ScanProsite (45, 56, 122) or taken from the primary literature (105), and removed from the intracellular pool to prevent potential false positives. Prosite did not identify any metal-binding motifs but recent bioinformatics work by Prangishvili *et al.* (105) identified Zn-finger motifs in A66 and C57 so these proteins were removed to be conservative. These results are presented in Table 2-10.

Summary

All ORFs within the STIV genome coding for more than 60 amino acids in length have been cloned into vectors suitable for protein expression in *E. coli*. The one exception, however, is ORF A78, which starts with a noncanonical GTG start codon and was only recently discovered through proteomics studies of purified STIV virions by Walid Maaty in Brian Bothner's lab (78). In addition, ORF A55 was cloned prior to narrowing of the focus to ORFs greater than 60 amino acids in length.

Varying degrees of soluble protein expression and purity were observed for several proteins. These are: B109, A197, A223, C381, B116, A81, A77, B264, A109, B345, and F93. In addition, Dirk Reiter, an exchange student from Tübingen, Germany who was doing research in our lab and helped with the initial cloning and expression experiments, cloned A510 into a yeast expression vector and was able to obtain soluble protein under low pH elution conditions. We did note that precipitated material eluted off the nickel column from A510 bacterial expression so it is possible that this is indeed

Table 2-10. STIV ORF Information Relevant to Disulfide Bonds in the Intracellular Proteins of the Viral Proteome.

Internal ORF ID #	ORF name	# amino acids	Known capsid protein ^a				predicted signal sequence ^b				Metal- binding motif from literature ^c				Filtered from genome?
			% Cys	# of Cys	% Ser	# of Ser	% Leu	Eukarya	Bacteria	Gram- pos. Bacteria	Gram- neg. Bacteria	predicted TM helices ^c	Prosite binding motifs ^d	prosited metal- binding motif	
1	B109	109	0	0.00%	3	2.75%	15	13.76%	yes	no	no	no	no	no	yes
2	C67	67	0	0.00%	12	17.91%	4	5.97%	no	no	no	37-59	no	no	yes
3	B130	130	0	0.00%	6	4.62%	14	10.77%	yes	no	no	36-58, 105-127	no	no	yes
4	A55	55	0	0.00%	1	1.82%	6	10.91%	yes	53-54	no	32-54	no	no	yes
X	A78	78	0	0.00%	2	2.56%	8	10.26%	yes	no	no	no	no	no	yes
5	A510	510	2	0.39%	32	6.27%	53	10.39%	no	no	no	no	no	no	no
6	C63	63	0	0.00%	11	17.46%	6	9.52%	no	no	no	no	no	no	no
7	C92	92	0	0.00%	6	6.52%	7	7.61%	no	26-27	no	5-27	no	no	yes
8	A61	61	0	0.00%	1	1.64%	9	14.75%	no	no	no	no	no	no	no
9	A197	197	2	1.02%	9	4.57%	20	10.15%	no	15-16	no	no	no	no	yes
10	A223	223	0	0.00%	21	9.42%	19	8.52%	yes	no	no	no	no	no	yes
11	C381	381	1	0.26%	43	11.29%	32	8.40%	yes	no	14-15	no	no	no	yes
12	B106	106	6	5.66%	6	5.66%	12	11.32%	no	no	no	no	no	no	no
13	B116	116	2	1.72%	5	4.31%	14	12.07%	no	no	no	no	no	no	no
14	A53	53	6	11.32%	1	1.89%	3	5.66%	no	no	no	no	no	no	no
15	C118	118	0	0.00%	5	4.24%	16	13.56%	no	no	no	no	no	no	no
16	A66	66	2	3.03%	1	1.52%	7	10.61%	no	no	no	no	no	C2H2 Zn finger	yes
17	C57	57	2	3.51%	0	0.00%	4	7.02%	no	no	no	no	no	C2H2 Zn finger	yes
18	A106	106	0	0.00%	2	1.89%	13	12.26%	no	no	no	no	no	no	no
19	A137	137	0	0.00%	10	7.30%	9	6.57%	no	no	no	no	no	no	no
20	B97	97	1	1.03%	6	6.19%	12	12.37%	no	no	no	no	no	no	no
21	A58	58	0	0.00%	0	0.00%	7	12.07%	no	no	no	no	no	no	no

Table 2-10 (cont.). STIV ORF Information Relevant to Disulfide Bonds in the Intracellular Proteins of the Viral Proteome.

Internal ORF ID	ORF #	name	# amino acids	Known capsid				predicted signal sequence ^b				Metal- binding motif				Filtered from genome?
				% Cys	# of Cys	% Ser	# of Ser	% Leu	protein ^a	Eukarya	Bacteria	Gram- pos.	Gram- neg.	TM helices ^c	binding motifs ^d	
22	B66b		66	4	6.06%	2	3.03%	9	13.64%	no	no	no	no	no	no	no
23	B66a		66	0	0.00%	16	24.24%	7	10.61%	no	no	no	14-15	no	no	yes
24	C121		121	0	0.00%	4	3.31%	11	9.09%	no	no	no	no	no	no	no
25	A81		81	0	0.00%	5	6.17%	11	13.58%	no	no	no	no	no	no	no
26	B124		124	0	0.00%	5	4.03%	6	4.84%	no	no	no	no	97-119	no	yes
27	A77		77	0	0.00%	3	3.90%	4	5.19%	no	no	no	no	no	no	no
28	B264		264	2	0.76%	12	4.55%	25	9.47%	no	no	no	no	no	no	no
29	A109		109	2	1.83%	6	5.50%	12	11.01%	no	no	no	no	no	no	no
30	B164		164	0	0.00%	11	6.71%	14	8.54%	yes	no	no	no	no	no	yes
31	B345 (MCP)		345	0	0.00%	27	7.83%	28	8.12%	yes	no	no	no	no	no	yes
32	C557		557	0	0.00%	50	8.98%	55	9.87%	yes	no	no	no	364-386, 538-556	no	yes
33	A51		51	0	0.00%	1	1.96%	10	19.61%	no	no	no	no	no	no	no
41	D66		66	2	3.03%	6	9.09%	10	15.15%	no	no	26-27	no	no	no	yes
42	F75		75	4	5.33%	7	9.33%	11	14.67%	no	19-20	no	no	no	no	yes
46	F93		93	1	1.08%	4	4.30%	12	12.90%	no	no	no	no	no	no	no
			total %	0.76%				6.65%				10.02%				

^a From: Rice, et. al. 2004. PNAS 101:7716-7720., Khayat, et. al. 2005. PNAS 102:18944-18949., Maaty, et. al. 2006. J. Virol. 80:7625-7635.^b Predictions done with SignalP server (<http://www.cbs.dtu.dk/services/SignalP/>).^c Predictions done with TMHMM server (<http://www.cbs.dtu.dk/services/TMHMM/>).^d Predictions done with ScanProsite server (<http://ca.expasy.org/tools/scanprosite/>).^e From: Prangishvili, et. al. 2006. Virus Research 117:52-67.

A510 and the bacterial purification should be re-examined under low pH conditions in light of Dirk's work in yeast.

Crystallization experiments were set up for all proteins that expressed any level of soluble protein with the exception of A77. This protein only expresses from shaker flasks and we were not able to produce very pure material despite efforts to add ion exchange or hydrophobic interaction chromatographic steps to the purification. It has been observed in some of the SSV proteins that simply moving the His tag from the C- to the N-terminus has resulted in much cleaner protein. Alternate expression constructs of A77 should be pursued.

Pure B109 and B264 were never obtained, as there were always two prominent bands that co-purified despite great efforts to separate them. Crystallization experiments were set up for both of these proteins in the double-band form anyway. Björn Tränkle, an exchange student from Tübingen, Germany who was doing research in our lab, subsequently found that the double-band material of B264 is in fact the result of the use of an alternate start codon at Met⁴². He recloned this ORF starting at Met⁴² and now pure protein is obtainable. This does not explain the two bands seen for ORF B109, however, because the start Met is the only one in the sequence. It is possible that the two bands seen for this protein are the result of a specific cleavage event. Efforts should be made to get N-terminal Edman or mass spectroscopic sequencing information about this protein so an alternate construct can be created for it as well. This protein is relatively small so perhaps a simple MALDI experiment to determine the mass of the two bands will be sufficient to identify the cleavage site. Alternate constructs should be explored.

Large, diffraction quality crystals were obtained for A197, B116, and F93. These three ORF products have had their structures determined and are the focus of this dissertation. The clone for B345, the major capsid protein (MCP), yielded soluble, homogenous protein following IMAC and size exclusion chromatography. The purified protein ran as a monomer on the S75 size exclusion column and was used for production of polyclonal antibodies. The clone was then provided to Jack Johnson's laboratory at the Scripps Research Institute and its structure was solved through the efforts of Dr. Reza Khayat (65).

Initial crystallization conditions were identified for A81, an apparent tetramer. Nathanael Lintner, an undergraduate in the Lawrence lab, has since been optimized these conditions to yield diffraction quality crystals. Nathanael has collected a high-resolution native data set. The protein only contains Met¹ and so is not suitable for selenomethionyl derivatization. Nathanael has thus spent some time trying to identify isomorphous heavy atom derivatives. He is also pursuing an alternative strategy; he has used site-directed mutagenesis to create three different mutants with a single Met introduced into three separate locations along the primary sequence. He has been able to produce isomorphous crystals of selenomethionyl-derivative protein and has collected a full dataset. Attempts at single isomorphous replacement (SIR) with this dataset and the native have provided initial phases, but a second derivative will be required. He has other mutants that could potentially be used for multiple isomorphous replacement (MIR). These possibilities are being pursued while he is awaiting synchrotron beam time. However, with a tunable x-ray source, he should be able to solve the structure with the single selenomethionyl derivative using single- or multiple-wavelength isomorphous diffraction (SAD and MAD,

respectively), possibly combined with single or multiple isomorphous replacement (SIRAS, MIRAS).

Björn Tränkle obtained initial crystallization conditions for A223. Anoop Sendamarai, a graduate student in the Lawrence lab, has been optimizing these crystals and has been able to obtain a low-resolution data set (~ 4 Å) from a very small crystal. He is continuing to optimize these crystallization conditions. This protein has a single Met, excluding Met¹, amounting to less than 0.5% Met content, making a detectable anomalous signal difficult to obtain. SAD/MAD methods with the selenomethionyl derivative will be pursued regardless, but this protein will likely require other heavy atom derivatives or the introduction of more methionines by mutagenesis to solve its structure.

Expression clones for A109 and A223 have been provided to Valerie Copie's lab to pursue structural and dynamic studies of these proteins using nuclear magnetic resonance spectroscopy. Jack Johnson's laboratory at the Scripps Research Institute is pursuing the structures of the structural proteins, i.e. those that make up the capsid as identified by the proteomic studies of Walid Maaty (78), while we are primarily focusing on the rest of the proteins, those that likely function within the host. As such, we have provided expression clones for B109, B130, A55, A223, C381, B164, B345 (MCP, whose structure was solved), and C557 to the Johnson lab. Despite our focus primarily being the non-structural proteins and the structural proteins being that of the Johnson lab, we are continuing to work with A223 and C381 because we had some success with those proteins prior to the knowledge that they are structural proteins. Various expression clones have been provided to the Young lab and Bothner lab for their studies as well.

Polyclonal antibodies (rabbit) have been made against several STIV proteins to be used in many applications, from Western Blot to immunoprecipitation. The Young lab has made antibodies against B345 and we have made antibodies against A510, A197, A223, C381, B116, and F93. These antibodies will be valuable in the functional characterization of these proteins and are currently being used by the Young lab to identify the proteins from virally infected cultures. They are available for use by any of our collaborators.

Currently, structures are being pursued for A81 and A223. Based on our hypotheses for the functions of A197, B116, and F93 that resulted from the structural studies of these proteins, we are in a position to start pursuing the functional studies of these proteins. Some of these functional studies are currently underway.

CHAPTER 3

STRUCTURAL CHARACTERIZATION OF A197; A GLYCOSYTRANSFERASE
DISPLAYING THE GT-A FOLDIntroduction

A197 is a 197-residue protein in open reading frame A of the STIV genome (Figure 3-1) whose sequence provides little insight into its function. For example, standard BLAST (4) and conserved domain (CD) searches (80) with default settings fail to identify proteins with similar sequence or structure. Further, the protein fold recognition algorithm that uses 1D and 3D sequence profiles coupled with secondary structure and solvation potential information utilized by the protein threading server 3D-pssm (64) fails to find meaningful matches. The highest scoring hit is nmh adenyltransferase from the archaeon *Methanococcus jannaschii* with an E-value of 2.09, however none of the prediction results (alignment, secondary structure matching, text matching, or structure/function classification) suggest any the level of confidence in this prediction and in fact many of these “results” are completely absent. However, as explained previously protein structure can lead to specific suggestions regarding protein function. We thus pursued the solution of the x-ray crystal structure of A197 with the goal of inferring its function and possible role in the viral life cycle. The protein expresses well in *E. coli* and lead to the formation of nice crystals suitable for X-ray diffraction analysis. Here we report on the structure of A197, one of the first STIV proteins to lend itself to x-ray crystallographic study. We identify it as a

glycosyltransferase. To our knowledge, this represents the first structure of a viral or archaeal glycosyltransferase displaying the GT-A fold and also one of the smallest glycosyltransferases. Potential roles for A197 in the STIV life cycle and their evolutionary implications are discussed.

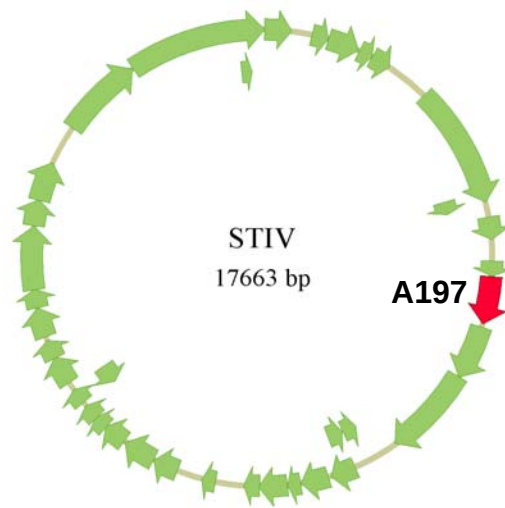


Figure 3-1. Location of A197 in the STIV Genome.

Results

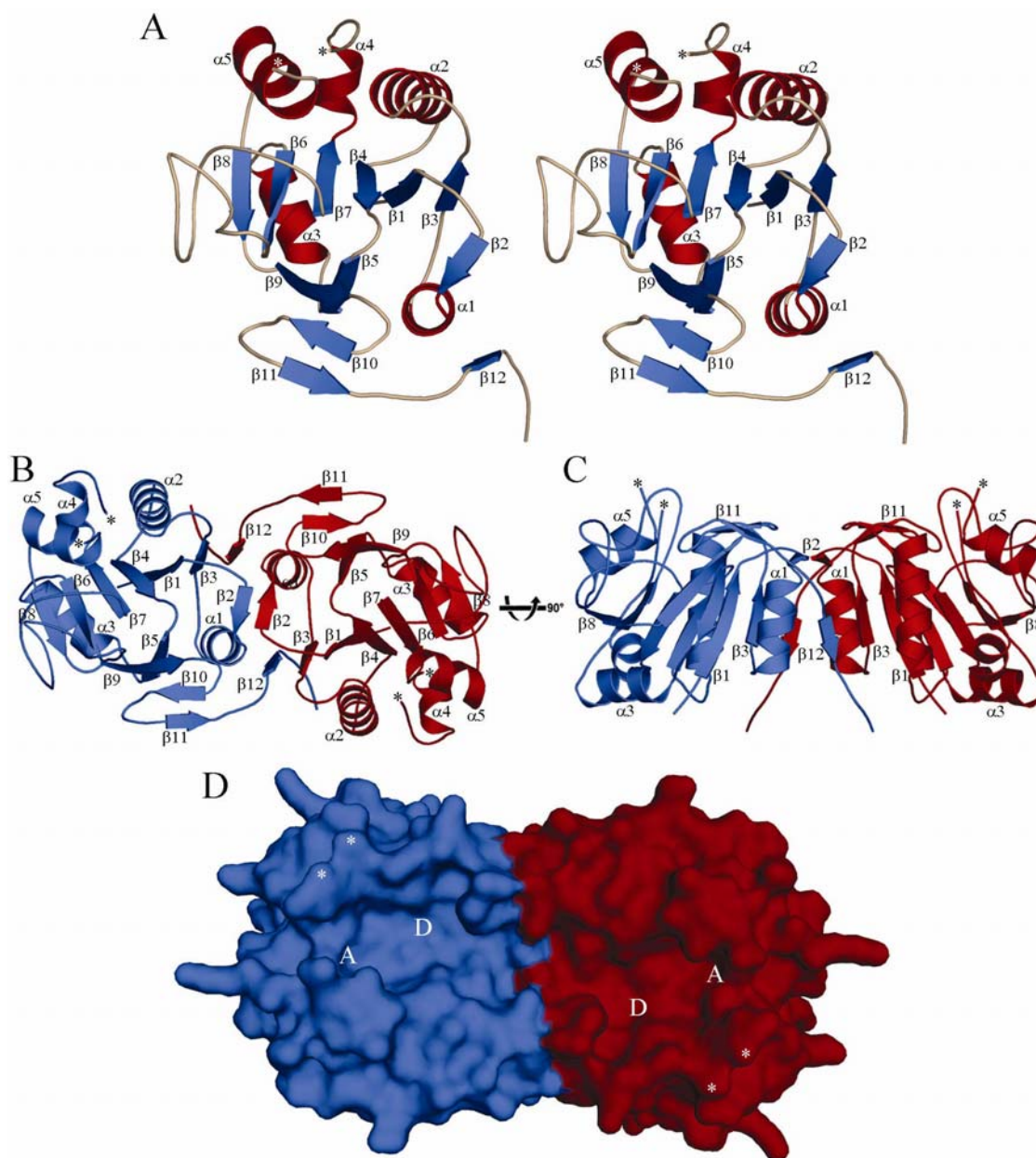
The A197 construct used in this study codes for the 197 amino acids of the native protein plus an additional C-terminal His-tag, for a total of 203 residues with a calculated mass of 24,090 Da. A197 elutes from the Superdex 75 size exclusion column as a single peak with an apparent molecular mass of approximately 20 kDa. In contrast, dynamic light scattering indicates a monodisperse solution with an average particle diameter of approximately 9.5 nm, suggesting that A197 is present as a higher-order oligomer in

solution. A197 crystallizes in space group I222 with one copy of the A197 polypeptide in the asymmetric unit. The structure was initially solved at a resolution of 2.28 Å by multiwavelength anomalous diffraction (MAD) using selenomethionyl-incorporated protein. The final model was refined against data collected on a second crystal to a resolution of 1.86 Å. Details on data collection and model refinement are presented in Tables 2-4 and 2-5, respectively, of Chapter 2. The structure has been deposited in the Protein Data Bank (www.pdb.org) under accession code 2C0N.

Structure of A197

The structure of the A197 monomer reveals a six-stranded, predominantly parallel, $\alpha/\beta/\alpha$ sandwich that is flanked by a four-stranded antiparallel β -sheet with an extended C terminus (Figure 3-2A). This C-terminal tail shows significant interaction with a neighboring molecule in the crystal (Figure 3-2B). Residues 139 to 147, which connect helices $\alpha 4$ and $\alpha 5$, are disordered in the crystal structure and thus are not included in the model. The topology of the central β -sheet is 3, 1, 4, 7, 6, 8, with strand 7 running in the antiparallel direction, while the flanking four-stranded antiparallel β -sheet is composed successively of β -strands 5, 9, 10, and 11.

In concert with a neighboring molecule of A197, β -strands 2, 3, and 12 are involved in the formation of intermolecular β -sheets (Figure 3-2B and C). β -strand 2 interacts across a crystallographic twofold axis, hydrogen bonding to an equivalent β -strand in the neighboring molecule to form a two-stranded antiparallel β -sheet. β -strand 12 interacts with the same neighboring molecule, hydrogen bonding to β -strand 3 in an



antiparallel fashion. This results in extension of the central β -sheet of the $\alpha/\beta/\alpha$ sandwich to seven strands, showing a topology of 12, 3, 1, 4, 7, 6, 8, with strands 12 and 7 running antiparallel to the remaining strands. These significant interactions suggest that the dimer seen in the crystal is related to the higher-order oligomer indicated by dynamic light scattering. This putative dimer exhibits overall dimensions of approximately 80 Å by 45 Å by 40 Å. Relative to known dimer interfaces, this putative dimer interface is extensive, with occlusion of 1,950 Å² of solvent-accessible surface area per monomer and a buried surface that is rich in hydrophobic residues.

The top edge of the central β -sheet, along with β -strand 2, creates the “floor” of a large pocket that is lined by elements of helices $\alpha 2$ and $\alpha 5$ and the $\beta 4$ - $\beta 5$ loop on one side and by the N-terminal half of the long $\beta 6$ - $\beta 7$ loop, β -strand 5, β -strand 9, and much of the $\beta 9$ - $\beta 10$ loop on the other. Formation of the crystallographic dimer results in an extended cleft that spans the dimer interface (Figure 3-2D). The largest pocket on a protein surface is usually involved in ligand recognition, and in the case of enzymes, will mark the active site (76). Thus, this pocket is likely to play an important role in the activity of A197.

Structural Homology with Glycosyltransferases

A DALI search (54) for structurally similar proteins identifies members of the glycosyltransferase superfamily as the closest structural homologues.

Glycosyltransferases catalyze the transfer of a sugar moiety from an activated donor sugar onto saccharide and non-saccharide acceptors (31). Members of the glycosyltransferase superfamily exhibit extreme diversity in their amino acid sequences, reflecting the large number of different donor and acceptor molecules that are utilized by

this class of enzymes. Sequence similarity and biochemical characterization have been used to group the glycosyltransferase superfamily into 87 distinct families (as of August 2006) designated GT1 to GT87 (30, 31). However, this number is certain to grow as more of these genes are identified and their products characterized. Characteristics of the various families along with current membership are available from the carbohydrate-active enzyme (CAZy) database (<http://afmb.cnrs-mrs.fr/CAZY/>) (30, 31).

In contrast to the diverse sequence composition of these enzymes, structural studies have identified only two predominant folds, designated GT-A and GT-B, though a third fold related to the GT-A fold has been found in sialyltransferase CstII from *Campylobacter jejuni* (23, 28, 33). Recent structural work in a number of laboratories has allowed the GT-A and GT-B folds to be assigned to many of the GT families (22). Importantly, our DALI search with A197 identified glycosyltransferases that exhibit the GT-A fold. Interestingly, three of the four most similar matches are to eukaryotic glycosyltransferases. These show highly significant Z scores (≥ 9) and root mean square deviations (RMSDs) near 3.0 Å for superposition of structurally equivalent C $_{\alpha}$ atoms, despite less than 15% sequence identity over equivalent residues.

The structure of A197 differs from the canonical GT-A fold in a few minor ways. First, the canonical fold shows a seven-stranded central β -sheet, whereas there are only six strands in A197. The β -strand running along the outside edge of the central β -sheet and the preceding α -helix are not present. In A197, however, dimer formation supplies a replacement in the form of β -strand 12, though it runs in the opposite direction (antiparallel). A second minor difference is the addition of β -strands 10 and 11 to extend the size of the flanking antiparallel β -sheet, though extended versions of this flanking β -

sheet have been seen before. For example, one the founding members of the GT-A fold family, the nucleotide-diphospho-sugar transferase from *Bacillus subtilis*, SpsA (27), also shows a four-stranded antiparallel β -sheet, though the connectivity differs from that in A197. Apart from the addition to this flanking β -sheet in A197, there is little decoration of its GT-A fold. In contrast, most enzymes with glycosyltransferase activity are significantly larger than A197 and often contain additional domains. Thus, this putative viral glycosyltransferase represents a minimal form of this large, extremely diverse, and highly important class of enzymes.

A197 shows greatest structural similarity to the catalytic domain of another founding member of the GT-A fold family, rabbit N-acetylglucosaminyltransferase I (GnT I) (136), an inverting glycosyltransferase in CAZy family GT13 (Figure 3-3B). DALI identifies 149 structurally equivalent residues, with 9% sequence identity, whose C_α atoms superpose with an RMSD of 2.9 Å. A refined superposition of the two proteins with Isqman (66) identifies a highly conserved structural core (Figure 3-4) consisting of 129 equivalent C_α atoms with 12.4% identity and an RMSD of 1.78 Å.

While the sequences of the many glycosyltransferase families are highly divergent, those that adopt the GT-A fold generally possess a DXD motif. The DXD typically serves to coordinate a divalent cation, usually Mn^{2+} , which in turn interacts with the donor substrate through the diphosphate group (22). This motif is invariantly found in a short loop or β -turn that connects an internal strand of the central β -sheet to the first β -strand in the flanking β -sheet (22). Such is the case for A197, where the DXD motif is situated in the short loop that connects strands β 4 and β 5 and is composed of residues Asp⁶¹, Glu⁶², and Asp⁶³. Superposition of A197 upon its closest structural neighbors

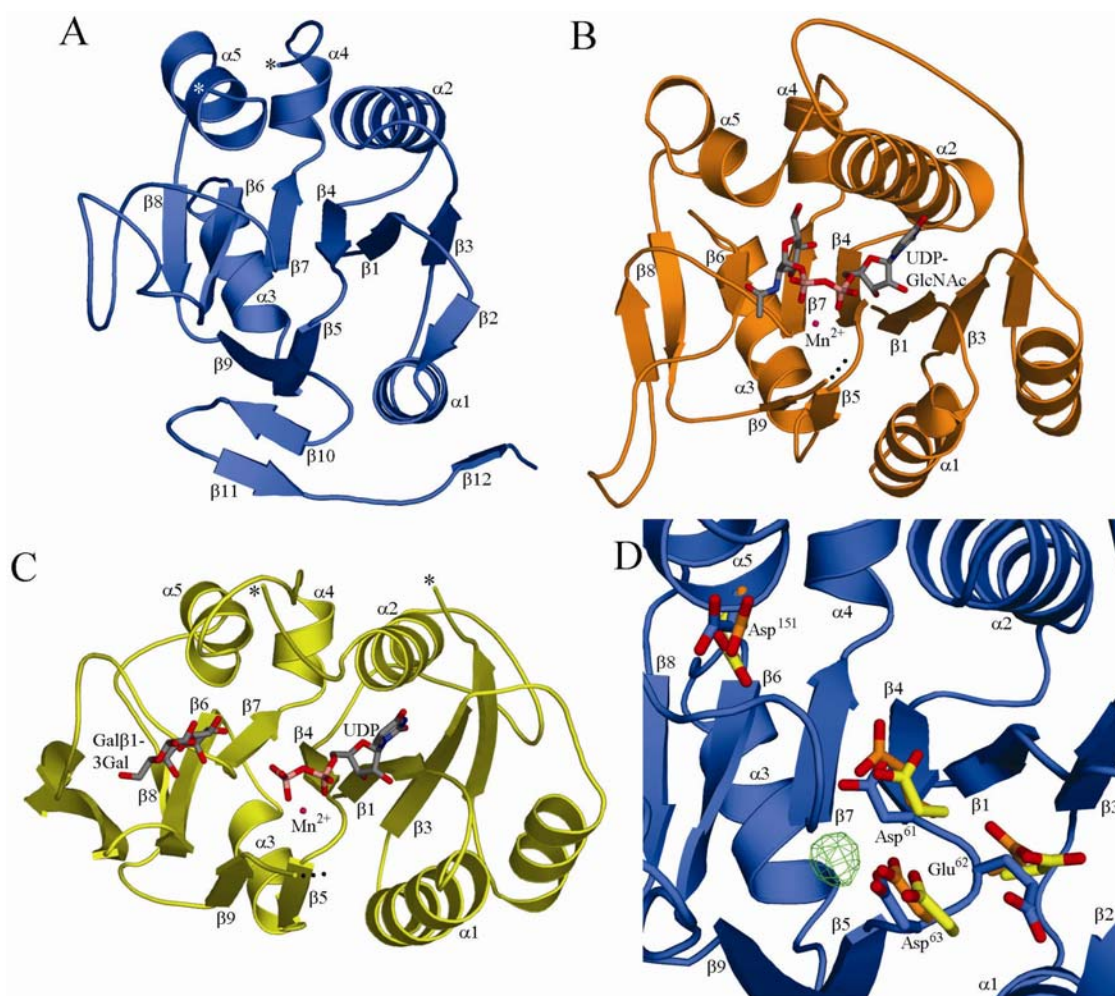


Figure 3-3. A197 and the Equivalent Domains of Structural Neighbors.

Orientations are the same in all panels and are identical to that in Fig. 3-2A. **(A)** Residues 1 to 197 of A197. **(B)** Residues 106 to 317 of rabbit GnT I (PDB ID 1FOA) (136). Secondary structural elements shared with A197 are labeled accordingly. The donor substrate (UDP-GlcNAc) is depicted as sticks and the coordinated divalent cation (Mn^{2+}) as a pink sphere in the active site. The C-terminal end of the structure is marked with dots to indicate the continuation of the polypeptide into an additional domain not present in A197. **(C)** Residues 75 to 310 of human GlcAT-I (PDB ID 1FGG) (97). Labels are as in panels A and B. UDP in the donor-binding site and Gal β 1-3Gal in the acceptor-binding site are shown as sticks and the coordinated Mn^{2+} as a pink sphere. Dots indicate the continuation of the polypeptide. **(D)** Superposition of key GT-A glycosyltransferase active-site features of A197 (blue) with GnT I (orange) and GlcAT-I (yellow). The DXD motif (Asp⁶¹, Glu⁶², and Asp⁶³) and catalytic base (Asp¹⁵¹) of A197 are shown as sticks, as are the equivalent residues in GnT I (Glu²¹¹, Asp²¹², Asp²¹³, and Asp²⁹¹) and GlcAT-I (Asp¹⁹⁴, Asp¹⁹⁵, Asp¹⁹⁶, and Glu²⁸¹). For clarity, only the ribbon diagram of A197 is shown. The green mesh depicts difference electron density (contoured at 6σ) for the Mn^{2+} -binding site. It is adjacent to the DXD motif and coincides nicely with expected metal binding site. In GnT I and GlcAT-I, the DXD motif is involved in coordination of the diphosphate moiety of the donor substrate through the intervening Mn^{2+} ion, while the catalytic base is appropriately positioned for proton extraction from the hydroxyl moiety of the acceptor sugar substrate. Superpositions were prepared with Isqkab (61). Images were prepared and rendered with PyMol (35).

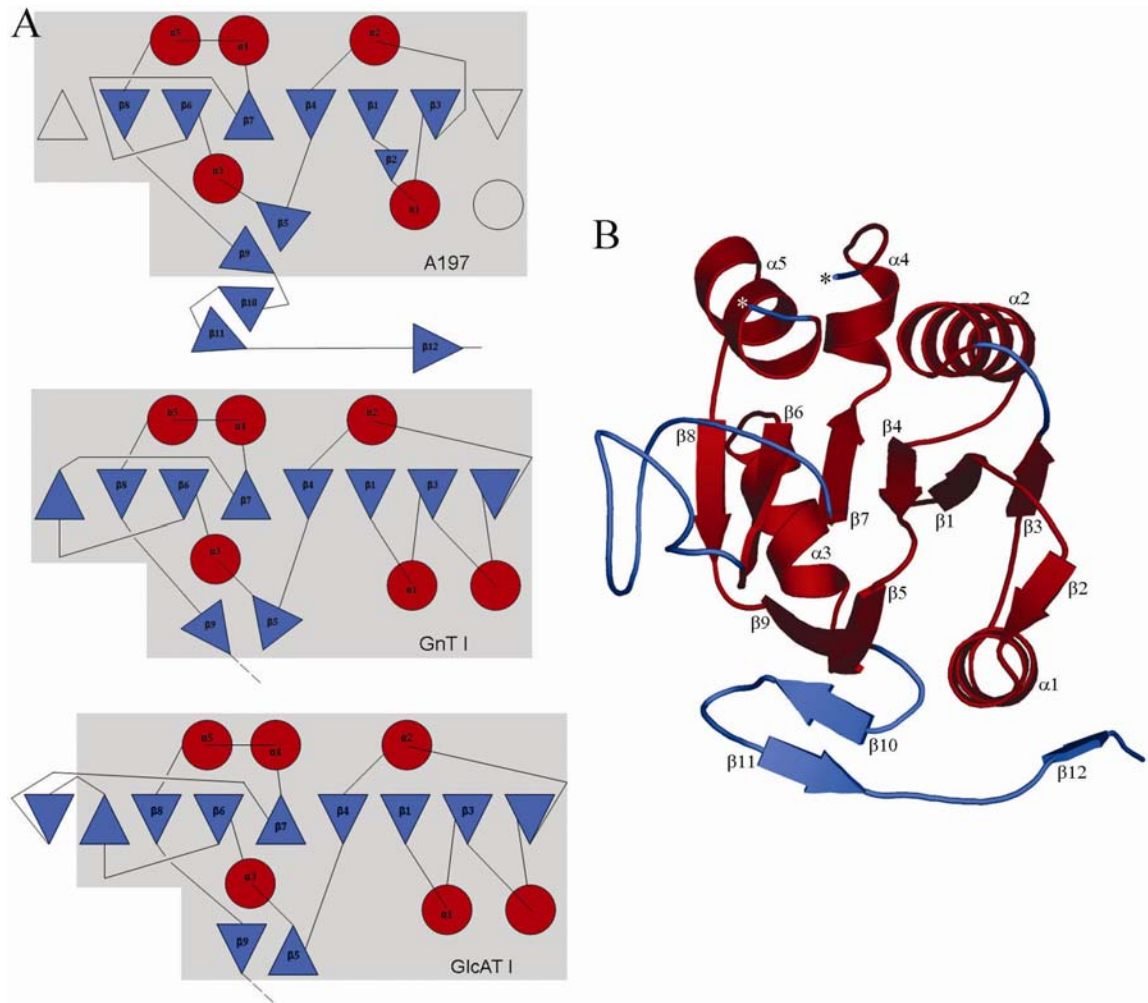


Figure 3-4. A197 and the GT-A Structural Core.

(A) Secondary structural maps of A197 and its nearest structural neighbors. β-strands are depicted as blue triangles and α-helices as red circles. Labels are as in Figures 3-2 and 3-3. Secondary structural elements that are absent in A197, but present in the canonical GT-A fold, are depicted as empty triangles or circles. The core GT-A fold is marked with a gray box. Elements outside the gray box are additions to the core. Dashes at the end of the maps in GnT I and GlcAT I indicate the continuance of the polypeptide into an additional domain that is not part of the GT-A fold. **(B)** Structure of A197 colored to highlight the conserved core. The regions of the structure colored red are highly conserved, while those in blue diverge.

(Figure 3-3D) and the resulting structure-based sequence alignment (Figure 3-5) show that this motif aligns very well, suggesting that the DXD motif in A197 will likewise interact with a Mn^{2+} ion. Accordingly, we see very strong difference density (a 6 σ peak)

at this position in a low-resolution (3.0 Å) difference map for crystals that were soaked with Mn^{2+} (Figure 3-3D).

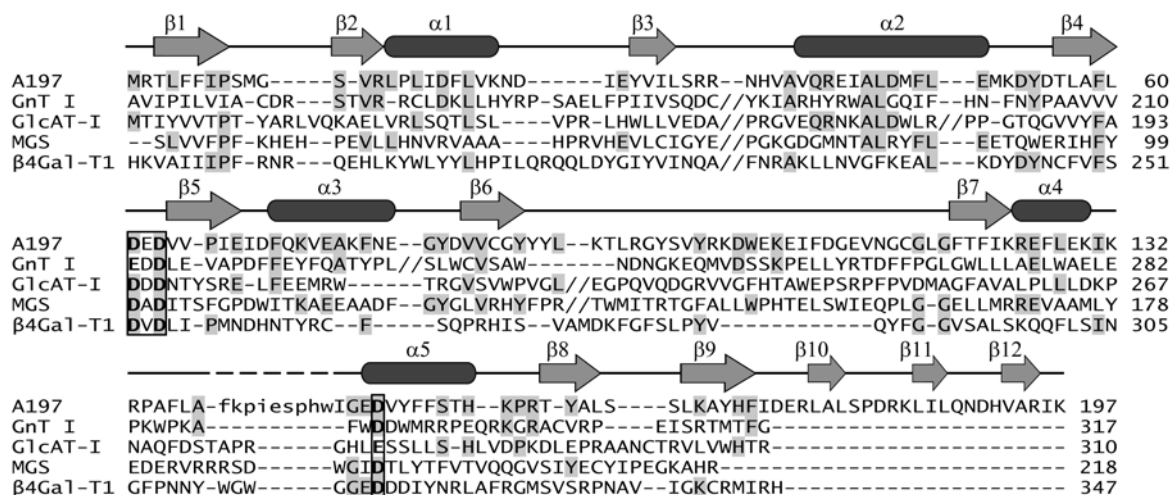


Figure 3-5. Structure-based Sequence Alignment of A197 Structural Neighbors.

Structure-based sequence alignment of A197 (PDB ID 2C0N) with GT-A domains of close structural neighbors. GnT I, residues 106 to 317 of rabbit N-acetylglucosaminyltransferase I (PDB ID 1FOA; family GT13; inverting mechanism) (136); GlcAT-I, residues 75 to 310 of human β 1,3-glucuronyltransferase I (PDB ID 1FGG; family GT43; inverting mechanism) (97); MGS, residues 2 to 99 of *Rhodobacter marinus* mannosylglycerate synthase (PDB ID 2BO4; family GT78; retaining mechanism) (41); β 4Gal-T1, residues 180 to 347 of bovine β 1,4-galactosyltransferase T1 (PDB ID 1FGX; family GT7; inverting mechanism) (44). The secondary structural elements of A197 are mapped above the alignment (arrows denote β -strands, and rounded rectangles denote α -helices). Long stretches of sequence that are not structurally equivalent to A197 have been removed from the alignment for conciseness and are marked by double slashes. The missing residues of the A197 α 4- α 5 loop are in lowercase in the alignment and are marked with dashes in the secondary structure map. Identical residues are highlighted in gray, and the DXD motif and catalytic base are in boldface and are boxed. The structure-based sequence alignment was created with the 3DCoffee@igs web server (<http://www.igs.cnrs-mrs.fr/Tcoffee/>) (100), using Msap_pair to compute the library and slight manual adjustment around the gaps with consideration of the pairwise structural alignments created by lsqman (66).

In addition to the DXD motif, glycosyltransferase enzymes exhibiting the GT-A fold commonly possess a catalytic base, Glu or Asp, that is thought to deprotonate a hydroxyl group on the acceptor substrate, thus assisting nucleophilic attack on the donor sugar (97, 136). The structural superposition (Figure 3-3D) and resulting structure-based

sequence alignment (Figure 3-5) identify Asp¹⁵¹ at the N-terminus of helix $\alpha 5$ as the putative catalytic base in A197. Thus, active-site features that are hallmarks of glycosyltransferase activity, the DXD motif and a putative catalytic base, are preserved within the proper structural context of the GT-A fold, strongly suggesting that A197 is indeed a glycosyltransferase.

As apparent from the structure-based sequence alignment (Figure 3-5), A197 has several other residues that are identical, or at least strongly conserved, within the structural context of the GT-A fold in addition to the DXD motif and the catalytic base. Many of these residues are known to be involved in various aspects of the specific enzymatic reaction, such as substrate specificity or stabilization. However, these similarities are not strong enough evidence to definitively predict the specificity, or possibly even the catalytic mechanism, of A197 because of the diversity seen within specificities of the structural neighbors in particular and within the entire glycosyltransferase class of enzymes in general.

Donor Substrate-binding Site

The structure of GnT I (136) was solved as a complex with its donor substrate, UDP-N-acetylglucosamine, and the coordinated Mn^{2+} (Figure 3-3B). Thus, superposition of GnT I and the donor substrate upon A197 suggests probable binding sites for an activated sugar substrate and a metal cofactor in the active site of A197 (Figures 2-3 and 2-6). The donor substrate is predicted to sit within the extended active-site pocket that runs across the surface of A197, with the uridine moiety situated over the C-terminal ends of $\beta 1$ and $\beta 3$, adjacent to the N-terminal end of $\alpha 2$. The diphosphate

moiety is predicted to span the C-terminus of $\beta 4$ and the N-terminus of $\beta 7$, with the sugar moiety situated over the N-terminus of $\beta 7$, adjacent to helices $\alpha 2$ and $\alpha 5$. The superimposed Mn^{2+} ion is positioned between the $\beta 4$ - $\beta 5$ loop and the phosphodiester bond of the UDP, a position that corresponds nicely with the major peak seen in the low-resolution difference maps from Mn^{2+} -soaked crystals of A197 (Figures 3-3D and 3-6).

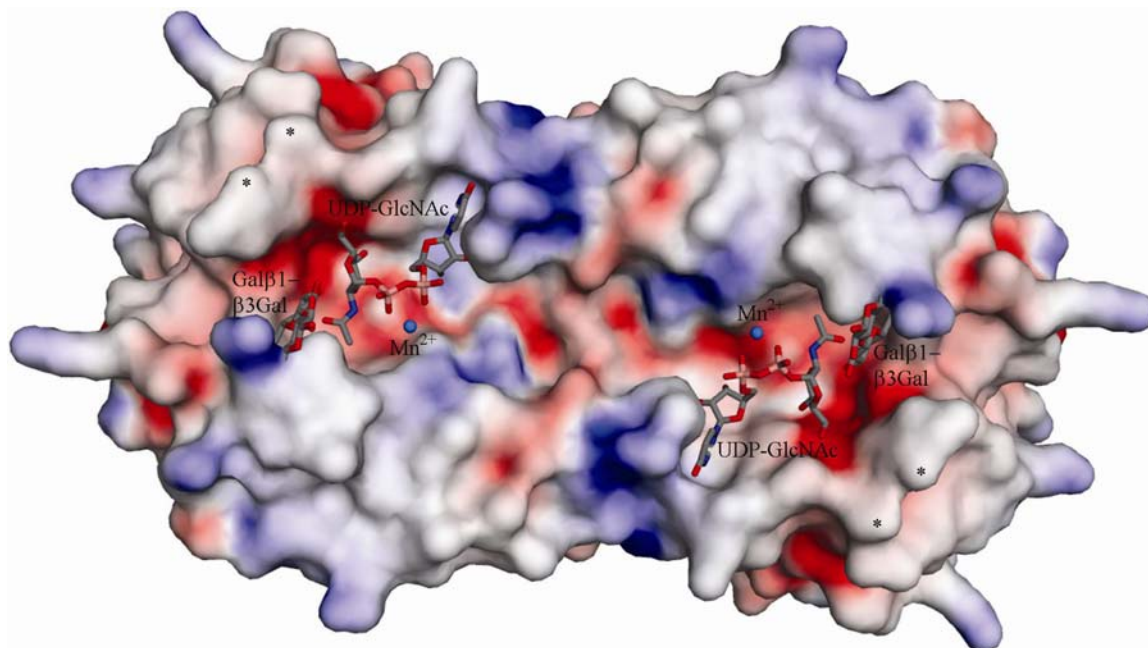


Figure 3-6. Electrostatic Surface of A197 Homodimer with Superposed Substrates.

The orientation is identical to that in Figure 3-2D, serving to highlight the putative active site. The termini of the disordered region within the $\alpha 4$ - $\alpha 5$ loop are marked by *. The putative donor substrate-binding sites are marked with stick representations of the superpositioned donor substrate of GnT I, UDP-N-acetylglucosamine. The putative acceptor substrate-binding sites are marked with stick representations of the superpositioned acceptor substrate of GlcAT I, Gal β 1-3Gal. The electrostatic potential was mapped to the surface of A197 with SPOCK (29), using a probe radius of 1.4 Å, a temperature of 353°K, an ionic strength of 0.15 M, and protein and solvent dielectric constants of 4 and 80, respectively. The color ramp of the surface is from -15 kT/e (red, acidic) to +15 kT/e (blue, basic). Superpositions were prepared with lsqkab (61). Images were prepared and rendered with PyMol (35).

Interestingly, there is a loop adjacent to the active site in GnT I that remains disordered until the donor substrate is bound. Closure of the loop over the donor substrate not only serves to hold it in place but also serves an important role in the binding of the acceptor substrate, providing a structural explanation for the ordered substrate binding seen in GnT I (136). Similarly, nine residues of the loop connecting helices $\alpha 4$ and $\alpha 5$ of A197 are disordered in the absence of substrate. However, relative to GnT I, this disordered loop is situated on the opposite side of the substrate-binding pocket. Nevertheless, its position would allow it to fulfill a similar role in closing over the substrate, possibly serving to complete the formation of an acceptor-binding site.

Disordered or highly flexible regions of a protein structure are often important for the function of that protein. Interestingly, PONDR (75, 114, 115), a program that predicts protein disorder from the linear sequence, does not predict disorder in the $\alpha 4/\alpha 5$ loop. In fact, it predicts that the entire protein is likely to be ordered (Figure 3-7A). The crystal structure, however, represents a population average of all the molecules that make up the crystal and actual information about conformational flexibility can be obtained as a result. Electron density will be absent from regions that are highly flexible or disordered and elevated B factors (temperature factors) will be apparent where the atoms in the protein structure have greater conformational freedom or flexibility. Electron density is absent for residues 139-147 of the $\alpha 4/\alpha 5$ loop and the B factors are greatest in this general vicinity of the protein structure (Figure 3-7B) suggesting this region is highly flexible. By analogy to the closest structural homolog, flexibility around the $\alpha 4/\alpha 5$ loop may be important to allow the loop to adopt an ordered conformation upon substrate binding, as described above for GnT I.

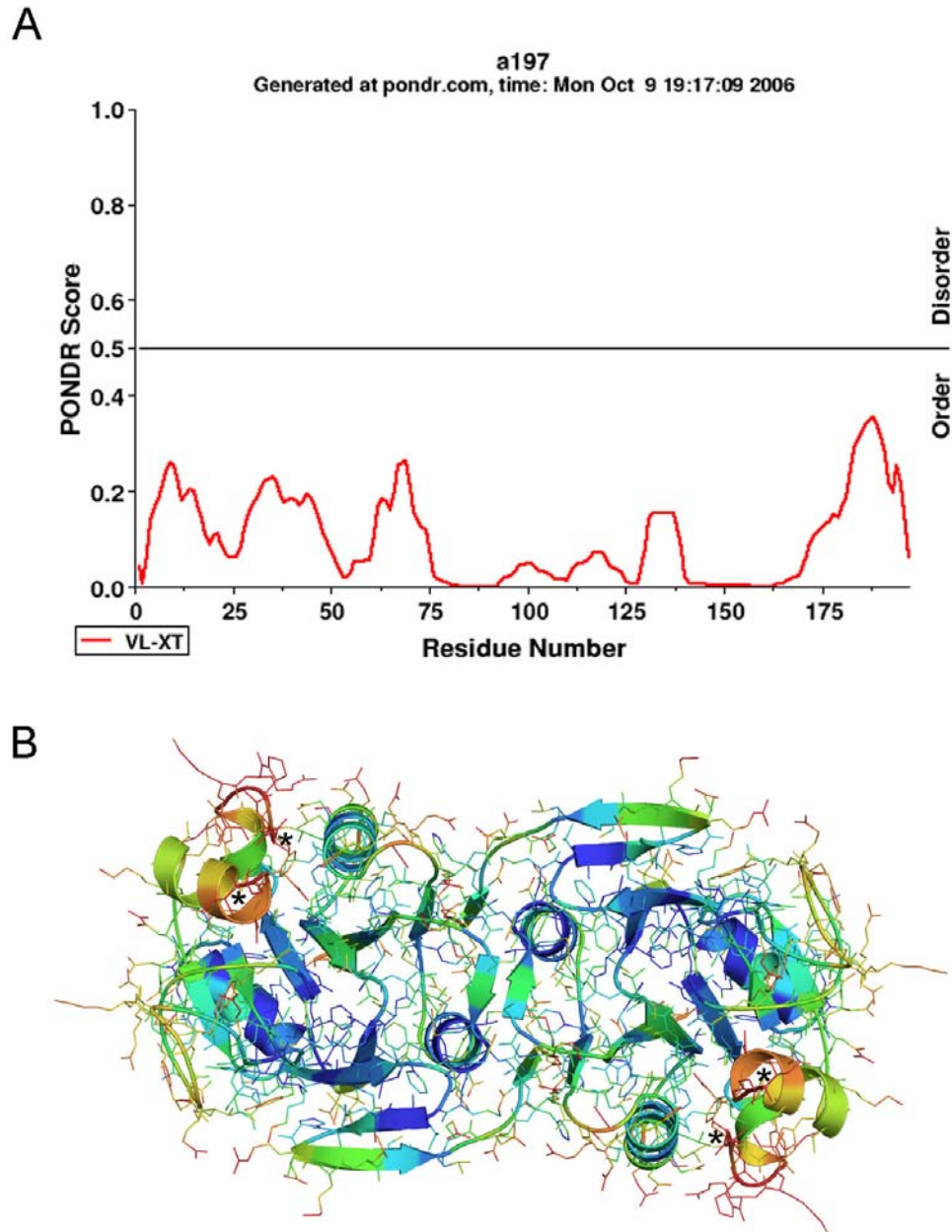


Figure 3-7. Predicted Disorder and Flexibility of A197.

(A) From the primary sequence, the VL-XT algorithm on the POND R web server (75, 114, 115) does not predict any disordered regions of A197. **(B)** The A197 dimer is colored by the B factor [from low (blue) to high (red)] to show regions of the protein that likely have increased flexibility. The mainchain is a cartoon and all atoms are shown as lines, to highlight the flexibility of the sidechains. Helices $\alpha 4$ and $\alpha 5$ and the intervening loop, residues 139 to 147 of which the electron density is absent (labeled with *), have the highest B factors compared to the rest of the model. This suggests that this region of the protein has increased flexibility, which may be important for the function. The closest structural neighbor of A197, GnT I, has a loop that is ordered upon donor substrate binding, holding it in place, and completing the binding pocket for the acceptor substrate. An analogous ordering of the $\alpha 4/\alpha 5$ loop of A197 is easily imagined and flexibility around the loop is likely important for facilitating this movement of the protein chain.

Acceptor Substrate-binding Site

The DALI search also identified human β 1,3-glucuronyltransferase I (GlcAT-I) (96, 97), an inverting glycosyltransferase in CAZy family GT43, as a close structural homologue (3.1-Å RMSD and 13% identity over 148 structurally equivalent residues) (Figure 3-3C). Once again, a refined superposition of GlcAT-I onto A197 with lsqman identifies the same highly conserved structural core (120 residues, 15% sequence identity, 1.83-Å RMSD) (Figure 3-4). Interestingly, while the structure of the C-terminal extension in A197 differs significantly from that seen in GlcAT-I, both proteins utilize the C-terminal tail in the formation of an extensive dimer interface. However, unlike A197, the active-site clefts of each monomer do not merge across the dimer interface.

Importantly, the GlcAT-I structure (97) was solved in complex with UDP, Mn^{2+} , and an acceptor substrate analog, Gal β 1-3Gal β 1-4Xyl (Figure 3-3C). The superposition of GlcAT-I on A197 suggests that the acceptor substrate for A197 will bind adjacent to strand β 6, in a groove formed predominantly by the long intervening region between β 6 and β 7 (Figure 3-6). The acceptor substrate will then be positioned adjacent to the N-terminus of α 5, an optimal arrangement relative to the catalytic base and the expected position of the donor substrate. A second structure of GlcAT-I has been solved with UDP-glucuronic acid, the full donor substrate, in the active site (96). Not surprisingly, the glucuronic acid moiety binds in a manner similar to that seen for the N-acetylglucosamine in GnT I, further indicating that the relative positions of the donor and acceptor substrates in A197 are reliably indicated by superposition of its closest structural homologues.

Discussion

Like the majority of the STIV genome, the predicted amino acid sequence of A197 provides little insight into its role in the viral life cycle. Thus, we initiated crystallographic studies of A197 with the specific goal of identifying structural homologues that might assist in making a functional assignment. The resulting structure does indeed identify close homologues; it is remarkably similar to several glycosyltransferases exhibiting the GT-A fold.

Due to the vast number of potential substrates, in both donor and acceptor, limited sequence similarity and three-dimensional structural information are generally insufficient to allow reliable prediction of substrates for a given glycosyltransferase. In the absence of biochemical characterization with the known substrates, it is also difficult to predict whether the enzyme will catalyze sugar-transfer with inversion or retention of stereochemistry at the anomeric carbon. For example, all of the structural homologues aligned in Figure 3-5 utilize different substrates, and three of the four are inverting enzymes (GnT I, GlcAT-I, and β 4Gal-T1), while mannosylglycerate synthase (MGS) is retaining. Yet the GT-A fold, the DXD motif, and the catalytic base are common to all. Commonality among the A197 structural neighbors is limited to displaying the GT-A fold and the use of a nucleotide diphosphate-activated sugar as the donor substrate. MGS illustrates the difficulty in assigning specific properties to a glycosyltransferase in the absence of biochemical characterization. Sequence-based assignment had placed it in inverting family GT2; however, subsequent characterization showed that it utilizes a retaining mechanism; hence MGS was reclassified as the founding member of family

GT78 (41). Thus, in the case of A197, the identity of donor and acceptor substrates and elucidation of mechanistic details await biochemical and genetic studies.

A197 is one of the smallest glycosyltransferases that we are aware of, composed of only the core catalytic GT-A fold and lacking additional functional domains. Thus, its structure may define the minimal components necessary for glycosyltransferase activity. The presence of minimal, “stripped-down” structural motifs is common in viruses (82), where the need for genetic efficiency within the viral genome is one common explanation for their occurrence. A197 has yet to be assigned to a GT family in the CAZy database. However, the lack of sequence similarity to other known glycosyltransferases suggests that A197 represents yet another sequence-based glycosyltransferase family. Prior to the publication of this work, A197 was absent from the CAZy database altogether; it was not recognized as a glycosyltransferase from its sequence. Recently, however, it was added as a non-classified glycosyltransferase (<http://afmb.cnrs-mrs.fr/CAZY/fam/GT0.html>). Many glycosyltransferases in this category (151 as of August 2006) display weak similarity to established GT families, but too distant to allow a reliable assignment. Some of these will serve as seeds to build new families in the future (31). Of all the non-classified glycosyltransferases, A197 is the only one to have its structure determined.

The CAZy database includes a number of viral glycosyltransferases, including annotated glycosyltransferases from other crenarchaeal viruses AFV1 (18), AFV2 (51), ARV1 (137), SIFV (8), SIRV1 (98, 99), SIRV2 (98), and STSV1 (143). These crenarchaeal viral glycosyltransferases have been placed in family GT4, a family that adopts the GT-B fold and utilizes the retaining mechanism of sugar transfer (31). We note that GT4 also includes a glycosyltransferase from PBCV-1 (48, 82), a eukaryotic

virus that is believed to be evolutionarily related to STIV . Thus, glycosyltransferases are not uncommon to viral genomes and are frequently found in viruses infecting the Crenarchaea but none of these putative crenarchaeal viral glycosyltransferases is related to A197; they are much larger and are likely to display the GT-B fold.

Viruses commonly decorate their proteins with sugars as a means of regulating interactions with their hosts. While viruses can accomplish this task by utilizing the host's glycosylation machinery, it is apparent that many viruses also encode their own proteins for specialized glycosylation needs (82). Virally encoded glycosyltransferases are known to serve a variety of functions and target a wide variety of acceptor substrates, including DNA, proteins, and small molecules. For example, some lytic bacteriophages glycosylate their DNA to protect it from host restriction enzymes while others alter the glycosylation of cell surface antigens to induce serotype conversion of the host bacteria during lysogeny or, in the case of lytic phage, to induce receptor conversion, preventing retention of progeny on the host cell debris (82).

In eukaryotic systems, many baculoviruses utilize virally encoded glycosyltransferases to glycosylate host ecdysteroids. This modification alters the biological activity of the hormone, suppressing development of the infected host. Glycosyltransferase-encoding genes are also common in the phycodnavirus family. This includes PBCV-1, where these virally encoded enzymes have been implicated in glycosylation of its major capsid protein, Vp54. While eukaryotic viral proteins are typically glycosylated by host-encoded enzymes located in the endoplasmic reticulum (ER) and Golgi apparatus, Vp54 is apparently glycosylated in an ER and Golgi apparatus-independent manner by virally encoded glycosyltransferases that are expressed

in the cytosol. Thus, it has been suggested that glycosylation of Vp54 reflects an ancestral pathway that existed prior to ER and Golgi apparatus formation (48, 82).

This is particularly interesting in light of the putative evolutionary relationship between PBCV-1 and STIV. The similarities between these viruses are not limited to their overall capsid architecture and the structures of their major capsid proteins. It has been shown through proteomic studies of purified virus that, like Vp54 in PBCV-1, the STIV major capsid protein (MCP) is glycosylated (Figure 3-8A and B) (78). Further, it was noted that the potential glycosylation sites on MCP, predicted using standard rules, correspond closely with the known glycosylation sites of Vp54 (Figure 3-8C) (78).

While A197 might serve any number of functional roles, the similarities to PBCV-1 with respect to capsid architecture, fold of the major capsid protein, glycosylation of the major capsid protein, and the occurrence of a virally encoded glycosyltransferase suggest that A197 participates in glycosylation of the major capsid protein in STIV and is likely involved in particle maturation.

It has further been observed that the STIV major capsid protein gives a horizontal spot train on two-dimensional gels (Figure 3-8B) (78), indicating a series of increasingly acidic species compared to the bacterially expressed protein, which is not glycosylated. This suggests that glycosylation introduces negative charge through addition of acidic sugar moieties such as galacturonic acid, glucuronic acid, and iduronic acid (38). Alternatively, the spot train might represent sulfation of neutral saccharides (38). Regardless of the nature of the attached sugar, MCP is glycosylated through posttranslational modification in its natural host but not when heterologously expressed in the bacteria, which may suggest the need for the virally encoded glycosyltransferase.

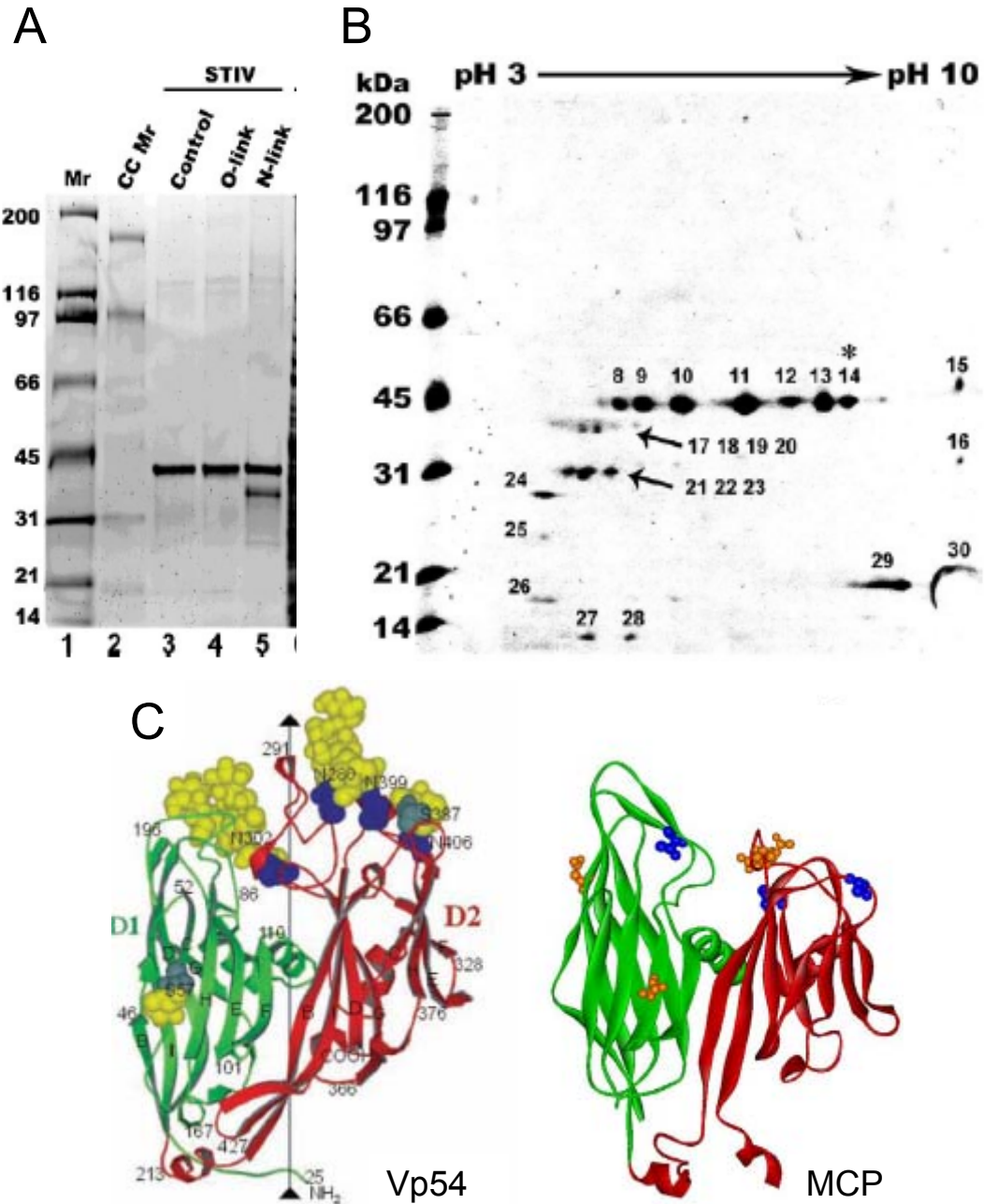


Figure 3-8. Glycosylation of the Major Capsid Protein (MCP) of STIV.

(A) Purified STIV was tested for glycosylation and the MCP stains positive. **(B)** 2-D gel analysis of purified STIV displays a horizontal spot train for the MCP (spots 8-14) showing increasingly acidic species as compared to bacterially expressed, non-glycosylated protein (which runs similarly to spot 14). **(C)** Predicted glycosylation sites (shown as spheres) of MCP from STIV (right) correspond nicely with known glycosylation sites of its structural homologue Vp54 from PBCV-1 (left). Figure was adapted from various figures in (78).

In addition to the obvious structural and functional roles for protein glycosylation, increased glycosylation has been shown to correlate with environmental extremes of temperature, pH, and salt concentration (1, 38, 53, 84, 138, 144). Thus, it has been suggested that increased protein glycosylation may stabilize extracellular proteins in these harsh environments (1, 38, 53, 84, 138, 144). In this light, it is not surprising that many crenarchaeal viruses apparently encode glycosyltransferases.

Glycosyltransferases are an enormous class of enzymes that display extreme diversity with respect to sequence and substrate specificity. Largely due to convergence of tertiary structure, however, they are thought to have evolved from a common ancestor, particularly those of the GT-A superfamily (43). As glycosyltransferases are ubiquitous to the three domains of life, they probably predate the evolutionary events that established these domains. Likewise, there is accumulating evidence that STIV is part of a viral lineage that predates the split of the three domains. Identification of a glycosyltransferase in STIV extends the observed similarities among double-stranded DNA viruses inhabiting the three domains of life beyond those of capsid architecture and the conserved fold of the major capsid protein. The last virus common to the STIV/PBCV-1 lineage is likely to have carried a glycosyltransferase gene in its genome. With the development of immunity, however, descendant viruses infecting multicellular eukaryotic organisms may have found it advantageous to lose this feature, relying instead on enzymes supplied by their hosts.

In summary, the structure of A197 provides substantial insight into its function and its relevance to the STIV life cycle. It is yet another demonstration that structure can reveal functional and evolutionary relationships.

Conclusions

The work presented in this chapter on the structural characterization of A197 from STIV provides an excellent example of how a protein's three-dimensional structure can provide insight into its function. The sequence of A197 is quite unique with the highest scoring hit from a PSI-BLAST search being the major merozoite surface protein precursor from the Eukaryotic parasite *Plasmodium falciparum* with a fairly insignificant E value of 0.74. Only after the structure was solved did it become apparent that A197 is a glycosyltransferase. Knowing this and going back to the PSI-BLAST results, we find that the most similar glycosyltransferase is a putative enzyme from *Sinorhizobium meliloti* with a highly insignificant E-value of 7.6. Similarly, going back to CD searches (80), we find glycosyltransferases with insignificant E-values of 6.1 and 17. Not only does this example show the value in using protein structure in the annotation of its function, it also points toward the danger of blindly annotating unknown genes based on very weak sequence or motif similarity.

A197 was recently added to the CAZy database (30, 31) as an unclassified glycosyltransferase, meaning its sequence is not similar enough to any other glycosyltransferases to warrant it being placed in an established GT family, and a family with one member does not make much sense. With the rapidly increasing number of completely sequenced genomes, glycosyltransferase genes are being discovered at a phenomenal rate, as evidenced by the ever-increasing number of genes in the CAZy database and the growing number of GT families. It is likely only a matter of time before

A197-like glycosyltransferases are discovered that either establish a link between it and a current family or necessitate the creation of a yet another GT family.

The work on A197 is a significant advance in the field of crenarchaeal virology because it is the first glycosyltransferase from a virus with an archaeal host to be structurally characterized, or to be studied in any detail at all for that matter. As mentioned above, putative glycosyltransferases have been annotated in the genomes of several other crenarchaeal viruses based on sequence comparisons but these are all different from A197 in that they are much larger and are expected to display the GT-B fold. A197 is the smallest glycosyltransferase that we are aware of, composed of only the core GT-A fold. As such, it may represent the minimal catalytic subunit necessary for glycosyltransferase activity and may thus provide a model system for this class of important enzymes. In this regard, the presumed evolutionary constraints placed on the overall structure of glycosyltransferases displaying the GT-A fold and the branching of the Crenarchaea near the “root” of the tree of life is particularly noteworthy.

In addition to providing insight into the function, the structure of A197 has allowed us to make specific suggestions about its role in the STIV lifecycle. Proteomics studies on purified virus from Brian Bothner’s lab coincided with this structural determination and revealed that the MCP is glycosylated. We expect that A197 is, at least in part, responsible for the observed glycosylation and thus plays a role in the maturation of the STIV particle. It is interesting that the bacterially expressed MCP lacks glycosylation. This may be because *E. coli* lacks the necessary machinery required and it would be worthwhile to see if co-expression of MCP and A197 is sufficient for recovering MCP glycosylation. By analogy, the major capsid protein of PBCV-1 is

glycosylated at nonstandard sites necessitating that it code for its own enzyme, which may be one reason why STIV must express its own glycosyltransferases.

Although we have identified A197 as a glycosyltransferase, much remains in its complete characterization that is beyond the scope of this research project. We do not know its substrates nor do we know whether it utilizes the retaining or inverting mechanism of catalysis. As explained in the body of this chapter, there is enormous diversity with respect to substrate specificity for this superfamily of enzymes and mechanistic details cannot be established without characterization of the substrates and products. The nature of the glycosylation on the MCP likewise remains unknown and its characterization can also be problematic. Establishing the details of either of these problems would undoubtedly help to elucidate the other. Furthermore, it will be important to demonstrate when in the viral lifecycle A197 is expressed. It is positioned in the STIV genome in a putative transcript that includes two known capsid components, A223 and C381. If A197 were indeed necessary for glycosylation of the MCP, it would make sense for its expression to coincide with the expression of the capsid components. Polyclonal antibodies against A197 are available and may aid in the detection of the protein in infected cultures.

CHAPTER 4

STRUCTURAL CHARACTERIZATION OF B116; A NEW FOLD IN A HIGHLY
CONSERVED CRENARCHAEAL VIRAL PROTEINIntroduction

B116 is a 116-residue protein in open reading frame (ORF) B of the viral genome (Figure 4-1) whose sequence provides little insight into its function. While the majority of the proteins encoded by STIV lack significant sequence similarity to other proteins, the hypothetical protein encoded by ORF B116 is unique. Though it does not show sequence similarity to any proteins of known function, it is common to the genomes of three additional hyperthermophilic viral families, the *Rudiviridae*, the *Lipothrixviridae*, and the *Bicaudaviridae*. B116 and its viral relatives thus constitute one of the most prominent clusters of orthologous proteins in the crenarchaeal viruses, one that spans three otherwise unrelated viral families (105, 112). Functional information about B116 is also not apparent from a conserved domain (CD) search (80) using a relaxed E-value threshold. The highest scoring hit (E-value 0.3) is to about 65 amino acids of the amidotransferase domain of the very large (1479 residues) alpha subunit of the glutamate synthase complex; and the glutamate synthase catalytic residues are not part of the aligned motif. Further, the protein fold recognition algorithm that uses 1D and 3D sequence profiles coupled with secondary structure and solvation potential information utilized by the protein threading server 3D-pssm (64) fails to find meaningful matches.

The highest scoring hit sounds interesting as it is to an archaeal elongation factor but the E-value is an insignificant 5.26.

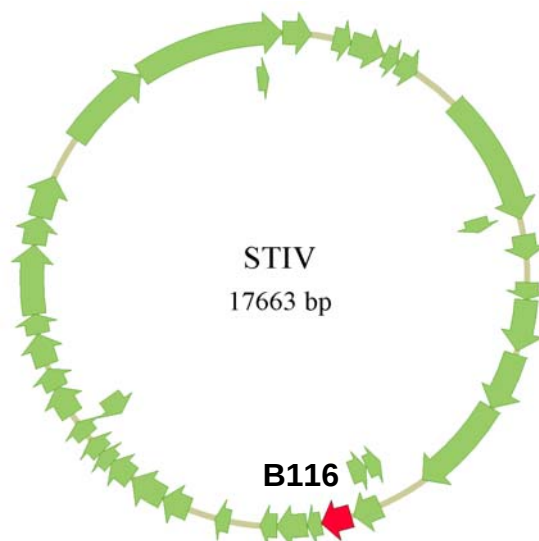


Figure 4-1. Location of the B116 Gene in STIV Genome.

Due to the lack of functional information provided by the sequence and the likely central importance of this protein family to the biology of these fascinating and unique viruses, we have undertaken biochemical and structural studies of STIV B116. Because B116 is broadly conserved across several families of crenarchaeal viruses, we hope that insight provided by its structure may be generally relevant to crenarchaeal viral lifecycles and their virus-host relationship. The protein expresses well in *E. coli* and diffraction quality crystals were produced to facilitate its characterization by x-ray crystallography.

Results

The B116 construct used in this study codes for the 116 amino acids of the native protein plus an additional C-terminal His-tag, for a total of 122 residues with a calculated mass of 14,153 Da. The purified protein elutes from a Superdex 75 size exclusion column as a single peak with an apparent molecular mass between 25 and 30 kDa, suggesting it is present as a homodimer in solution. B116 crystallizes in space group $P2_12_12_1$ with two copies of the protein, a potential homodimer, in the asymmetric unit. The structure was determined at a resolution of 2.4 Å by single-wavelength anomalous diffraction (SAD) using selenomethionyl-incorporated protein. Details on data collection and model refinement are presented in Tables 2-6 and 2-7, respectively, in Chapter 2. The structure has been deposited in the Protein Data Bank (www.pdb.org) under accession code 2J85.

Structure of B116

The B116 polypeptide folds to form a five-stranded, predominantly parallel β -sheet lined on one side by three α -helices (Figure 4-2). The polypeptide traces through β -strand 1 (β 1), β -strand 2 (β 2), α -helix 1 (α 1), β -strand 3 (β 3), α -helix 2 (α 2), β -strand 4 (β 4), α -helix 3 (α 3) and β -strand 5 (β 5). This results in a β -sheet topology of 3, 1, 4, 5, 2, with strand β 5 running antiparallel to the remaining strands. Three of the connections between these secondary elements appear to be structurally important. Extended traverses are required to connect β 1 to β 2, α 2 to β 4, and β 4 to α 3. Interestingly, a disulfide bond is also found; Cys33 and Cys62 contribute their sulfurs to it. This covalent link between the α 1 and α 2 helices is likely to enhance the thermostability of the

B116 fold (13, 73). The significance of this disulfide bond is discussed in more detail in Chapter 5 of this dissertation.

Two copies of the B116 polypeptide are found in the asymmetric unit, giving rise to the homodimer suggested by size exclusion chromatography. An important component of these subunit interactions are the $\beta 2$ strands, which are found along the edge of each subunit (Figure 4-2). These strands come together in an antiparallel fashion to unite the 5-stranded beta sheet of each subunit into a larger 10-stranded β -sheet. Interestingly, the topology of the six central strands ($\beta 4$, $\beta 5$, and $\beta 2$) of this 10-stranded mixed β -sheet resemble that of an unclosed antiparallel β -barrel (Figure 4-2B). The remaining strands and helices flank the unclosed barrel, giving rise to a saddle-shaped protein. A deep cleft, the “seat” of the saddle, spans the dimer interface, while the extended $\beta 4$ - $\alpha 3$ loop rises above to form the “horn” of the saddle. Formation of the dimer occludes 882 Å² (greater than 12%) of the solvent-accessible surface area per monomer, with an interface rich in hydrophobic residues.

Structural comparisons performed using the DALI (54) and VAST (46) servers fail to find structures that superimpose meaningfully upon B116, the highest scoring results are barely above the level of insignificance. For the most part, the identified structures show some similarity to the β -sheet in a single B116 subunit, but the similarities end there. Thus, it seems the structure of B116 is fairly unique. This lack of structural homology is unfortunate; it complicates the effort to formulate a functional hypothesis. However, structural analysis should also include an analysis of the associated surface properties, including the location of significant clefts or pockets, an examination

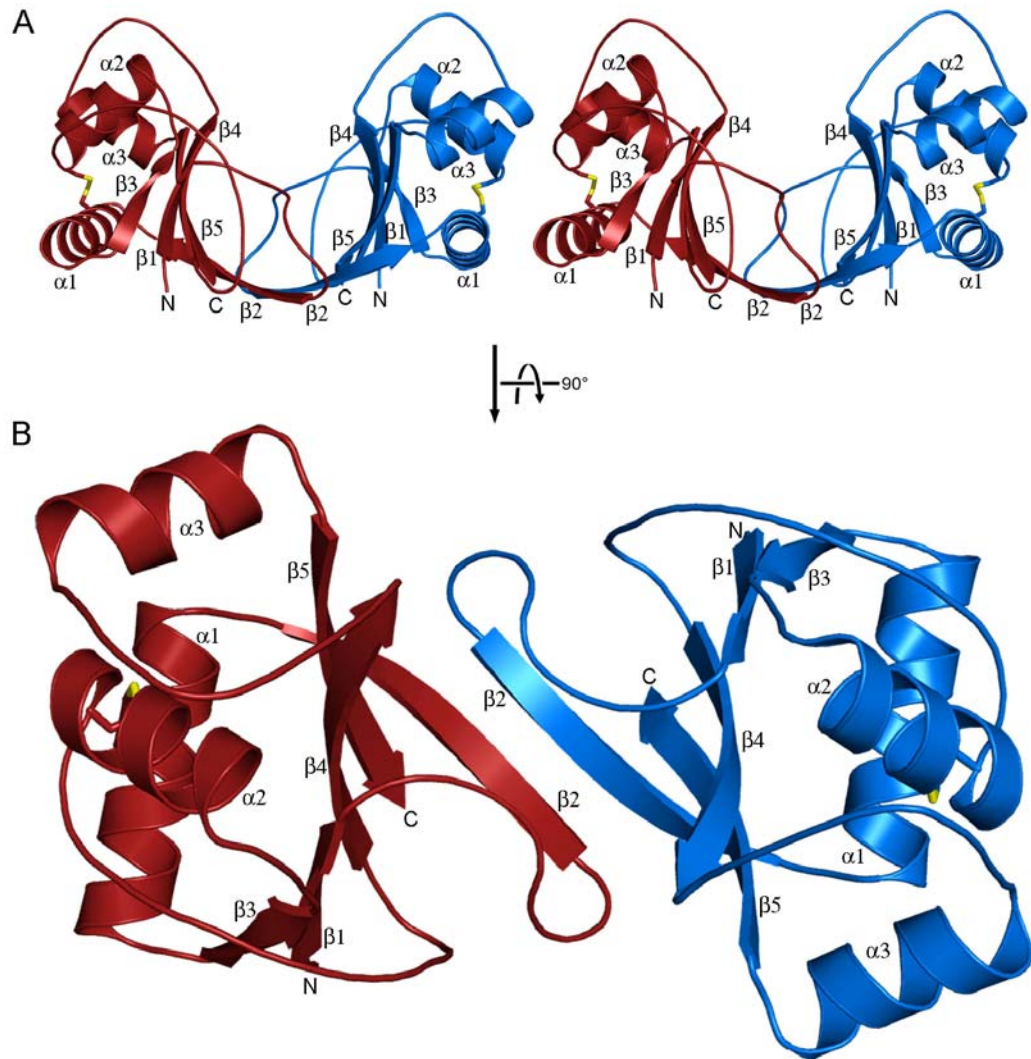


Figure 4-2. Structure of B116.

(A) Stereo image of the B116 homodimer. One monomer is colored red and the second is blue. The ribbon diagram depicts the secondary structural elements of B116, which are labeled in ascending order from N- to C-terminus. The polypeptide traces from the N-terminus through $\beta 1$, $\beta 2$, $\alpha 1$, $\beta 3$, $\alpha 2$, $\beta 4$, $\alpha 3$, and $\beta 5$ of each subunit. The dimer displays a central 10-stranded β -sheet in which the central six strands adopt a topology resembling an unclosed antiparallel β -barrel. The missing staves result in formation of a deep cleft that spans the dimer interface, giving rise to a saddle shaped homodimer. Note that Cys³³ in helix $\alpha 1$ forms a disulfide bond (yellow) with Cys⁶² at the end of helix $\alpha 2$. **(B)** Relative to panel A, the homodimer has been rotated 90° about the depicted horizontal axis. The view is now looking into the deep cleft that spans the dimer interface, the “seat” of the saddle. Labels are as in panel A.

of the conserved residues, and relationships between these features and the electrostatic surface potential. Considered together, these properties can substantially aid development of a functional hypothesis.

Sequence Homology in B116-like Proteins

Considering proteins encoded within the STIV genome, B116 is somewhat unique because it shares clear sequence similarity to proteins from other crenarchaeal viruses (Figure 4-3). In addition to STIV, a PSI-BLAST (4) search identifies B116-like proteins in six other crenarchaeal viruses, spanning three additional viral families. These include SIRV1, SIRV2 and ARV1 (*Rudiviridae*), SIFV and AFV1 (*Lipothrixviridae*), and ATV (*Bicaudaviridae*). Interestingly, homologues are also found in three species of Bacteria; two from the family *Clostridiaceae* [*Alkaliphilus metalliredigenes* (ZP_00799325.1)(A. Copeland *et al.*, unpublished), *Clostridium beijerincki* (ZP_00907466.1) (A. Copeland *et al.*, unpublished)], and one from the family *Bacillaceae* [*Bacillus subtilis* (yddF, CAB12302.1)(70)]. To our knowledge, none of these viral or bacterial proteins has been characterized.

A multiple sequence alignment of these viral and bacterial proteins is shown in Figure 4-3. The alignment shows significant conservation along the entire linear sequence. To get a better idea of the extent of amino acid conservation, and thus a better picture of regions of the protein that are likely to be important for its function, the multiple sequence alignment of B116 and its homologues in other crenarchaeal viruses can be mapped onto the B116 structure (Figure 4-4). An algorithm using an amino acid substitution matrix is used to compute a score that reveals the level of conservation for

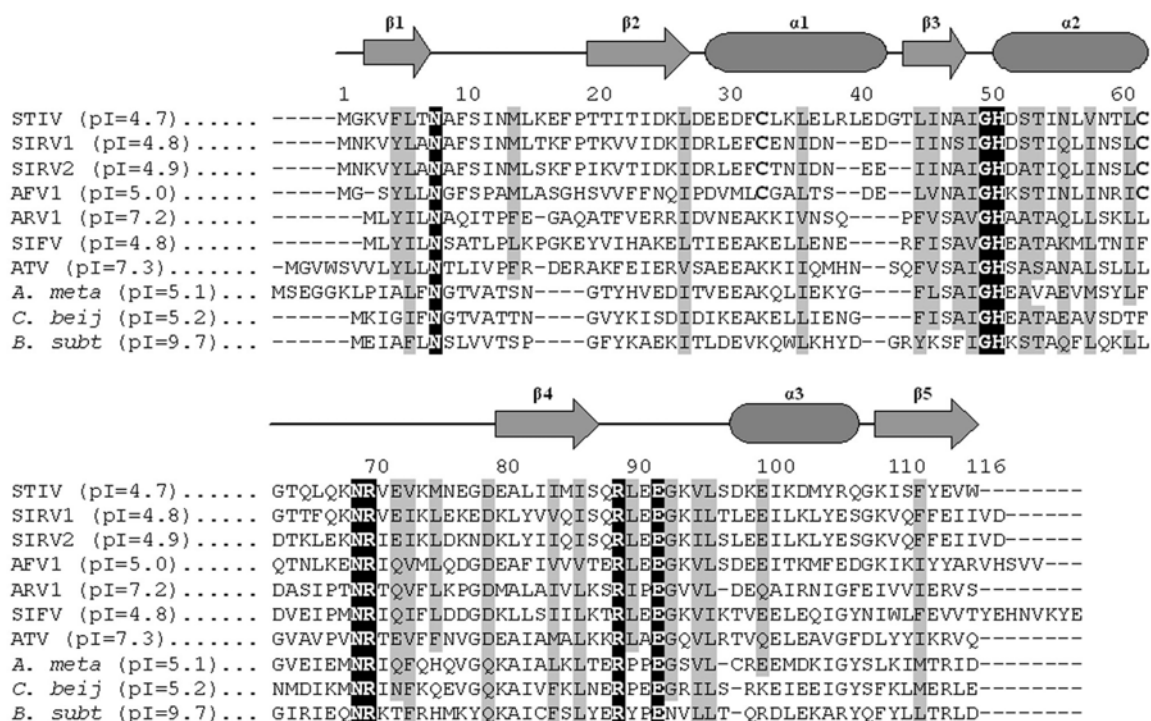


Figure 4-3. Multiple Sequence Alignment of the B116 Family.

B116 from STIV was aligned with orthologs from other crenarchaeal viruses and the bacterial homologs using default settings in ClustalX (131, 132). A few manual adjustments were made around the gaps using the alignment of the viral proteins alone as a guide. Note the absence of gaps in the sequence of B116 itself. Residues that are strictly conserved are shaded in black with white lettering; strongly conserved residues are shaded in gray. The relative positions of the secondary structural elements in B116 are indicated above the alignment, where arrows represent β -strands and rounded rectangles represent α -helices. The numbering corresponds to that of B116. B116-like proteins show sequence conservation throughout their primary structure. Note the occurrence of the cysteine pair (bold type) among the viral orthologs from SIRV1, SIRV2, and AFV1, while the remaining sequences lack both cysteines. The *Rudoviridae* sequences include SIRV1 [accession number NP_666617.1 (98)], SIRV2 [CAC87312.1 (98)] and ARV1 [CAI44168.1 (137)], *Lipothirixviridae* include AFV1 [YP_003762.1 (18)] and SIFV [NP_445691.1 (8)], while ATV is a member of the *Bicaudaviridae* [YP_319837.1 (52)]. The creation of a new viral family for STIV is pending. Surprisingly, homologs are also found in three species of Bacteria; two from the family *Clostridiaceae* [*Alkaliphilus metalliredigenes* (ZP_00799325.1) (A. Copeland *et al.*, unpublished), *Clostridium beijerincki* (ZP_00907466.1) (A. Copeland *et al.*, unpublished)], and one from the family *Bacillaceae* [*Bacillus subtilis* (yddF, CAB12302.1)(70)]. The presence of a B116-like protein in unrelated crenarchaeal viruses and select bacteria makes a strong case for horizontal gene transfer (105).

each amino acid in a multiple sequence alignment and this score is then related to a structural representative (47, 71). When the B116 model is viewed in this manner, it is clear that the surface containing the cleft is much more conserved than the rest of the

protein. It is also apparent that the residues that line the entrance to the cleft make up the most highly conserved region of the entire protein (Figure 4-4).

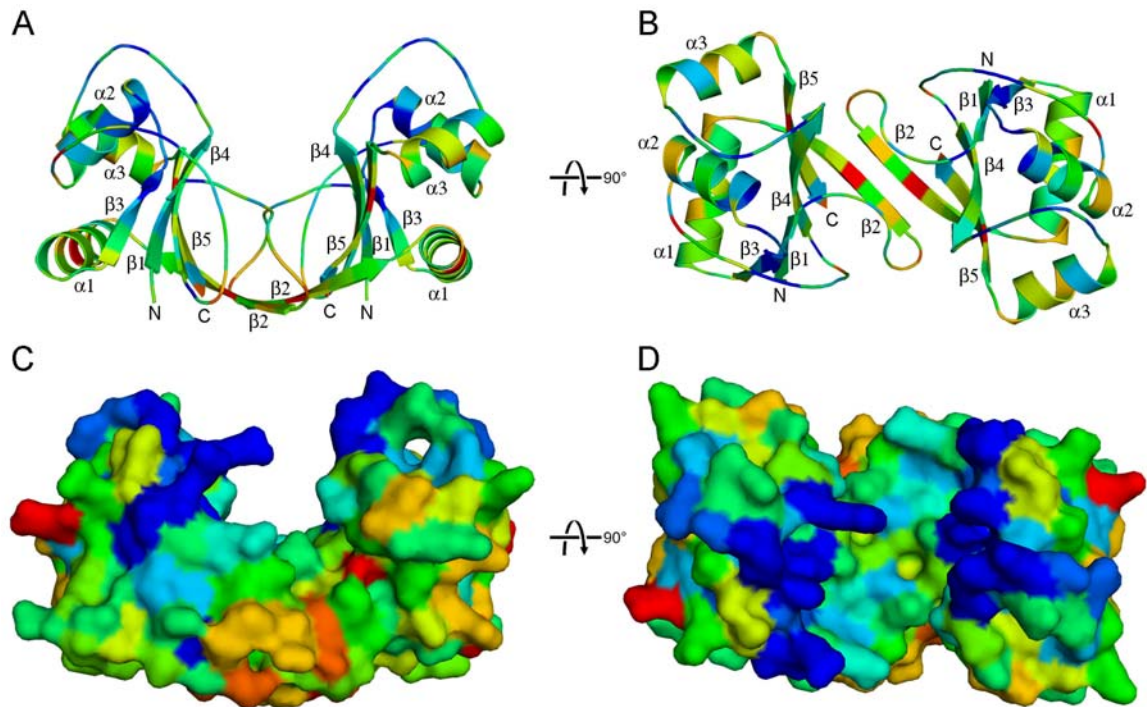


Figure 4-4. Conserved Surface of the Viral B116-like Proteins.

Ribbon diagram (**A and B**) and surface rendering (**C and D**) of B116 structure colored according to amino acid conservation among the crenarchaeal viral B116-like proteins. The color ramps from blue (strictly conserved) to red (variable). Panels B and D are in identical orientations and are rotated 90° about the depicted horizontal axes relative to panels A and C, which are also in identical orientations. Amino acid conservation among a family of proteins suggests regions that are evolutionarily constrained and thus are likely to be important for structure and/or function. In the case of B116, the top of the saddle in general and the “horn” region in particular are the most highly conserved. This surface is likely important for the protein’s function. The ClustalX multiple sequence alignment and B116 coordinate file were submitted to the ConSurf Server (47, 71) (<http://consurf.tau.ac.il>) to compute amino acid conservation scores for each residue using the Bayesian method, which was then substituted for the B factor in the resultant coordinate file. PyMOL (35) and the color_b.py script (<http://adelie.biochem.queensu.ca/~rlc/work/pymol/>) were then used to visualize the modified coordinate file.

Despite this large conserved surface, however, only seven residues are strictly conserved. These residues are likely the most important for the function of the protein.

The regions of greatest conservation are in the $\beta 4$ - $\alpha 3$ loop and the $\beta 3$ - $\alpha 2$ loop and strictly conserved residues are also found in the $\beta 1$ - $\beta 2$ loop and the $\alpha 2$ - $\beta 4$ loop. Thus, all seven of the strictly conserved residues in the B116-like family are found in surface-exposed loops and, with the exception of the conserved glycine residue in the $\beta 3$ - $\alpha 2$ turn, they are unlikely to be of structural importance. Therefore, the strict conservation of Asn⁸, His⁵¹, Asn⁶⁹, Arg⁷⁰, Arg⁸⁹, and Glu⁹² is most likely related to the function of B116. These regions of discontinuous linear sequence come together in the tertiary structure to form the top of the saddle with the $\beta 4$ - $\alpha 3$ loop comprising the horn. Importantly, when only these strictly conserved residues are mapped to the surface of B116, it becomes clear that they form a spatially contiguous surface that runs along the rim of the central cleft, extending down along the walls of the cleft, and upwards to the horn of the saddle (Figure 4-5).

The calculated isoelectric point (pI) for B116 is 4.7. With the exception of the B116-like proteins from ATV and ARV1, most of the other viral proteins show comparable pIs (Figure 4-3). Similarly, two of three bacterial proteins show acidic pIs of approximately 5, while yddF from *B. subtilis* is quite basic. This suggests the presence of a significant amount of negative charge on most of these proteins. However, for B116, projection of the calculated electrostatic field onto the protein surface reveals a significant segregation of the positive and negative charge.

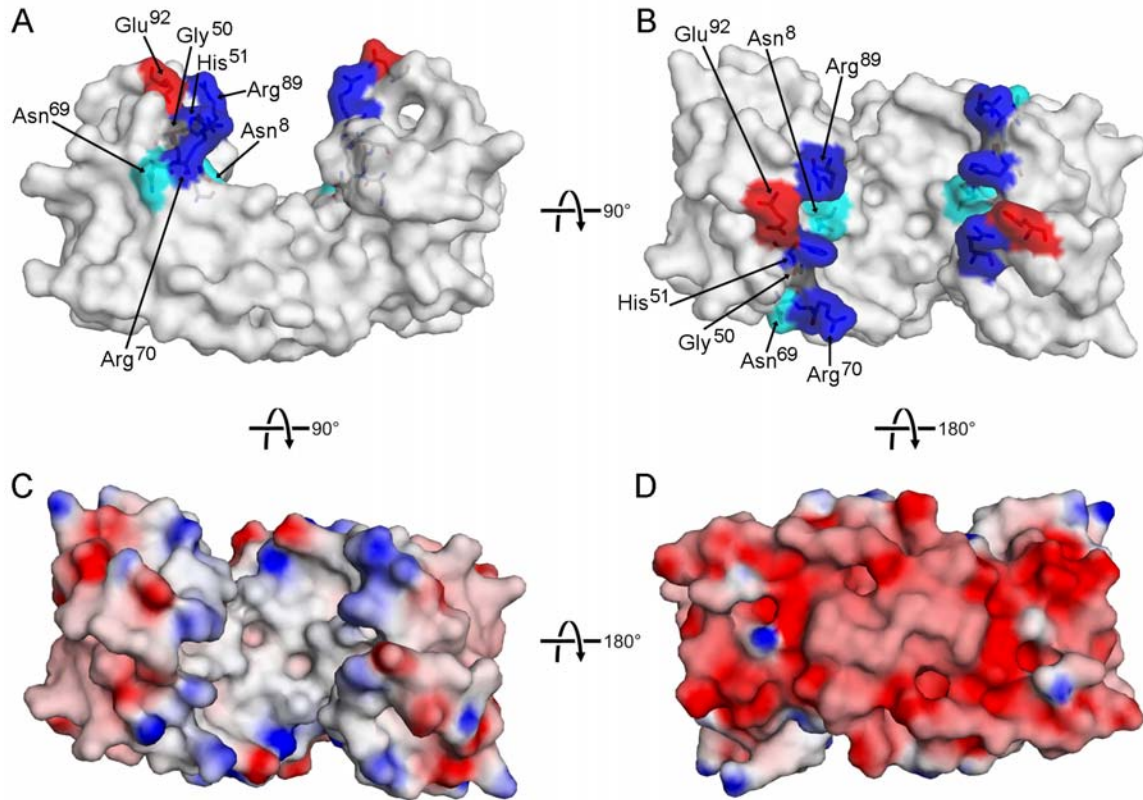


Figure 4-5. Surface Features of the B116 Family.

(A) The surface of the B116 dimer is shown with strictly conserved residues colored according to residue type. Asn⁸ and Asn⁶⁹ are cyan; Gly⁵⁰ is gray; His⁵¹, Arg⁷⁰, and Arg⁸⁹ are blue; Glu⁹² is red. These strictly conserved residues form a patch of contiguous surface on each of the symmetry related subunits. Arg⁸⁹ and Glu⁹² are contributed by the $\beta 4$ - $\alpha 3$ loop that forms the horn of the saddle; His⁵¹ comes from the $\beta 3$ - $\alpha 2$ loop, while Arg⁷⁰ lies in the $\alpha 2$ - $\beta 4$ loop. The basic residues, colored in blue (His⁵¹, Arg⁷⁰ and Arg⁸⁹), lie along the lip and sides of the central cleft. **(B)** The strictly conserved surface features are shown from a different perspective. Relative to panel A, the dimer has been rotated by 90° about the depicted horizontal axis. **(C)** The electrostatic potential is mapped to the surface of B116 dimer; the perspective is identical to that in panel B. The color ramp on the surface is from -15 kT/e (red, acidic) to 15 kT/e (blue, basic). The surface exhibits a mix of positive and negative potential. The positive potential along the edge of the central cleft is due, in part, to the strictly conserved basic residues. Positive potential is also seen in the upper left and lower right corners of the protein surface in this panel. To our knowledge, the intracellular pH of *Sulfolobus* has not been accurately determined. However, because *Sulfolobus* is an acidophilic organism, it is possible that the intracellular pH is also more acidic than in most organisms. This might accentuate the positive potentials depicted in panels C and D. The electrostatic potentials at the surface were calculated with SPOCK (29), using a probe radius of 1.4 Å, a temperature of 353 K, an ionic strength of 0.15 M, pH equal to 6.0, with protein and dielectric constants of 4 and 80, respectively. The image itself was prepared with PyMOL. **(D)** Relative to panel C, B116 has been rotated by 180° about the depicted horizontal axis to reveal a preponderance of negative charge on the non-conserved face.

For the surfaces surrounding the strictly conserved regions, a mixed pattern of positive and negative potential is observed, with the positive potential clustered around the rim and walls of the cleft (Figure 4-5C). In contrast, an 180° rotation about the horizontal axis reveals a surface with strong negative potential (Figure 4-5D). This surface is distant from the conserved residues pictured in panels 4-5A and 4-5B. Importantly, the more neutral and basic pIs of a few B116-like proteins suggest that the negative potential on this surface of B116 is not conserved across the B116-like family of proteins.

Disordered or highly flexible regions of a protein structure are often important for the function of that protein. As the primary sequence determines protein structure, it must also determine the lack of structure. Thus programs such as PONDR (75, 114, 115) can predict protein disorder from the linear sequence. The crystal structure represents a population average of all the molecules that make up the crystal and some information about conformational flexibility can be obtained as a result. Elevated B factors (temperature factors) result from greater conformational freedom of the atoms in the protein structure. An examination of the PONDR disorder prediction results and the B factors of the B116 structure itself may therefore suggest regions that may contribute to the protein's function (Figure 4-6). With the exception of β 4, the predicted disorder from PONDR (Figure 4-6A) agrees fairly well with elevated B factors in the structure (Figure 4-6C). The surface of the protein suggested to have the greatest conformational flexibility corresponds with the conserved surface further suggesting that it is likely involved in the function of the protein.

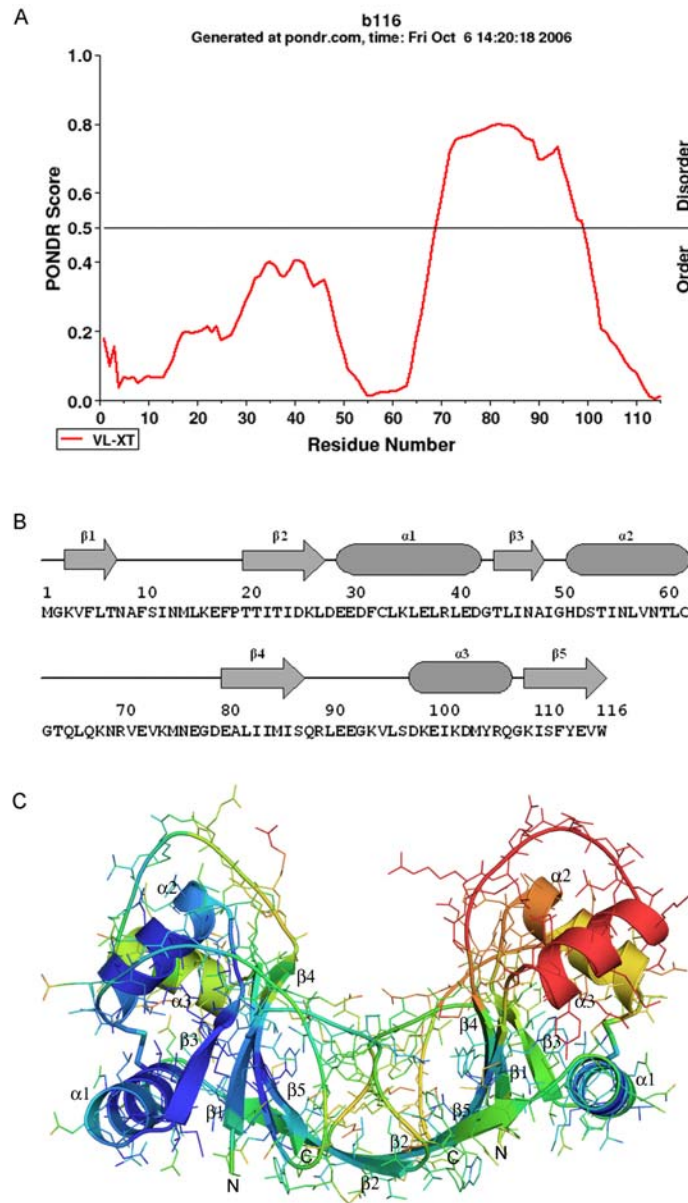


Figure 4-6. Predicted Disorder and Flexibility of B116.

(A) Residues 71 to 100 of the B116 primary sequence are predicted to be disordered by the VL-XT algorithm on the POND web server (75, 114, 115). **(B)** B116 sequence with the corresponding secondary structure map. The predicted disordered region encompasses half of the $\alpha 2/\beta 4$ loop through $\beta 4$ and the $\beta 4/\alpha 3$ loop and into the N-terminus of $\alpha 3$. **(C)** The B116 structure is colored by the B factor [from low (blue) to high (red)] to show regions of the protein that likely have increased flexibility. The mainchain is a cartoon and all atoms are shown as lines to highlight the flexibility of the sidechains. The regions predicted to be disordered by POND agree fairly well with elevated B factors with the exception of $\beta 4$, which displays fairly average B factors and is likely stabilized by adjacent β -strands. The most elevated B factors are apparent in $\alpha 2$, much of the $\alpha 2/\beta 4$ loop, the $\beta 4/\alpha 3$ loop, and $\alpha 3$. These surface-exposed regions may need to be flexible to facilitate an interaction with another molecule. The B factors of the left subunit are noticeably lower than those of the right subunit. This is due to the presence of crystal contacts with a neighboring molecule.

Discussion

While the B116 surface is quite convoluted, especially in the areas surrounding the conserved surface features, it is devoid of any obvious pockets that might play a role in recognition of small molecules. Thus, it seems likely that B116 will interact with a macromolecule. Given the inherent two-fold symmetry in the B116 homodimer, interacting macromolecules should also possess (pseudo)two-fold symmetry, at least when bound to the B116 homodimer. While the overall pI of the protein is acidic, the strictly conserved basic residues (His⁵¹, Arg⁷⁰, and Arg⁸⁹) impart a significant positive potential along the lip and sides of the central cleft. This suggests that the lip and walls of the central cleft might interact with a macromolecule carrying significant negative charge. The overall shape of the conserved surface and the intervening cleft is also a consideration. This surface is relatively convoluted, whereas most protein-protein interactions utilize relatively flat surfaces. The spacing between symmetry related residues in the highly conserved β 4- α 3 loop is also relevant. In fact, using standard B-form DNA as a metric, one can see that the separation between the loops is roughly equivalent to the spacing of the major groove along the DNA axis (Figure 4-7). In addition, the side-by-side comparison also highlights the apparent complementarity between the conserved surface features of B116 and double stranded DNA (Figure 4-7). An interaction between B116 and double stranded DNA can be envisioned in which the β 3- α 3 loops inserts into adjacent tracks of the DNA major groove, while the intervening ribose-phosphate backbone is accommodated by the central cleft of B116 with its conserved basic residues. The increased flexibility of this region of the protein,

suggested by the elevated B factors, may be important in facilitating a specific interaction.

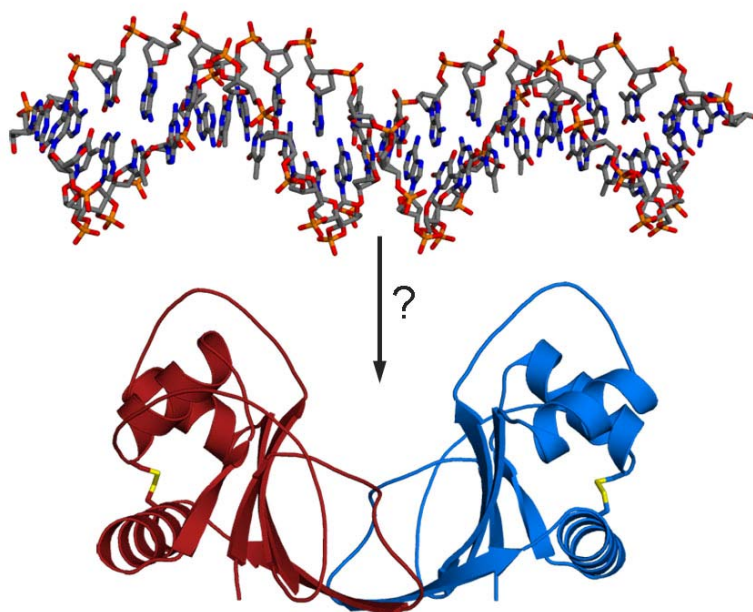


Figure 4-7. Spacing Between the $\beta 4/\alpha 3$ Loops and the Complementarity to dsDNA. The B116 homodimer is depicted as in Figure 4-2A. A 21-bp fragment of B-form DNA is positioned above B116 as a metric. The spacing between the two symmetry related $\beta 4$ - $\alpha 3$ loops is roughly equivalent to the major groove spacing of the DNA. Complementarity between the saddle shaped surface of B116 and the DNA is apparent. This extends to complementarity with respect to charge, as well (Figure 4-5C). The ribose-phosphate backbone flanking the minor groove in the center of the DNA fragment might be accommodated by the conserved positive charge flanking the central cleft of B116. Note that the upper right and lower left corners of Figure 4-5C also show areas of positive potential that might serve to recognize the ribose-phosphate backbone present at the outer edges of the DNA.

Thus, the overall surface features of the B116 homodimer are quite suggestive of a protein-DNA interaction in which B116 recognizes a (pseudo-)palindromic DNA sequence. Accordingly, preliminary evidence from electrophoretic mobility shift assays shows a nonspecific interaction between B116 and double-stranded DNA (Figure 4-8; Brian Eilers and Martin Lawrence, unpublished observation). DNA-binding proteins that

recognize a specific DNA sequence with high affinity will often show nonspecific interactions with DNA at elevated DNA or protein concentrations. Efforts are currently underway to identify a higher affinity, sequence specific interaction for B116. Until then, it is difficult to say whether this potential B116-DNA interaction might serve to regulate gene expression, or is indicative of a role in the modification and or synthesis of nucleic acid, either alone, or in a larger protein-nucleic acid complex.

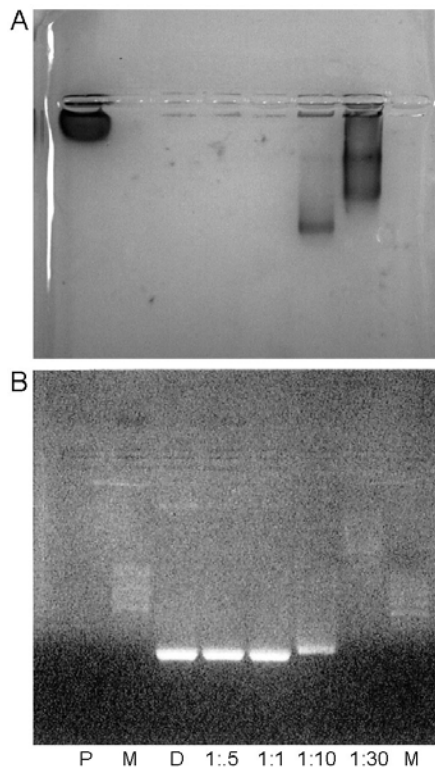


Figure 4-8. Nonspecific Interaction between B116 and DNA.

Electrophoretic mobility shift assay of a constant amount of DNA and an increasing amount of B116. Panels A and B are the same gel with panel A stained for protein and panel B stained for DNA. P = protein only, M = DNA markers, D = DNA only, 1:5 = 1 to 0.5 molar ratio of DNA to protein concentrations, 1:1 = equal molar ratio of DNA to protein concentrations, 1:10 = 1 to 10 molar ratio of DNA to protein concentrations, 1:30 = 1 to 30 molar ratio of DNA to protein concentrations. A band shift is observed for both the protein (down from protein alone) and the DNA (up from DNA alone) when the protein concentration is 10-fold greater or more than the DNA concentration (see lanes labeled 1:10 and 1:30 vs. lanes labeled P and D). The change in the electrophoretic mobility of the species indicates an interaction between the protein and the DNA. Eilers and Lawrence, unpublished observation.

As noted earlier, B116 is orthologous to proteins encoded by six other crenarchaeal viruses representing three additional viral families. Because the similarities between these families are quite limited with respect to morphology and genome sequence, the extent of the conservation of this gene product is remarkable, especially when considering the higher mutation rates typically observed in viruses. There are some similarities with respect to these viruses however. Namely, they all inhabit acidic hyperthermophilic environments and infect hosts that belong to the family *Sulfolobaceae*. The maintenance of the observed sequence conservation in these diverse viruses might then suggest evolutionary pressure exerted by a feature common to their respective hosts. That there are no homologues in the respective host genomes but there are in a few bacterial species provides strong evidence for horizontal gene transfer (105).

The intramolecular disulfide bond in B116 is also worthy of comment. B116 is not among the nine proteins identified in the purified viral particle (78), however, it is found in STIV infected *Sulfolobus solfataricus* (A. Ortmann, E. Larson, and M. Young, *et al.*, unpublished observation). This suggests that B116 is a cellular protein, expressed in the host cytoplasm upon infection with STIV. Though disulfide bonds are generally rare in the intracellular proteins of most prokaryotes (62), there is strong genomic evidence that suggests disulfide bonds are commonly employed to stabilize the intracellular proteins of *Sulfolobus solfataricus* (13) and its associated viruses (73). Thus, the likely role of the disulfide bond is to enhance the thermostability of the B116 fold.

Despite the absence of a structural homolog with known function, a preliminary understanding of the function of B116-like proteins has emerged. These results will guide future biochemical and genetic studies of the B116-like family of proteins and help

to decipher their role in a process of central importance to the crenarchaeal viruses. As this occurs, the structure of B116 will continue to contribute to a more detailed understanding of the structure-function relationships inherent in this newly discovered protein fold. Furthermore, as this structure is somewhat unique, it is valuable to the biological community as a whole in this time of abundant sequence data as it fills a gap of the primary sequence to tertiary structure fold space and will aid in future structure predictions.

Conclusions

The work presented in this chapter on the structural characterization of B116 from STIV demonstrates the utility of pursuing the structures of unknown proteins. Although it was envisioned that functional clues would primarily come from structural homology to a protein of known function, this was not the case for B116. Using knowledge gained from a detailed understanding of the protein on the molecular level, however, provided valuable insight that led to specific suggestions about its potential role in the viral lifecycle.

The structural characterization of B116 is a significant advance in the field of crenarchaeal virology because it is a ubiquitous protein among these unique viruses. Knowledge gained from its study will therefore go beyond STIV and help shed light on processes that are important to the viral life cycles and virus-host relationships of the crenarchaeal viruses as a whole. We hypothesize that B116 functions at the interface of virus and host so it is possible that study of this protein will also help to shed light on the host's cellular processes. It is highly likely that all the B116-like proteins from these

diverse viruses will adopt a very similar three-dimensional fold and will accordingly play an analogous role within the context of each of the viral lifecycles and their relationship with the host.

Obviously, more work needs to be done in the characterization of this protein that is beyond the scope of this research project. The structure of B116 has clearly provided insight into its potential function and now further research must be done to verify our hypotheses and lead to a deeper understanding of the STIV lifecycle. The structural analysis of the overall fold of B116 leads us to believe that it interacts with DNA and its conservation across several diverse viral families suggests that this DNA is derived from the host. Although we favor the binding partner to be DNA due to surface characteristics, we cannot rule out a protein binding partner so both will be pursued. Transcript mapping and expression profile experiments being pursued in Mark Young's laboratory will provide valuable insight into when in the viral lifecycle B116 is active.

To investigate the possibility of a DNA binding partner, systematic evolution of ligands by exponential enrichment (SELEX) (19, 134) will be used. This process involves the selection of specific sequences of DNA that interact with the target protein from an initial pool of random oligonucleotides containing defined ends to allow for subsequent amplification. The selection process involves cycles of binding the oligonucleotide pool to the protein, washing away those that do not interact, elution of those that do interact, and amplification of the resulting pool to create a new pool for the next round. With each consecutive round, the pool is enriched for specifically interacting sequences. After a number of enrichment cycles have been performed, the resulting oligonucleotides are sequenced and a consensus binding sequence is identified.

Knowledge of the virus and host genome sequences will allow putative binding sites to be identified. Once a specific binding motif is identified, the interaction between protein and DNA can be investigated in more detail and the structure of the complex will be pursued.

To pursue the identification of possible protein binding partners several strategies may be employed (81). Polyclonal antibodies have been produced against B116 and may be used in conjunction with STIV-infected cultures of *Sulfolobus solfataricus* P2 to isolate B116 in complex with its binding partner. Knowing the timing of B116 expression in the viral lifecycle will be a great help in this effort. Alternatively, bacterially expressed and His-tagged B116 is easy to obtain and may be used to identify putative binding partners. The His-tag is located on the opposite face from the presumed business end of the protein and so is not expected to interfere with binding. The His-tagged protein can be bound to an IMAC column and lysate from infected or uninfected *S. solfataricus* P2 is poured over the column. If B116 has a high affinity-binding partner and there is enough of it in the lysate, B116 complexes may be eluted. In both cases, mass spectrometry and N-terminal Edman sequencing can identify the binding partner since the genomes of *S. solfataricus* P2 and STIV are known.

Further, since we believe that disulfide bonds contribute to the stability of the intracellular proteins of the Crenarchaea (discussed in detail in Chapter 5), detailed studies of the affect of the disulfide bond on the thermostability of B116 should be investigated. Cysteine mutants of B116 should be constructed or cysteine-free B116 homologues should be cloned so direct comparisons can be made in thermal denaturation experiments. The cysteines of B116 are not in close proximity to the presumed active

site of B116 nor are the cysteines conserved in all homologues so they are not expected to be important for the function of the protein. However, binding studies should also be performed on cysteine-free proteins to determine if they do have an affect on the binding interaction.

STRUCTURAL CHARACTERIZATION OF F93; A WINGED-HELIX PROTEIN
THAT POINTS TOWARD STABILIZING DISULFIDE BONDS IN THE
INTRACELLULAR PROTEINS OF A HYPERTHERMOPHILIC VIRUS

Introduction

F93, a 93-residue protein in reading frame F of the STIV genome (Figure 5-1). The sequence of F93 shows weak similarity (well below the default PSI-BLAST threshold value) to predicted PadR-like transcriptional regulators from a few Archaea and Bacteria in a PSI-BLAST search (4). Likewise, if the expect values are greatly relaxed in a conserved domain (CD) search (80), similarity to predicted PadR-like family members is detected. Members of this family are transcriptional regulators that appear to be related to the MarR family. PadR itself is involved in negative regulation of phenolic acid metabolism. Though this suggests that F93 is a winged helix DNA-binding protein, the sequence and motif similarities are weak and they are to proteins that have not been characterized biochemically. Furthermore, the protein fold recognition algorithm that uses 1D and 3D sequence profiles coupled with secondary structure and solvation potential information utilized by the protein threading server 3D-pssm (64) fails to find meaningful matches. The highest scoring hit is to an engineered three-helix bundle with an insignificant E-value of 2.86. The lack of concrete functional information about this protein from its sequence led us to seek its structure to obtain functional information.

This protein expresses well and diffraction quality crystals were produced so the structural characterization was pursued.

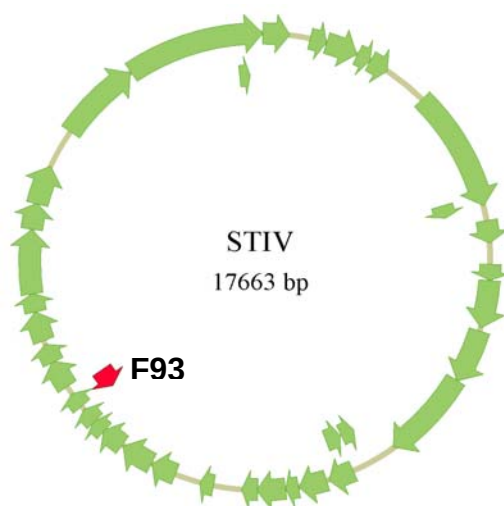


Figure 5-1. Location of F93 in the STIV Genome.

Structural and biochemical characterization of F93 reveals a homodimeric winged-helix protein that is likely to function as a transcriptional regulator, confirming the sequence-based predictions. Notably, we find an interchain disulfide bond that spans the dimer interface. The occurrence of this disulfide bond in a presumptive intracellular protein prompted further analysis of the cysteine distribution in the STIV proteome. The analysis reveals a pattern consistent with the frequent use of disulfide bonds in small intracellular proteins encoded by STIV, strongly suggesting that intracellular disulfide bonds make a significant contribution to the thermostability of the viral proteome. While stable disulfide bonds are rarely found in the cytoplasm of most organisms, our conclusion is consistent with the pioneering observations out of Todd Yeates's lab on the

occurrence of stabilizing disulfide bonds in cytoplasmic proteins of certain Archaea (13, 79, 88). Our work strongly supports these conclusions and extends the observed occurrence of stabilizing disulfide bonds in cytoplasmic proteins to the genome of a hyperthermophilic virus.

Results

The F93 construct used in this study codes for the 93 amino acids of the native protein plus an additional C-terminal His-tag, resulting in 99 residues with a calculated mass of 11,970 Da. Purified F93 elutes from a Superdex™ 75 size exclusion column as a single peak with an apparent molecular weight of approximately 25 kDa, suggesting that it is present as a homodimer in solution. The protein crystallizes in space group C2221 with two copies of the F93 polypeptide in the asymmetric unit. The structure was initially solved at 2.4 Å resolution by single-wavelength anomalous diffraction (SAD) using selenomethionyl-incorporated protein, with subsequent refinement at 2.2 Å resolution using native data. Details on data collection and model refinement are presented in Tables 2-8 and 2-9, respectively, in Chapter 2. The structure has been deposited in the Protein Data Bank (www.pdb.org) under accession code 2CO5.

Structure of F93

The structure of F93 (Figure 5-2) reveals a winged-helix (or winged-helix-turn-helix, wHTH) fold, a subclass of the helix-turn-helix (HTH) protein superfamily. Members of the wHTH family share a bundle of three α -helices, which in F-93 is comprised of helices $\alpha 1$, $\alpha 2$, and $\alpha 3$. Helices $\alpha 2$, $\alpha 3$ and the intervening turn comprise

the HTH motif, in which $\alpha 3$ serves as the recognition helix. Following $\alpha 3$, two β -strands connect through a reverse turn to form a small flanking antiparallel β -sheet that constitutes the “wing” of this structural motif. These features are important to the DNA-binding function of these proteins; the recognition helix generally makes base-specific contacts within the DNA major groove, while the wing interacts with the ribose-phosphate backbone and/or the minor groove (6, 21).

Variations on the wHTH DNA-binding domain that are common to other wHTH proteins are also seen in F93. Namely, a small β -strand ($\beta 1$) inserted between $\alpha 1$ and $\alpha 2$, and a fourth α -helix ($\alpha 4$) C-terminal to the wing. The additional β -strand hydrogen bonds in an antiparallel fashion to β -strand 3 of the wing and results in a three-stranded, antiparallel β -sheet with a topology of 1,3,2. Whereas the C-terminal helix ($\alpha 4$) extends away from the wing, roughly parallel to helix $\alpha 2$, and antiparallel to helix $\alpha 4$ of the neighboring molecule in the asymmetric unit. In most cases, C-terminal extensions to the core wHTH domain are involved in dimerization (6), and this is indeed the case in F93.

These interactions with a neighboring molecule result in formation of the homodimer (Figure 5-2) identified by size exclusion chromatography, MALDI-TOF mass spectrometry (results not shown), and SDS-PAGE under nonreducing conditions (Figure 5-5). Much of the dimer interface is composed of interactions with, and between, the C-terminal half of $\alpha 4$ from each monomer. The N-terminus and beginning of $\alpha 1$, the C-terminus of $\alpha 2$, and the $\alpha 2$ - $\alpha 3$ loop are also involved. The dimer interface is composed of a hydrophobic core, and approximately 1,100 Å² of accessible surface area is buried upon formation of the dimer.

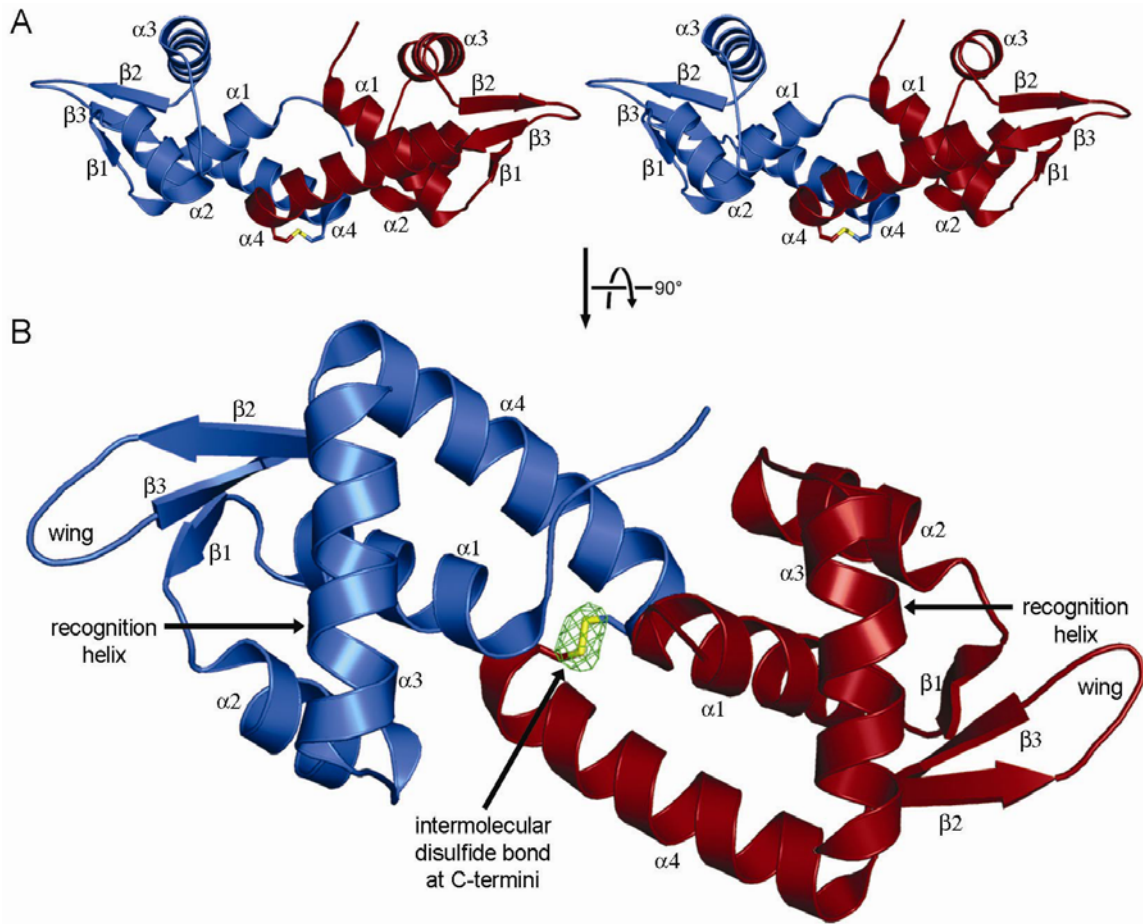


Figure 5-2. Ribbon Diagrams of the F93 Homodimer.

(A) Stereo image of the side view of the covalent homodimer with α -helices and β -strands labeled in ascending order from N-terminus to C-terminus. Chain A is shown in red, chain B in blue. The polypeptide traces from the N-terminus of the subunit through $\alpha 1$, $\beta 1$, $\alpha 2$, $\alpha 3$, $\beta 2$, $\beta 3$, and $\alpha 4$. The disulfide bond covalently linking Cys⁹³ at the C-terminus of each subunit is highlighted in yellow. The dimer interface is composed primarily of the N-terminus, the N-terminal end of $\alpha 1$, and the C-terminal half of $\alpha 4$. **(B)** The F93 dimer is rotated 90° relative to the orientation in panel A, about the depicted horizontal axis. The view now looks down upon the recognition helices ($\alpha 3$) of each monomer, rather than end-on. The secondary structural elements are depicted as in panel A with additional labels to highlight key features common to the winged-helix-turn-helix (wHTH) fold, namely the recognition helices ($\alpha 3$) and the wings ($\beta 2$ - $\beta 3$). The disulfide bond is again shown as a yellow stick, surrounded now by a difference density map contoured at 5 σ (green mesh) in which the sulfur atoms were omitted from model refinement.

Interestingly, an intermolecular disulfide bond is present between the two subunits of the dimer. The disulfide bond is formed by Cys⁹³, which is present at the C-terminal end of helix $\alpha 4$, and thus explains the covalent dimer seen by MALDI-TOF

mass spectrometry and SDS-PAGE under nonreducing conditions (discussed below). Homodimeric winged-helix proteins generally possess more extensive dimer interfaces than that seen here for F93. These generally result from additional domains C-terminal to helix $\alpha 4$ (Figure 5-4B) or N-terminal to $\alpha 1$. In contrast, F93 appears to rely on the disulfide bond as a substitute for the more extensive interface.

Searches for nearest structural neighbors using VAST (46) and DALI (54) indicate a close resemblance to MarR-like families of transcriptional regulators. For example, DALI identifies 82 structurally equivalent C α positions between F93 and MarR (PDB ID 1JGS) (2) that superpose with a root mean square deviation of 1.8 Å. Members of the MarR family are typically involved in transcriptional regulation, where the core structural motif, present in F93, binds to a specific DNA target sequence. These similarities suggest that F93 likewise plays a role in transcriptional regulation. However, structural homology to other members of the WHTH families should not be overlooked as they may indicate alternative functions for this protein. For example, F93 also shows significant structural similarity to replication terminator proteins (RTP) from various sources. Thus, a role in replication of the viral genome might also be considered for F93.

Disordered or highly flexible regions of a protein structure are often important for the function of that protein. As the primary sequence determines protein structure, it must also determine the lack of structure. Thus programs such as PONDR (75, 114, 115) can predict protein disorder from the linear sequence. The crystal structure represents a population average of all the molecules that make up the crystal and some information about conformational flexibility can be obtained as a result. Interpretable electron density will be absent for the most flexible regions and elevated B factors (temperature

factors) result from greater conformational freedom of the atoms. PONDR and the F93 structure itself suggest conformational flexibility at the N-termini and the wings (Figure 5-3). The flexibility in these regions is likely important for the function of F93 as they need to interact with the ribose-phosphate backbone upon DNA-binding.

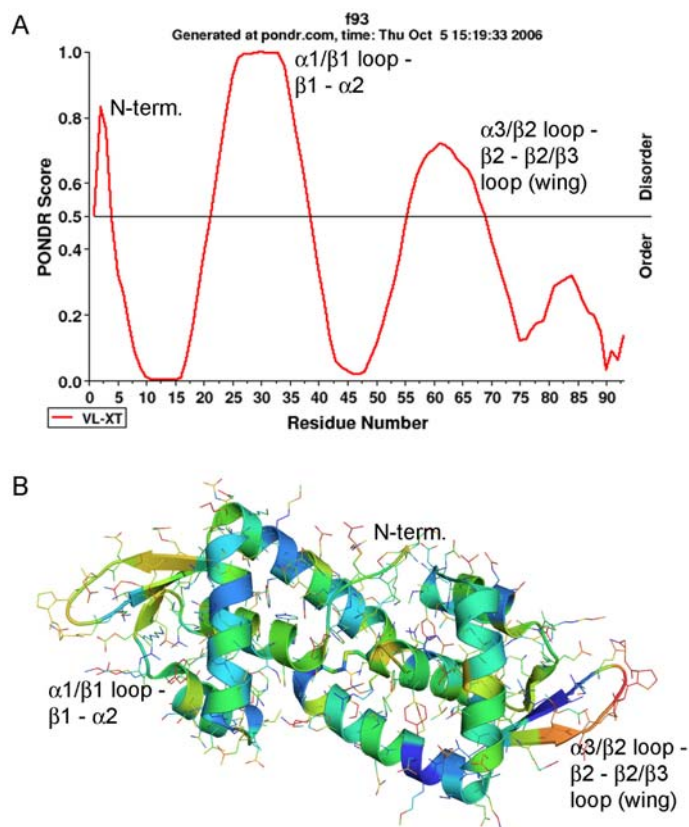


Figure 5-3. Predicted Disorder and Flexibility of F93.

(A) Disordered regions predicted from the F93 primary sequence are labeled with the secondary structural elements from the crystal structure in which they belong. The VL-XT algorithm on the PONDR web server (75, 114, 115) was used to calculate the predicted disorder. **(B)** The F93 structure is colored by the B factor [from low (blue) to high (red)] to show regions of the protein that likely have increased flexibility. TLS refinement was used so B factors are residual B factors resulting in slightly reduced values. The mainchain is a cartoon and all atoms are shown as lines to highlight the flexibility of the sidechains. The predicted disorder agrees fairly well with elevated B factors. The N-terminus is very flexible as demonstrated by the lack of interpretable electron density for the first 4 residues of the A chain. The wings display the highest mainchain B factors and the density is difficult to interpret in places. PONDR predicts the greatest disorder for the region of the $\alpha 1/\beta 1$ loop through $\beta 1$ and through $\alpha 2$. This is not the case although parts of this region, particularly the $\alpha 1/\beta 1$ loop through $\beta 1$, do display slightly elevated B factors.

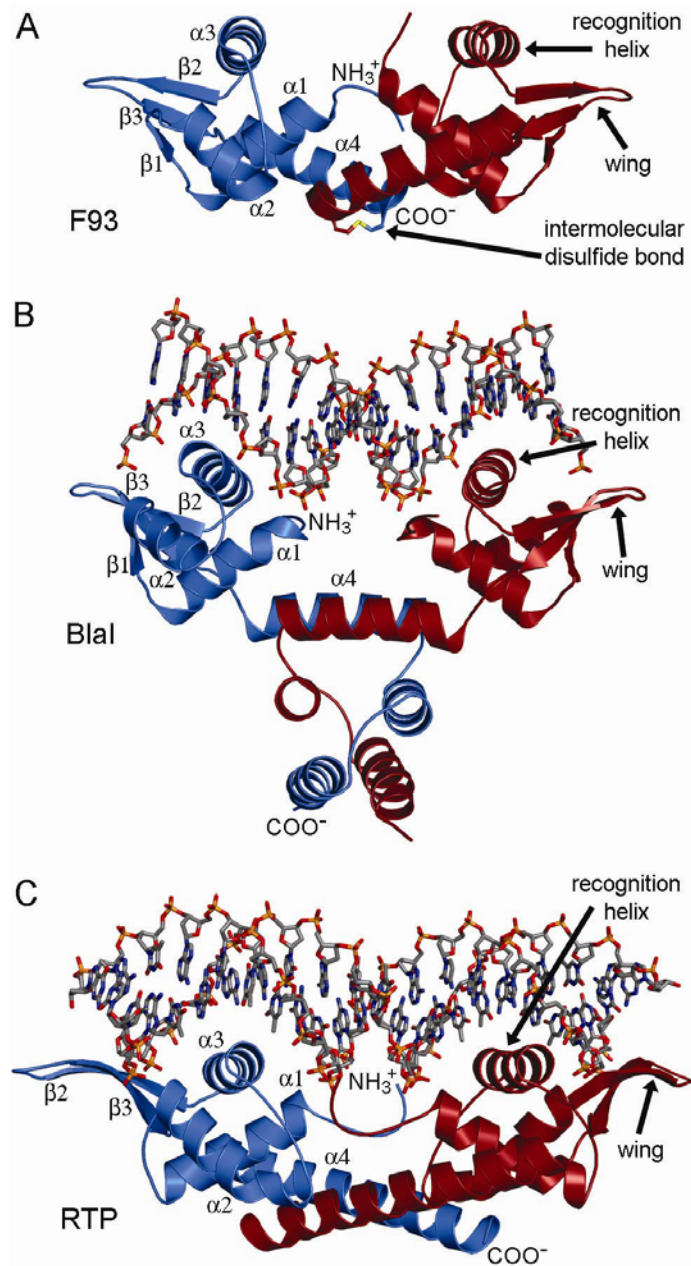


Figure 5-4. F93 and Representative Structural Neighbors in Complex with Target DNA. The orientation in all panels correspond to that depicted in Figure 5-2A. Elements of all models that are common to F93 and the wHTH fold are labeled accordingly. **(A)** Structure of F93 homodimer. The N-termini of each chain adapt different conformations and interpretable electron density is absent for the first four amino acids of chain A (red). Similarly, the wings adapt slightly different orientations between chain A and chain B and the density is relatively poor. This indicates flexibility within these regions, which may be important for DNA recognition and/or specificity (120). **(B)** Structure of the BlaI-DNA complex (PDB ID 1XSD) (118). **(C)** Structure of RTP-DNA complex (PDB ID 1F4K) (140). In panels B and C, it is apparent that the N-termini, the recognition helices, and the wings facilitate DNA recognition. Similar interactions can be inferred for F93 with respect to its unknown target DNA sequence. Panels A through C also illustrate diversity in the relative orientation of the winged-helix motif that is used to tune the fit

Figure 5-4 (cont.).

between homodimeric transcription factors and their target DNA. In addition to sequence variability, the relative pitch of the recognition helices and the spacing between them can be adjusted to recognize (pseudo-)palindromic sequences of various lengths. The winged-helix motif found in BlaI is obviously accompanied by a more elaborate dimer interface. Similarly, the extended $\alpha 4$ helix in RTP also results in a more extensive dimer interface than that seen in F93. F93 may have incorporated the disulfide bond (shown in stick representation in panel A) to compensate for the smaller dimer interface.

Putative Interactions with DNA

Structural homology clearly suggests that F93 will function as a DNA-binding protein. However, additional properties inherent in F93 lend further support to this putative role. The protein, with a basic pI, is rich in arginine and lysine residues. This is particularly true for residues at the N-terminus, which lie between the two recognition helices of the homodimer. These basic residues are well positioned to interact with the ribose-phosphate backbone of DNA. In addition, the recognition helices contain solvent exposed aspartate and tyrosine side chains, residues that are capable of forming base-specific interactions.

Superposition of F93 onto any of several structural homologues in complex with DNA provides a model for the putative interaction between F93 and its target DNA (Figure 5-4). For example, structures of the β -lactamase repressor (BlaI, PDB ID 1XSD)(118) from *Staphylococcus aureus* (Figure 5-4B) and the replication terminator protein (RTP, PDB ID 1F4K) (140) of *Bacillus subtilis* (Figure 5-4C) have been solved in complex with their respective target DNA sequences. Each of these structures shares the core winged-helix fold and helix $\alpha 4$ seen in F93. However, BlaI has additional secondary structural elements C-terminal to helix $\alpha 4$, resulting in a more extensive dimer interface than that in F93. Likewise, a greater dimer interface is seen in RTP from *B. subtilis* due

to the longer $\alpha 4$ helix. In addition, RTP also shows significantly longer wings. As expected, each of these dimeric WHTH proteins bind DNA in a similar manner, with the recognition helices forming base-specific interactions in adjacent major grooves of the DNA, while the positive N-termini and the wings interact with the ribose-phosphate backbone. An analogous complex between the F93 dimer (Figure 5-4A) and its target DNA is easily imagined.

The Intermolecular Disulfide Bond

The intermolecular disulfide bond linking the C-termini of the A and B chains is a striking feature in F93 (Figure 5-2). Because this putative DNA-binding protein is not found in purified virus (78), it is expected to be an intracellular protein. However, disulfide bonds are not generally found in intracellular proteins, perhaps suggesting that the disulfide bond is an artifact of crystallization. In contrast, the solvent-exposed disulfide bond is present in the crystal structure despite crystal growth in the presence of 0.5 mM TCEP, a strong reducing agent, suggesting that the disulfide is in fact biologically relevant. To further investigate the presence of this disulfide bond, freshly purified F93 was analyzed by SDS-PAGE, with and without the use of reducing agent in the sample buffer (Figure 5-5A). While F93 runs at the expected molecular weight in the presence of reducing agent (DTT, Figure 5-5A, lane 1), in the absence of reducing agent it is present largely as the homodimer (Figure 5-5A, lane 3). Further, size exclusion chromatography also indicates a dimeric protein and subsequent analysis by MALDI-TOF mass spectrometry indicates the material is present as the covalent homodimer (results not shown). Surprisingly, western analysis of the bacterial expression using

polyclonal antiserum against F93 in the absence of reducing agent shows that the monomer is largely absent even within the *E. coli* (Figure 5-5B, lane 3). Perhaps most importantly, this disulfide bond stabilizes a quaternary structure common to homodimeric winged helix proteins (Figure 5-4). Combined, these results strongly suggest the presence of an intracellular covalent homodimer *in vivo*.

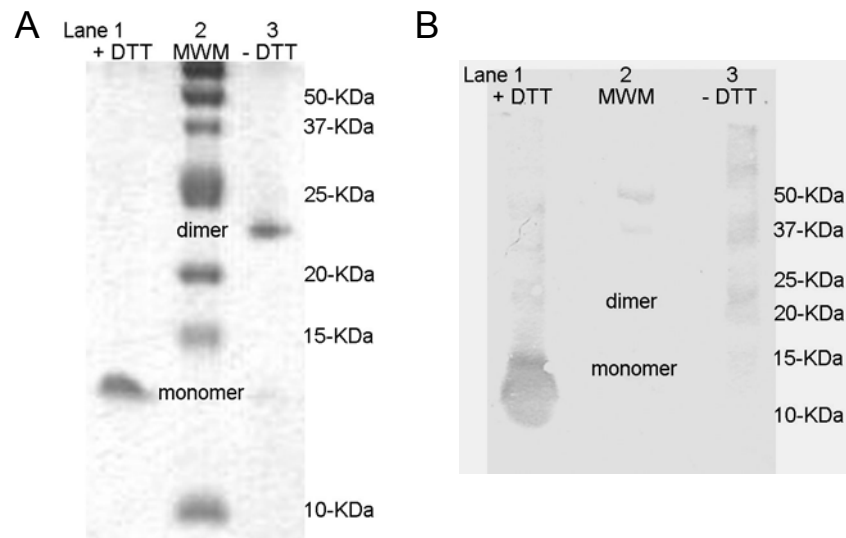


Figure 5-5. Analysis of F93 under Reducing and Non-reducing Conditions.

(A) Analysis of freshly purified F93 with SDS-PAGE. **(B)** Analysis of bacterial expression of F93 using polyclonal antiserum against F93. For each, reducing (+ DTT, lane 1) and non-reducing conditions (- DTT, lane 3) were used. An equal amount of purified F93 (panel A) or bacterial cell pellet (panel B) was loaded in lanes 1 and 3. The corresponding sizes of the molecular weight markers (MWM, lane 2) are shown to the right. Reduced F93 runs at an approximate molecular weight of 12,000 kDa (monomer) while non-reduced F93 gives a predominant band at approximately 24,000 kDa (dimer), the expected size of the covalently cross-linked homodimer.

Genomic Analysis of Cysteine Content

The presence of the F93 intermolecular disulfide bond prompted us to examine the cysteine content encoded within the STIV genome. Similar to Mallick *et al.* (79), we

looked for a cysteine pattern suggestive of disulfide bonds in intracellular proteins; specifically, an abundance of proteins with an even, rather than an odd, number of cysteine residues. Towards this end, a recent proteomic investigation of the purified viral particle has identified nine proteins (78) (labeled by asterisks in Figure 5-6A), which we consider as extracellular proteins. These proteins have been removed from the genome in order to yield a pool enriched in intracellular proteins. In addition, proteins exhibiting detectable signal sequences, determined by the program SignalP (14), were also considered extracellular and removed. Similarly, putative integral membrane proteins were identified using the program TMHMM (69, 125). Finally, because even numbers of cysteine residues are frequently employed in metal binding, these motifs were identified using the primary literature (105) or the program ScanProsite (45, 56, 122), and removed from the intracellular pool to prevent these potential false positives. The remaining proteins represent a genomic pool enriched in intracellular proteins.

When the number of cysteine residues in this pool of putative intracellular proteins is determined, a strong preference for even numbers of cysteine residues is indeed observed (Figure 5-6B, solid line). This strongly suggests an abundance of disulfide bonds in the intracellular proteins encoded in the STIV genome. Interestingly, this trend is also readily apparent in the unfiltered genome, as the particle-associated proteins generally lack cysteine and removing them from the genome to enrich for intracellular proteins had little affect on the general trend of the curve (Figure 5-6B, dashed line). It should be pointed out that proteins containing a single cysteine residue may participate in intermolecular disulfide bonds, as seen in F93.

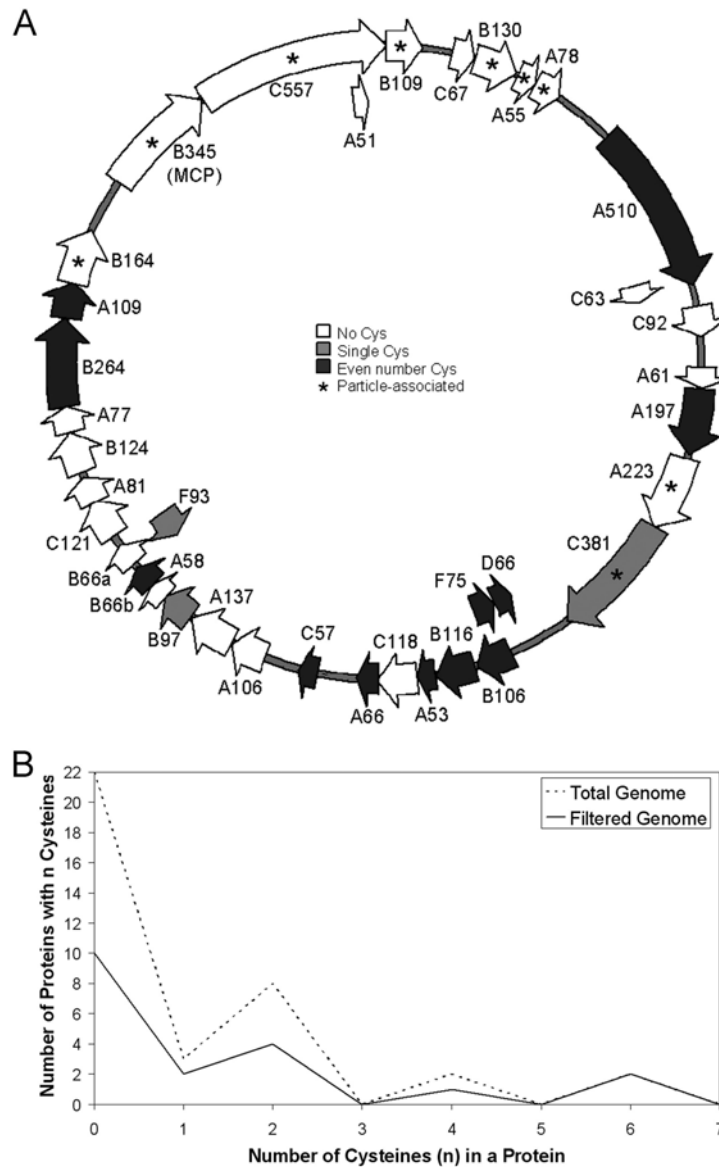


Figure 5-6. Analysis of the Genomic Cysteine Content.

(A) The STIV genome map showing ORFs coding for proteins that lack cysteine (white arrows), contain a single cysteine (gray arrows), or contain an even number of cysteines (black arrows). ORFs coding for proteins associated with purified virus (78) are labeled with an asterisk (*). With the exception of C381 and A223, these proteins cluster in one region of the genome and are noticeably lacking in cysteine. The map was created with Vector NTI Advance 10.1.1 (Invitrogen). **(B)** The distribution of cysteines within the STIV proteome. An abundance of disulfide bonds within intracellular proteins is suggested by the preference for even numbers of cysteine residues in a pool enriched for intracellular proteins (filtered genome, solid line). The genome was enriched for intracellular proteins by filtering out known extracellular proteins (78) (8 proteins lacking cysteine, 1 protein with a single cysteine, all labeled with * in panel A), proteins that contain predicted signal sequences or transmembrane helices (4 cysteine free proteins, 2 proteins with a pair of cysteines and 1 protein with 4 cysteines), and proteins that are predicted to contain a metal-binding motif (105) (2 proteins containing a pair of cysteines). The trend is also apparent in the unfiltered genome (dashed line).

Discussion

wHTH DNA-Binding Protein

F93 shows greatest similarity to the MarR-like families of winged helix proteins. In addition to the winged helix motif, members of these families possess a characteristic fourth helix that follows the wing (helix α_4 in F93). However, F93 differs slightly from MarR in that the connection between helices α_1 and α_2 is extended to include a short β -strand, resulting in a three-stranded wHTH as opposed to the two-stranded wHTH fold of MarR. Members of the MarR-like families are quite common in the Archaea, and most of the major Archaea-specific wHTH transcription factors are members of this larger family (6). The members of these families, barring a few exceptions, function as transcription factors (6), and thus suggest a similar activity for F93

The organization of the STIV genome suggests the presence of several multi-gene transcripts (112) (Figure 5-6A). Interestingly, the relative position of F93 to one potential transcript resembles the arrangement of several bacterial and archaeal operons to their regulatory transcription factors. Specifically, the gene for F93 is found adjacent to the 5'-end of a potential transcript beginning with open reading frame C121, but is encoded on the opposite strand (Figure 5A). This suggests that F93 may regulate transcription of C121 and the proteins downstream of it. However, other target sequences, including sequences within the host genome, are certainly possible, as are altogether different functions for a wHTH DNA-binding protein. For example, F93 shows similarity to RTP family members and thus might conceivably play a role in

replication of the viral genome. Future work to identify the DNA binding site and activity of the protein will be a priority.

In contrast to many MarR-like wHTH proteins, which show significant N- or C-terminal embellishments that often serve to modulate the activity of the protein in response to a signal, additional structural elements are notably absent from F93. Thus, the minimal structure of F93 suggests a constitutively active transcriptional regulator. However, regulation could be provided in *trans* by a viral or host protein, or through binding of a small molecule that blocks the DNA-binding surface, as is the case for MarR itself (2). In this regard, it is noteworthy that there is a short sequence motif (LTEKG) between strand $\beta 3$ and helix $\alpha 4$ that is conserved in STIV F93, SSV1 F-93 (and its homologues in other SSVs), and select MarR-like and PadR-like proteins in various archaeal and bacterial genomes (68). This conserved patch is distinct from the DNA-binding surface and could thus participate in an interaction with other viral or host proteins. The interaction might play a regulatory role, or serve to recruit additional proteins to the bound DNA.

The structure of F93 provides definitive evidence for a wHTH protein in the STIV genome. In conjunction with the previous work of Kraft *et al.* (68) and the more recent work of Menon *et al.* (unpublished) with wHTH proteins from SSV1, it is now clear that this structural motif is common to crenarchaeal viral genomes. Not surprisingly then, bioinformatics approaches by our lab (unpublished) and others (105) note the presence of weak signals suggestive of the wHTH motif in other hyperthermophilic viral genomes. As is the case for STIV F93, conserved domain (CD) searches with default values generally fail to identify this domain. However, using increased expect values, the

occurrence of the wHTH motif is suggested for three proteins from *Pyrobaculum* spherical virus (accession numbers: YP_015525.1, YP_015569.1, YP_015526.1) and for one protein from *Acidianus* rod-shaped virus 1 (accession number: CAI44193.1). While in some cases the individual CD searches may not yield statistically meaningful matches, the demonstration of the wHTH fold by structural studies of three such proteins from SSV1 and STIV suggests that these four proteins from *Pyrobaculum* spherical virus and *Acidianus* rod-shaped virus 1 are also bona fide members of the wHTH family of DNA-binding proteins.

Intracellular Disulfide Bonds

Before our work on F93, we noted the presence of an intramolecular disulfide bond in the structure of another STIV protein, B116 (PDB ID 2J85, Larson *et al.*, submitted; see Chapter 4 of this dissertation). While the function of B116 is unknown, it appears to be an intracellular protein, as it has not been found in purified virus (78), but is present in virally infected cells (A. Ortmann and M. Young, unpublished observation). Homologues of B116 are clearly present in other crenarchaeal viruses spanning several viral families (105, 112), including the *Rudiviridae*, the *Lipothrixviridae*, and the *Bicaudaviridae*. The two cysteines of the disulfide bond are present in the primary sequence of most of these B116 homologues. In a select few, however, both cysteines are absent and in no case is there only a single cysteine of the pair present. Hence, the presence of the two cysteine residues is highly coupled, consistent with formation of a disulfide bond and a selective pressure against a free cysteine in this protein. These

observations on B116, along with the intermolecular disulfide bond in F93, prompted us to examine the cysteine content of the STIV genome.

Interestingly, eight of the nine proteins found in purified viral particles lack cysteine altogether. The exception is open reading frame C381, in which a single cysteine is found (Figure 5-6). The general lack of cysteine in proteins constituting the viral particle may suggest that the solfataric environment into which the virus is secreted is not favorable for disulfide bond formation. Thus, there has been little evolutionary pressure for the maintenance of cysteine pairs in viral proteins exposed to this environment. In any event, these extracellular proteins clearly rely on other mechanisms to achieve thermostability (65). We note, however, that these proteins are relatively large in relation to the rest of the STIV proteome (see discussion below), and might be further stabilized upon formation of the viral particle (65).

In contrast to proteins constituting the viral particle, the cellular proteins encoded by STIV show increased cysteine content and a strong preference for even numbers of cysteine residues (Figures 5-6 and 5-7). A similar pattern was observed by Mallick *et al.* (79) in their work with a size restricted set of proteins from *P. aerophilum*. Specifically, a pool of putative intracellular proteins between 100 and 200 amino acids in size gave an even-odd cysteine distribution similar to that presented here for STIV, indicating an abundance of intracellular disulfide bonds in this hyperthermophilic crenarchaeal organism. In an alternative approach, they used sequence-structure mapping to estimate the abundance of disulfide bonds in the intracellular proteins of *P. aerophilum* and a number of additional prokaryotic organisms that have been completely sequenced. In contrast to their genome-based analysis, the sequence-structure mapping indicated

appreciable levels of intracellular disulfide bonds in at least 15 thermophilic and hyperthermophilic organisms (13, 79). Of these, *P. aerophilum* is predicted to have the greatest percentage of intracellular cysteines involved in disulfide bonds (44%) by genomic analysis and one of the greatest by sequence-structure mapping (greater than 10 times that expected by chance) (13, 79). Thus, sequence-structure mapping and genomic analysis single out the extraordinary cysteine distributions in *P. aerophilum*. While the sequence-structure mapping appears to be the more sensitive technique, it cannot be applied to the STIV proteome as it is largely devoid of significant sequence similarity to known structures. However, the genome-based analysis is apparently the more stringent test, and thus serves to underscore the significance of the even-odd cysteine distribution seen here for STIV.

Why does the STIV proteome display such a pronounced even-odd distribution? We believe this is due to the relatively small size of the proteins found within the genome. Smaller proteins necessarily possess a smaller hydrophobic core relative to their hydrophilic surfaces. Thus, the magnitude of the hydrophobic effect may be correspondingly smaller and additional means of stabilizing a particular fold may be required. In fact, proteins that are less than 100-residues in size are frequently found to be metal rich or, in the case of extracellular proteins, rich in disulfide bonds (113).

Indeed, when we examined the frequency of cysteine residues as a function of protein size, we found a pronounced preference for cysteine in the smaller proteins of the filtered (Figure 5-7) and unfiltered STIV genome (results not shown). However, decreased cysteine content as a function of protein size might be expected, whether or not the cysteines are involved in disulfide bond formation. At least for spherically shaped

proteins, increased size will result in a decreased surface to volume ratio. Thus, the relative occurrence of hydrophilic residues such as cysteine should decrease while at the same time there is a relative increase in the occurrence of hydrophobic residues. Importantly, though, the cysteine content of the STIV proteome shows a steeper decrease as a function of protein size than is predicted by the surface to volume ratio of a growing sphere (Figure 5-7A).

Alternatively, a simple empirical approach to predicting cysteine content as a function of protein size can be taken. In the absence of disulfide bonds, the cysteine content might be expected to parallel that of a similar amino acid, such as serine. Thus, we compared the decrease in cysteine content to the change in serine content as a function of protein size (Figure 5-7B). As a counterpoint, relative amounts of hydrophobic amino acids such as leucine might indicate changes in the size of the hydrophobic core. Thus, the relative serine and leucine content might accurately reflect the ratio of surface exposed hydrophilic residues to buried hydrophobic residues as a function of protein size. In this regard, we find the serine and the leucine content to be relatively flat as a function of protein size, in sharp contrast to the cysteine content (Figure 5-7B). This suggests that the globular approximation for proteins in the STIV proteome is inaccurate, and further highlights the pronounced preference for cysteine in the smaller STIV proteins.

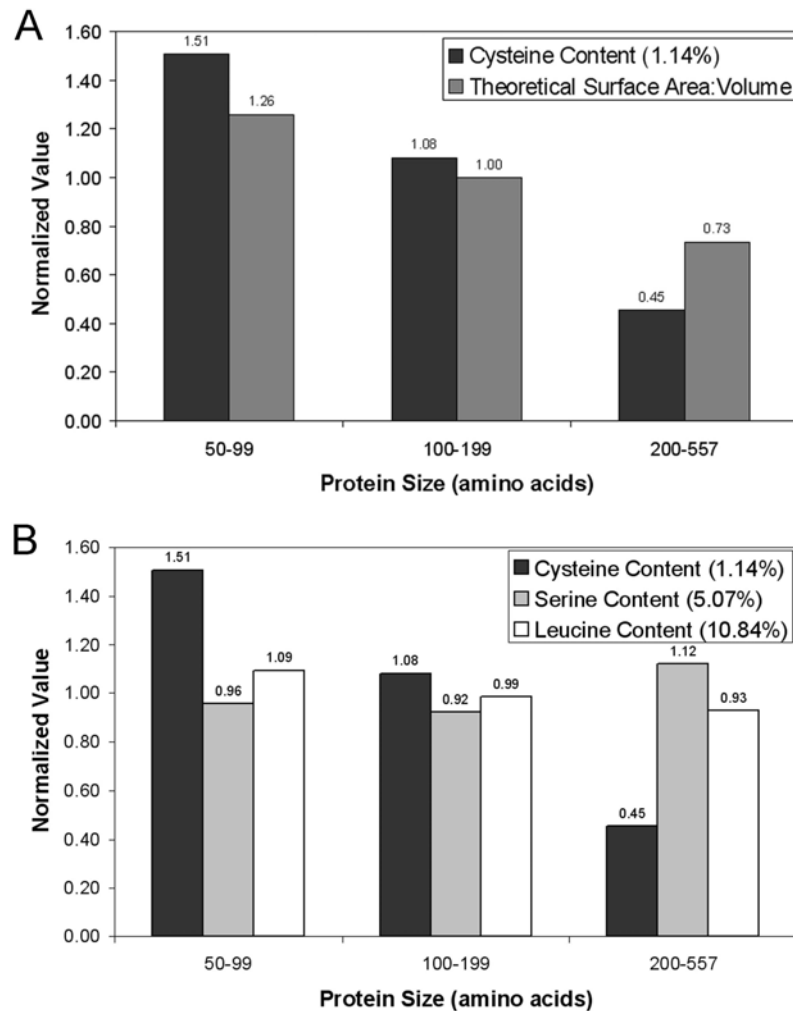


Figure 5-7. Cysteine, Serine, and Leucine Content as a Function of Protein Size.

(A) The filtered STIV genome was divided into bins according to protein size (51-99, 100-199, and 200-557 amino acids in length) and the cysteine content (dark boxes) of each bin was calculated (number of cysteines in the group divided by the total number of residues in the group) and normalized relative to the cysteine content of the filtered genome (1.14%). The theoretical surface area to volume ratio (light boxes) was calculated based on a spherical approximation of the average-sized protein from each bin. The surface to volume ratios were then normalized to that determined for the 100-199 amino acid-sized bin, putting them on a relative scale. The smaller proteins clearly show an enriched cysteine content. In contrast, the larger proteins display a paucity of cysteine residues. More importantly, the cysteine content falls as a function of protein size more rapidly than predicted by the theoretical surface to volume ratio of a spherical protein. This suggests that intracellular disulfide bonds are more common in the smaller intracellular proteins. **(B)** Cysteine content (dark boxes) compared to the serine (light boxes) and leucine (white boxes) content as a function of protein size in the filtered genome. The serine and leucine contents were calculated as described above for cysteine. Serine and leucine content were both normalized to their respective frequency in the filtered genome (serine content of 5.07%, leucine content of 10.84%). The change in cysteine content as a function of protein size may be expected to mirror that of a similarly hydrophilic residue such as serine and to display a trend opposite that of a hydrophobic residue such as leucine; however, this is not the case in the STIV genome, again suggesting that the smaller intracellular proteins show enriched cysteine content. In both panels, the normalized values are shown at the tops of the boxes.

Structural and biochemical support for the occurrence of stabilizing disulfide bonds in the intracellular proteins of thermophilic organisms has been provided by several laboratories. For instance, the activity and denaturation constants (both thermal and chemical) of methylthioadenosine phosphorylase from several crenarchaeal sources have been found to be greatly decreased in the presence of reducing agent (5, 24-26). The thermostability of TATA-binding protein from the crenarchaeal organism *P. woesei* was also found to be greatly decreased and the cysteine pair believed to be responsible for the increased thermostability is notably absent from mesophilic homologues, for which reducing agent has no effect on stability (34). Mutational studies of glycosyltrehalose trehalohydrolase from *S. solfataricus* (40) and isocitrate dehydrogenase from *A. pernix* (63) show that the disulfide bond is important for enzymatic activity and thermostability. Structural comparisons of adenylosuccinate lyase from *P. aerophilum* with those from mesophilic sources, lacking cysteines, suggest that the 3 disulfide bonds are strategically placed to limit the mobility of the peripheral domains (133). Further, it was found that reducing agents greatly reduce both thermal and chemical stability of the protein (133). Despite this mounting evidence, it is the genomic studies from Yeates and coworkers (13, 79, 88) that have revealed the surprising abundance of disulfides in select thermophilic organisms and the true extent of this phenomenon.

The ability of these thermophilic organisms to form disulfide bonds in the presumably reducing environment of their cytoplasm may be somewhat disconcerting because it is contrary to the current biological paradigm. However, work over the past ten years on the pathways regulating disulfide bond formation in *E. coli*, namely the DsbA/DsbB and glutathione reductase/thioredoxin reductase pathways, has shown that

relatively minor changes to these tightly regulated systems can result in cytoplasmic disulfide bond formation (17, 83, 108). This work with arguably the most well understood organism, culminating in the commercial availability of cell lines to allow for bacterial cytosolic expression of disulfide-rich proteins (Origami™ from Novagen), underscores the possibility that pathways exist to allow for cytoplasmic disulfide bond formation within the relatively poorly understood thermophile. Indeed, Beeby *et al.* (13) have identified a protein, protein disulfide oxidoreductase (PDO), in the genomes of many thermophilic organisms that may fill this role. The presence of PDO happens to correlate with both the predicted abundance of intracellular proteins containing disulfide bonds and elevated optimal growth temperature strongly suggesting that it is involved in the formation and maintenance of cytoplasmic disulfide bonds in these organisms (13).

Beeby *et al.* (13) also find that some thermophiles do not contain a predicted abundance of intracellular disulfide bonds, namely the methanogens and many sulfur-reducing organisms. Many of these are anaerobes that grow in very low redox potentials. *Thermoproteus tenax*, the host of TTSV1, is a hyperthermophilic, sulfur-reducing anaerobe so I would predict that it does not have an abundance of intramolecular disulfide bonds or a PDO gene. Its genome has not been sequenced so was not investigated by Yeates. However, Prangishvili *et al.* (105) note that TTSV1 codes for a thiol-disulfide isomerase. Interestingly, Menon, *et al.* (unpublished) find that the TTSV1 genome suggests an abundance of intracellular disulfide bonds. That its host may not suggests a real need for the virus to code its own protein involved in disulfide bond pathways.

Because viruses often provide a window into key biochemical processes of their host organisms, the abundance of disulfide bonds in the cellular STIV proteins suggests that the same might be true of its host. At the least, it implies disulfide bond formation in the cytoplasm of its host. Not surprisingly, *Sulfolobus solfataricus* is among those organisms predicted to carry significant numbers of intracellular disulfide bonds (13). This suggests that other crenarchaeal viruses, particularly those infecting *Sulfolobus*, may likewise utilize intracellular disulfide bonds to stabilize their cellular proteins. Indeed, structural work of Smita Menon and Brian Eilers on intracellular proteins from SSV1 shows that this is in fact the case (Menon *et al.*, manuscript in preparation; Eilers *et al.*, manuscript in preparation). Taken together, these studies present substantial new evidence at both the structural and genomic level for an abundance of intracellular disulfide bonds in hyperthermophilic organisms, and extend these observations to their equally intriguing viruses.

Conclusions

The structure of F93 is an excellent example of what we envisioned the structural analysis of the STIV proteome would provide. From the structure, we were able to determine that the likely function of the protein is DNA binding and its role in the viral lifecycle is in transcriptional regulation. Further, aspects of the structure provide a glimpse into the forces at work within the host cell and one of the mechanisms by which it may cope with living in a high temperature, acidic environment. Namely, many of the small cytoplasmic proteins encoded by the virus, and by extension the host as well, are stabilized by disulfide bonds. Though this statement is contrary to the current biological

paradigm, the structural and genomic evidence presented in this chapter lend support to the notion that hyperthermophilic organisms are rich in intracellular proteins with disulfide bonds. Disulfide bonds are certainly not the only force at work to stabilize the viral proteome. This work also demonstrates that oligomerization is likely to play an important role in stabilization of both intra- and extracellular proteins.

F93 displays the winged helix fold and structural homology shows that it is closely related to the MarR-like families of homodimeric transcriptional regulators. This strongly suggests a DNA-binding function for F93 and a role in regulating viral transcription. However, this role still needs to be confirmed. Though we favor this function, it is also possible that F93 regulates the transcription of host genes that are important for the viral lifecycle or that it is not involved in transcription at all. The fold of F93 is notably similar to a family of replication termination proteins so it is possible that it plays a role in replication of the viral genome.

The structural characterization of F93 was the first step. It is now important to test the hypotheses afforded by the structure and to clearly establish its DNA-binding function and role in the viral lifecycle. F93 does appear to bind DNA in very crude electrophoretic mobility shift assays (EMSA). However, this binding appears to be nonspecific under the conditions tested because a shift is observed with all the DNA fragments from the viral genome that were tested and even with the random plasmid fragments intended to be the negative controls. All DNA-binding proteins would be expected to have at least very weak affinity for all DNA sequences, but the observed binding is questionable and perhaps may be the result of nonspecific aggregation of the DNA and the protein. The binding conditions for these EMSA experiments will need to

be fine-tuned to better represent those found within the host cell to find a specific binding sequence. Since it is also possible that the recognition sequence for F93 is host-specific, rather than pursuing conditions that may allow for a specific DNA interaction with viral sequences in EMSA experiments, we will try to identify the specific recognition sequence through alternative means.

To look much more broadly for the specific DNA recognition sequence of F93, systematic evolution of ligands by exponential enrichment (SELEX) (19, 134) will be used. This process involves the selection of specific sequences of DNA that interact with the target protein from an initial pool of random oligonucleotides containing defined ends to allow for subsequent amplification. The starting pool is large enough to theoretically contain all possible nucleotide combinations for a random region of the desired length, approximately 16-22 nucleotides in this case. The selection process involves cycles of binding the oligonucleotide pool to the target protein, washing off the unbound oligonucleotides, elution of the bound oligonucleotides, and amplification of the bound oligonucleotides to create a new, enriched, pool for the next round. With each consecutive cycle, the oligonucleotide pool is enriched for sequences that specifically interact with the target protein. After a number of enrichment cycles have been performed, the resulting oligonucleotides are sequenced and a consensus binding sequence is identified. Knowledge of the viral and host genome sequences will allow the putative binding site to be identified.

Following the identification of the specific binding sequence, the complex will be further characterized by foot printing and experiments and attempts will be made to co-crystallize the protein-DNA complex. To investigate the affect of the disulfide bond on

the thermostability of F93, thermal denaturation experiments using differential scanning calorimetry should be done with wild-type protein in the presence and absence of reducing agent and on the cysteine-free mutant.

Though not directly related to STIV, Beeby *et al.* (13) find that some thermophiles do not contain a predicted abundance of intracellular disulfide bonds, namely the methanogens and many sulfur-reducing organisms. Many of these are anaerobes that grow at very low redox potentials. *Thermoproteus tenax* is a hyperthermophilic, sulfur-reducing anaerobe that is the host of the crenarchaeal virus TTSV1. It seems possible to speculate that *T. tenax* will not show an abundance of intramolecular disulfide bonds nor a PDO gene, though its genome has not been sequenced and so was not investigated by Beeby *et al.* Smita Menon in the Lawrence lab, however, has been looking at disulfide abundance within the crenarchaeal virus metagenome and has noted that TTSV1 appears to have an abundance of intracellular disulfide bonds. Further, Prangishvili *et al.* have reported that TTSV1 codes for a thiol-disulfide isomerase (105). That this host may not have an abundance of intracellular disulfide bonds but the virus does suggests a real need for the virus to code its own protein involved in disulfide bond pathways. If this TTSV1 protein really does help to form disulfide bonds in the cytoplasm, it would be very interesting to study because, as is the case for many viral proteins, it may be a much less complex system to study biochemically in comparison to the host's system.

CONCLUDING REMARKS

The last few years have seen a dramatic increase in the number of viruses known to infect hyperthermophilic organisms. These viruses generally infect the Crenarchaea and exhibit unprecedented diversity with respect to both morphology and genomic sequence, necessitating the creation of new viral families to reflect their unique characteristics (104). Despite their increasing numbers, the study of archaeal viruses is still in its infancy compared to viruses infecting the domains Eukarya and Bacteria. New morphologies are being described and new genome sequences are being released but the study of these viruses on the molecular level is falling behind. Little is known regarding archaeal viral life cycles, virus-host relationships, genetics, or biochemistry. This is particularly true for viruses of the Crenarchaea, and further study of these unique viruses is clearly needed. Importantly, while such studies will lead to a greater understanding of the viruses themselves, they are also expected to provide genetic, biochemical and evolutionary insight into their Crenarchaeal hosts and the requirements for life in the harsh environments in which these organisms often thrive. The research described here is the start of a path toward the complete structural characterization of one such virus, *Sulfolobus* turreted icosahedral virus, STIV.

The sequences of the predicted products coded for by the STIV genome show very little similarity to sequences of known function. To understand the virus at the molecular level, it is important to have detailed knowledge of the functions of its proteins. Since protein function is intimately related to its 3-dimensional structure and

the structure of a protein evolves much slower than its sequence, we believed that we could uncover the functions of these unknown proteins through structural knowledge that would be afforded by the solution of their x-ray crystal structures. The structural analysis presented here for three unknown STIV proteins clearly demonstrates that this is the case, protein structure does suggest function (Figure 6-1). Furthermore, the structures of these proteins have led to specific suggestions of their roles within the STIV lifecycle and even shed light on a biochemical process that may be important for the adaptation to life in a boiling, acidic hot spring.

The structure of A197 revealed that it is a glycosyltransferase. Specifically, it is a glycosyltransferase of the GT-A family. Viruses encode their own glycosyltransferases for many specialized glycosylation needs. Work in the Bothner lab found that MCP derived from purified virus is glycosylated and that it is difficult to deglycosylate. Furthermore, bacterially expressed MCP is not glycosylated (78). The glycosylation of MCP appears to be nonstandard and suggests that the virus relies on a specialized enzyme. We believe A197 serves this purpose; it is involved in the glycosylation of the MCP and thus, in the maturation of the viral particle.

The structure of B116 is unique and structural homology did not lead to a suggestion of function. The structure itself however proved invaluable nonetheless in leading to a hypothesis of its function. B116 is common to proteins from several other unrelated crenarchaeal viruses and a close inspection of the strictly conserved residues in relation to surface features of the protein led to the hypothesis that it is involved in DNA-binding. The structure of this protein is also very significant because its conservation across family lines suggests it is important to crenarchaeal viral lifecycles as a whole.

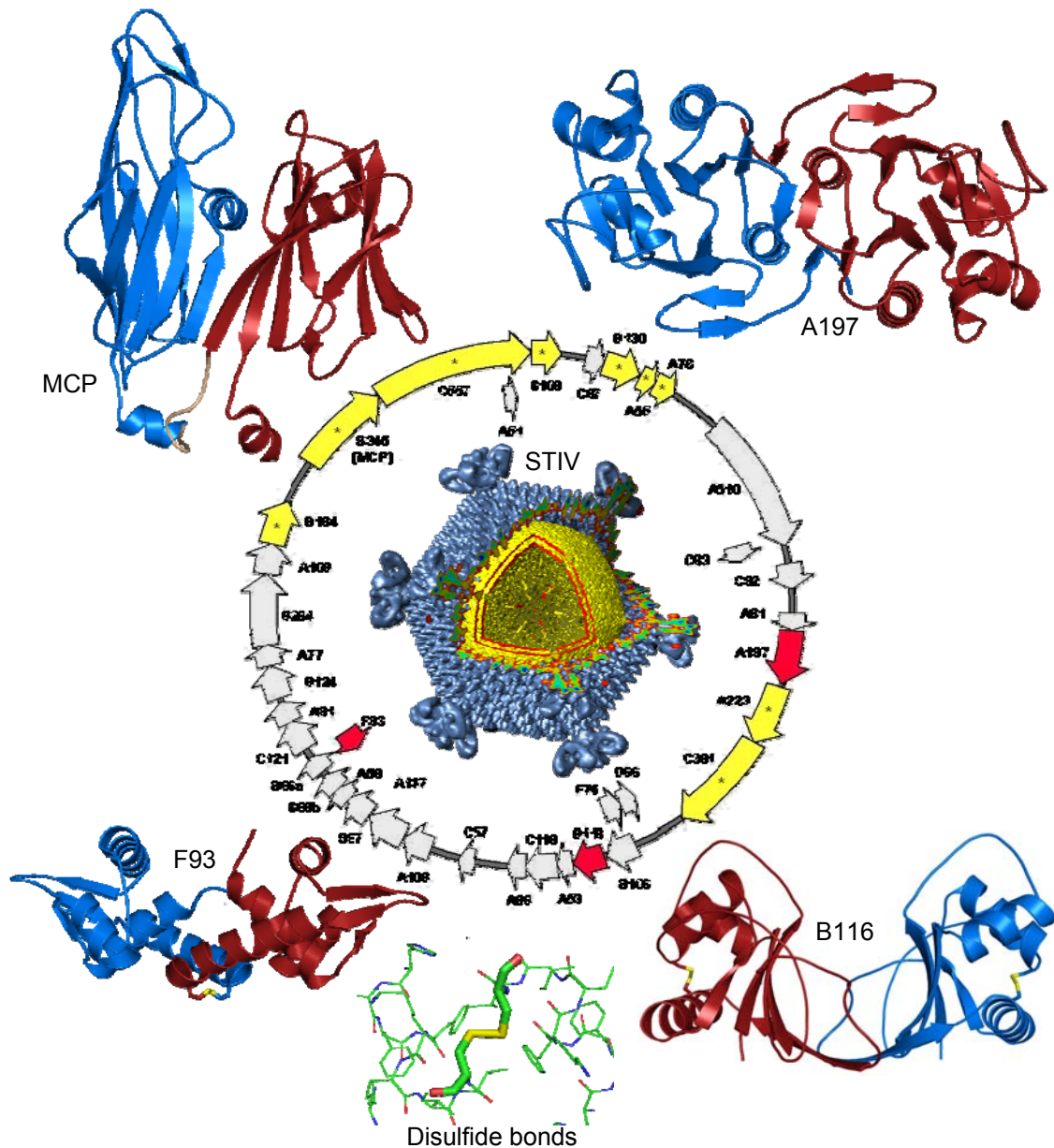


Figure 6-1. The Progress of the Structural Characterization of STIV.

The cryo-EM reconstruction of STIV and the structure of its major capsid protein, MCP, have revealed its relationship to a viral lineage that predates the split of the three domains of life. The structure of A197 revealed that it is a glycosyltransferase that likely plays a role in particle maturation. The structure of B116 is unique but has suggested that it has a function in DNA-binding. Due to its high conservation across unrelated viral families, it may be of critical importance to the crenarchaeal viral lifecycle. The structure of F93 revealed a winged helix DNA-binding protein that has a likely role in regulation of viral transcription. The structures of B116 and F93 have provided structural evidence that STIV, and by extension, its host, utilizes disulfide bonds to provide stability for the intracellular proteins.

The structure of F93 revealed a homodimeric winged helix protein with notable similarity to the MarR-like families of transcriptional regulators. Thus, this protein likely serves a role in the regulation of the viral genome. This structure, however, also revealed an intermolecular disulfide bond, which led us to genomic studies of cysteine distributions and the conclusion that disulfide bonds are frequently employed by crenarchaeal viruses as a strategy to stabilize their small, intracellular proteins. This supports the hypothesis that intracellular disulfide bonds are common among thermophilic organisms, particularly the Crenarchaea (13, 79), and extends it into their viruses as well.

Despite the advancement in our knowledge of this interesting and unique virus because of the work presented here, there is still much to uncover about STIV. The structural studies of A197, B116, and F93 are just the beginning as the majority of the STIV ORFs still code for proteins of unknown function. The structural characterization of the rest of the proteome will undoubtedly continue to lead to hypotheses of their functions. The structural characterization of each protein is not the end, however. While the structures of the proteins lead to specific suggestions of their functions, they also lead to more specific questions that can only be addressed biochemically. What are the substrates of the glycosyltransferase? Does it use the inverting or retaining mechanism of catalysis? Does B116 really bind DNA? What is the specific role of this protein in the crenarchaeal viral lifecycle to lead to its conservation in so many unrelated viral families? What is the specific recognition sequence of F93? Is its function dependent upon the dimer stabilization afforded by the disulfide bond? These questions and more are being raised and the structures have put us in a position to design specific experiments to

address them. Furthermore, the structures will no doubt be used continuously to provide insight as the functional hypotheses evolve.

REFERENCES CITED

1. Albers, S.-V., S. M. Koning, W. N. Konings, and A. J. M. Driessen. 2004. Insights into ABC transport in Archaea. *Journal of Bioenergetics and Biomembranes* 36:5-15.
2. Alekshun, M. N., S. B. Levy, T. R. Mealy, B. A. Seaton, and J. F. Head. 2001. The crystal structure of MarR, a regulator of multiple antibiotic resistance, at 2.3 Å resolution. *Nat Struct Mol Biol* 8:710-714.
3. Allen, M., D. Willits, M. Young, and T. Douglas. 2003. Constrained synthesis of cobalt oxide nanomaterials in the 12-subunit protein cage from *Listeria innocua*. *Inorg. Chem.* 42:6300-6305.
4. Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25:3389-3402.
5. Appleby, T. C., Mathews, II, M. Porcelli, G. Cacciapuoti, and S. E. Ealick. 2001. Three-dimensional structure of a hyperthermophilic 5'-deoxy-5'-methylthioadenosine phosphorylase from *Sulfolobus solfataricus*. *J Biol Chem* 276:39232-42.
6. Aravind, L., V. Anantharaman, S. Balaji, M. M. Babu, and L. M. Iyer. 2005. The many faces of the helix-turn-helix domain: Transcription regulation and beyond. *FEMS Microbiology Reviews* 29:231-262.
7. Arnold, H. P., U. Ziese, and W. Zillig. 2000. SNDV, a novel virus of the extremely thermophilic and acidophilic archaeon *Sulfolobus*. *Virology* 272:409-16.
8. Arnold, H. P., W. Zillig, U. Ziese, I. Holz, M. Crosby, T. Utterback, J. F. Weidmann, J. K. Kristjanson, H. P. Klenk, K. E. Nelson, and C. M. Fraser. 2000. A novel Lipothrixvirus, SIFV, of the extremely thermophilic crenarchaeon *Sulfolobus*. *Virology* 267:252-266.
9. Bailey, S. 1994. The CCP4 suite - programs for protein crystallography. *Acta Cryst.* D50:760-763.
10. Bamford, D. H. 2003. Do viruses form lineages across different domains of life? *Research in Microbiology* 154:231-236.
11. Bamford, D. H., J. M. Grimes, and D. I. Stuart. 2005. What does structure tell us about virus evolution? *Current Opinion in Structural Biology* 15:655-663.

12. Barns, S. M., C. F. Delwiche, J. D. Palmer, and N. R. Pace. 1996. Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. *PNAS* 93:9188-9193.
13. Beeby, M., B. D. Connor, C. Ryttersgaard, D. R. Boutz, L. J. Perry, and T. O. Yeates. 2005. The genomics of disulfide bonding and protein stabilization in thermophiles. *PLoS Biology* 3:1549-1558.
14. Bendtsen, J. D., H. Nielsen, G. von Heijne, and S. Brunak. 2004. Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* 340:783-95.
15. Benson, S. D., J. K. H. Bamford, D. H. Bamford, and R. M. Burnett. 2004. Does common architecture reveal a viral lineage spanning all three domains of life? *Molecular Cell* 16:673-685.
16. Benson, S. D., J. K. H. Bamford, D. H. Bamford, and R. M. Burnett. 1999. Viral evolution revealed by bacteriophage PRD1 and human adenovirus coat protein structures. *Cell* 98:825-833.
17. Bessette, P. H., F. Aslund, J. Beckwith, and G. Georgiou. 1999. Efficient folding of proteins with multiple disulfide bonds in the *Escherichia coli* cytoplasm. *PNAS* 96:13703-13708.
18. Bettstetter, M., X. Peng, R. A. Garrett, and D. Prangishvili. 2003. AFV1, a novel virus infecting hyperthermophilic Archaea of the genus *Acidianus*. *Virology* 315:68-79.
19. Blackwell, T. K., and H. Weintraub. 1990. Differences and similarities in DNA-binding preferences of MyoD and E2A protein complexes revealed by binding site selection. *Science* 250:1104-1110.
20. Bradford, M. M. 1976. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal Biochem.* 72:248-54.
21. Brennan, R. G. 1993. The winged-helix DNA-binding motif: Another helix-turn-helix takeoff. *Cell* 74:773-776.
22. Breton, C., L. Snajdrova, C. Jeanneau, J. Koca, and A. Imberty. 2006. Structures and mechanisms of glycosyltransferases superfamily. *Glycobiology* 16:29R-37R.
23. Buehner, M., G. C. Ford, D. Moras, K. W. Olsen, and M. G. Rossmann. 1973. D-glyceraldehyde-3-phosphate dehydrogenase: three-dimensional structure and evolutionary significance. *PNAS* 70:3052-3054.

24. Cacciapuoti, G., C. Bertoldo, A. Brio, V. Zappia, and M. Porcelli. 2003. Purification and characterization of 5'-methylthioadenosine phosphorylase from the hyperthermophilic archaeon *Pyrococcus furiosus*: substrate specificity and primary structure analysis. *Extremophiles* 7:159-68.
25. Cacciapuoti, G., M. A. Moretti, S. Forte, A. Brio, L. Camardella, V. Zappia, and M. Porcelli. 2004. Methylthioadenosine phosphorylase from the archaeon *Pyrococcus furiosus*. Mechanism of the reaction and assignment of disulfide bonds. *Eur J Biochem* 271:4834-44.
26. Cacciapuoti, G., M. Porcelli, C. Bertoldo, M. De Rosa, and V. Zappia. 1994. 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* 269:24762-9.
27. Charnock, S. J., and G. J. Davies. 1999. Structure of the nucleotide-diphospho-sugar transferase, SpsA from *Bacillus subtilis*, in native and nucleotide-complexed forms. *Biochemistry* 38:6380-6385.
28. Chiu, C. P. C., A. G. Watts, L. L. Lairson, M. Gilbert, D. Lim, W. W. Wakarchuk, S. G. Withers, and N. C. J. Strynadka. 2004. Structural analysis of the sialyltransferase CstII from *Campylobacter jejuni* in complex with a substrate analog. *Nat Struct Mol Biol* 11:163-170.
29. Christopher, J. A. 1998. SPOCK: The Structural Properties Observation and Calculation Kit. <http://quorum.tamu.edu/>.
30. Coutinho, P. M., E. Deleury, G. J. Davies, and B. Henrissat. 2003. An evolving hierarchical family classification for glycosyltransferases. *Journal of Molecular Biology* 328:307-317.
31. Coutinho, P. M., and B. Henrissat. 1999. Carbohydrate-Active Enzymes server at URL: <http://afmb.cnrs-mrs.fr/CAZY/>.
32. Cowtan, K. 1994. DM: An automated procedure for phase improvement by density modification. *Joint CCP4 and ESF-EACBM Newsletter on Protein Crystallography* 31:34-38.
33. Davies, G. J., T. M. Gloster, and B. Henrissat. 2005. Recent structural insights into the expanding world of carbohydrate-active enzymes. *Current Opinion in Structural Biology* 15:637-645.
34. DeDecker, B. S., R. O'Brien, P. J. Fleming, J. H. Geiger, S. P. Jackson, and P. B. Sigler. 1996. The crystal structure of a hyperthermophilic archaeal TATA-box binding protein. *J Mol Biol* 264:1072-84.

35. DeLano, W. L. 2002. The PyMOL Molecular Graphics System. <http://www.pymol.org>.
36. DeLong, E. F. 1992. Archaea in coastal marine environments. PNAS 89:5685-5689.
37. Dyall-Smith, M., S.-L. Tang, and C. Bath. 2003. Haloarchaeal viruses: how diverse are they? Research in Microbiology 154:309-313.
38. Eichler, J., and M. W. W. Adams. 2005. Posttranslational protein modification in Archaea. Microbiol. Mol. Biol. Rev. 69:393-425.
39. Emsley, P., and K. Cowtan. 2004. Coot: model-building tools for molecular graphics. Acta Crystallographica Section D 60:2126-2132.
40. Feese, M. D., Y. Kato, T. Tamada, M. Kato, T. Komeda, Y. Miura, M. Hirose, K. Hondo, K. Kobayashi, and R. Kuroki. 2000. Crystal structure of glycosyltrehalose trehalohydrolase from the hyperthermophilic archaeum *Sulfolobus solfataricus*. J Mol Biol 301:451-64.
41. Flint, J., E. Taylor, M. Yang, D. N. Bolam, L. E. Tailford, C. Martinez-Fleites, E. J. Dodson, B. G. Davis, H. J. Gilbert, and G. J. Davies. 2005. Structural dissection and high-throughput screening of mannosylglycerate synthase. Nat Struct Mol Biol 12:608-614.
42. Fox, G. E., L. J. Magrum, W. E. Balch, R. S. Wolfe, and C. R. Woese. 1977. Classification of methanogenic Bacteria by 16S ribosomal RNA characterization. PNAS 74:4537-4541.
43. Franco, O. L., and D. J. Rigden. 2003. Fold recognition analysis of glycosyltransferase families: further members of structural superfamilies. Glycobiology 13:707-712.
44. Gastinel, L., C. Cambillau, and Y. Bourne. 1999. Crystal structures of the bovine β 4galactosyltransferase catalytic domain and its complex with uridine diphosphogalactose. EMBO J. 18:3546-3557.
45. Gattiker, A., E. Gasteiger, and A. Bairoch. 2002. ScanProsite: a reference implementation of a PROSITE scanning tool. Appl Bioinformatics 1:107-8.
46. Gibrat, J.-F., T. Madej, and S. H. Bryant. 1996. Surprising similarities in structure comparison. Current Opinion in Structural Biology 6:377-385.
47. Glaser, F., T. Pupko, I. Paz, R. E. Bell, D. Bechor-Shental, E. Martz, and N. Ben-Tal. 2003. ConSurf: identification of functional regions in proteins by surface-mapping of phylogenetic Information. Bioinformatics 19:163-164.

48. Graves, M. V., C. T. Bernadt, R. Cerny, and J. L. Van Etten. 2001. Molecular and genetic evidence for a virus-encoded glycosyltransferase involved in protein glycosylation. *Virology* 285:332-345.
49. Haring, M., X. Peng, K. Brugger, R. Rachel, K. O. Stetter, R. A. Garrett, and D. Prangishvili. 2004. Morphology and genome organization of the virus PSV of the hyperthermophilic archaeal genera *Pyrobaculum* and *Thermoproteus*: a novel virus family, the Globuloviridae. *Virology* 323:233-42.
50. Haring, M., R. Rachel, X. Peng, R. A. Garrett, and D. Prangishvili. 2005. 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* 79:9904-11.
51. Haring, M., G. Vestergaard, K. Brugger, R. Rachel, R. A. Garrett, and D. Prangishvili. 2005. Structure and genome organization of AFV2, a novel Archaeal Lipothrixvirus with unusual terminal and core structures. *J. Bacteriol.* 187:3855-3858.
52. Haring, M., G. Vestergaard, R. Rachel, L. Chen, R. A. Garrett, and D. Prangishvili. 2005. Independent virus development outside a host. *Nature* 436:1101-1102.
53. Hettmann, T., C. L. Schmidt, S. Anemuller, U. Zahringer, H. Moll, A. Petersen, and G. Schafer. 1998. Cytochrome *b*_{558/566} from the Archaeon *Sulfolobus acidocaldarius*. A novel highly glycosylated, membrane-bound b-type hemoprotein. *J. Biol. Chem.* 273:12032-12040.
54. Holm, L., and C. Sander. 1993. Protein structure comparison by alignment of distance matrices. *J Mol Biol* 233:123-38.
55. Huber, H., M. J. Hohn, R. Rachel, T. Fuchs, V. C. Wimmer, and K. O. Stetter. 2002. A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. *Nature* 417:63-67.
56. Hulo, N., A. Bairoch, V. Bulliard, L. Cerutti, E. De Castro, P. S. Langendijk-Genevaux, M. Pagni, and C. J. Sigrist. 2006. The PROSITE database. *Nucleic Acids Res* 34:D227-30.
57. Invitrogen Corporation. 2003. Gateway Technology: a universal technology to clone DNA sequences for functional analysis and expression in multiple systems.
58. Iyer, L. M., K. S. Makarova, E. V. Koonin, and L. Aravind. 2004. Comparative genomics of the FtsK-HerA superfamily of pumping ATPases: implications for the origins of chromosome segregation, cell division and viral capsid packaging. *Nucl. Acids Res.* 32:5260-5279.

59. Jones, T. A., J. Y. Zou, S. W. Cowan, and M. Kjeldgaard. 1991. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr.* A47:110-119.
60. Jonuscheit, M., E. Martusewitsch, K. M. Stedman, and C. Schleper. 2003. A reporter gene system for the hyperthermophilic archaeon *Sulfolobus solfataricus* based on a selectable and integrative shuttle vector. *Mol Microbiol* 48:1241-52.
61. Kabsch, W. 1976. A solution for the best rotation to relate two sets of vectors. *Acta Cryst.* A32:922-923.
62. Kadokura, H., F. Katzen, and J. Beckwith. 2003. Protein disulfide bond formation in prokaryotes. *Annu Rev Biochem* 72:111-35.
63. Karlstrom, M., R. Stokke, I. H. Steen, N. K. Birkeland, and R. Ladenstein. 2005. Isocitrate dehydrogenase from the hyperthermophile *Aeropyrum pernix*: X-ray structure analysis of a ternary enzyme-substrate complex and thermal stability. *J Mol Biol* 345:559-77.
64. Kelley, L. A., R. M. MacCallum, and M. J. E. Sternberg. 2000. Enhanced genome annotation using structural profiles in the program 3D-PSSM. *Journal of Molecular Biology* 299:499-520.
65. Khayat, R., L. Tang, E. T. Larson, C. M. Lawrence, M. Young, and J. E. Johnson. 2005. Structure of an archaeal virus capsid protein reveals a common ancestry to eukaryotic and bacterial viruses. *PNAS* 102:18944-18949.
66. Kleywegt, G. 1996. Use of non-crystallographic symmetry in protein structure refinement. *Acta Crystallographica Section D* 52:842-857.
67. Kraft, P., D. Kummel, A. Oeckinghaus, G. H. Gauss, B. Wiedenheft, M. Young, and C. M. Lawrence. 2004. 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.* 78:7438-7442.
68. Kraft, P., A. Oeckinghaus, D. Kummel, G. H. Gauss, J. Gilmore, B. Wiedenheft, M. Young, and C. M. Lawrence. 2004. Crystal structure of F-93 from *Sulfolobus* Spindle-Shaped Virus 1, a winged-helix DNA binding protein. *J. Virol.* 78:11544-11550.
69. Krogh, A., B. Larsson, G. von Heijne, and E. L. Sonnhammer. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* 305:567-80.

70. Kunst, F., N. Ogasawara, I. Moszer, A. M. Albertini, G. Alloni, V. Azevedo, M. G. Bertero, P. Bessieres, A. Bolotin, S. Borchert, R. Borriss, L. Boursier, A. Brans, M. Braun, S. C. Brignell, S. Bron, S. Brouillet, C. V. Bruschi, B. Caldwell, V. Capuano, N. M. Carter, S. K. Choi, J. J. Codani, I. F. Connerton, N. J. Cummings, R. A. Daniel, F. Denizot, K. M. Devine, A. Dusterhoft, S. D. Ehrlich, P. T. Emmerson, K. D. Entian, J. Errington, C. Fabret, E. Ferrari, D. Foulger, C. Fritz, M. Fujita, Y. Fujita, S. Fuma, A. Galizzi, N. Galleron, S. Y. Ghim, P. Glaser, A. Goffeau, E. J. Golightly, G. Grandi, G. Guiseppi, B. J. Guy, K. Haga, J. Haiech, C. R. Harwood, A. Henaut, H. Hilbert, S. Holsappel, S. Hosono, M. F. Hullo, M. Itaya, L. Jones, B. Joris, D. Karamata, Y. Kasahara, M. Klaerr-Blanchard, C. Klein, Y. Kobayashi, P. Koetter, G. Koningstein, S. Krogh, M. Kumano, K. Kurita, A. Lapidus, S. Lardinois, J. Lauber, V. Lazarevic, S. M. Lee, A. Levine, H. Liu, S. Masuda, C. Mauel, C. Medigue, N. Medina, R. P. Mellado, M. Mizuno, D. Moestl, S. Nakai, M. Noback, D. Noone, M. O'Reilly, K. Ogawa, A. Ogiwara, B. Oudega, S. H. Park, V. Parro, T. M. Pohl, D. Portetelle, S. Porwollik, A. M. Prescott, E. Presecan, P. Pujic, B. Purnelle, et al. 1997. The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. *Nature* 390:249-256.
71. Landau, M., I. Mayrose, Y. Rosenberg, F. Glaser, E. Martz, T. Pupko, and N. Ben-Tal. 2005. ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res.* 33:W299-W302.
72. Larson, E. T., D. Reiter, M. Young, and C. M. Lawrence. 2006. Structure of A197 from *Sulfolobus* turreted icosahedral virus: a crenarchaeal viral glycosyltransferase exhibiting the GT-A fold. *J. Virol.* 80:7636-7644.
73. Larson, E. T., D. Reiter, M. J. Young, and C. M. Lawrence. Submitted. A winged-helix protein from *Sulfolobus* Turreted Icosahedral Virus points toward stabilizing disulfide bonds in the intracellular proteins of a hyperthermophilic virus.
74. Laskowski, R. A., M. W. MacArthur, D. S. Moss, and J. M. Thornton. 1993. PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Cryst.* 26:283-291.
75. Li, X., P. Romero, M. Rani, A. K. Dunker, and Z. Obradovic. 1999. Predicting protein disorder for N-, C-, and internal regions. *Genome Informatics* 10:30-40.
76. Liang, J., H. Edelsbrunner, and C. Woodward. 1998. Anatomy of protein pockets and cavities: Measurement of binding site geometry and implications for ligand design. *Protein Sci* 7:1884-1897.
77. Luft, J., R. Collins, N. Fehrman, A. Lauricella, C. Veatch, and G. DeTitta. 2003. A deliberate approach to screening for initial crystallization conditions of biological macromolecules. *Journal of Structural Biology* 142:170-179.

78. Maaty, W. S. A., A. C. Ortmann, M. Dlakic, K. Schulstad, J. K. Hilmer, L. Liepold, B. Weidenheft, R. Khayat, T. Douglas, M. J. Young, and B. Bothner. 2006. 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.* 80:7625-7635.
79. Mallick, P., D. R. Boutz, D. Eisenberg, and T. O. Yeates. 2002. Genomic evidence that the intracellular proteins of archaeal microbes contain disulfide bonds. *PNAS* 99:9679-9684.
80. Marchler-Bauer, A., and S. H. Bryant. 2004. CD-Search: protein domain annotations on the fly. *Nucleic Acids Res.* 32:W327-331.
81. Markillie, L. M., C. T. Lin, J. N. Adkins, D. L. Auberry, E. A. Hill, B. S. Hooker, P. A. Moore, R. J. Moore, L. Shi, H. S. Wiley, and V. Kery. 2005. Simple protein complex purification and identification method for high-throughput mapping of protein interaction networks. *J. Proteome Res.* 4:268-274.
82. Markine-Goriaynoff, N., L. Gillet, J. L. Van Etten, H. Korres, N. Verma, and A. Vanderplasschen. 2004. Glycosyltransferases encoded by viruses. *J Gen Virol* 85:2741-2754.
83. Masip, L., J. L. Pan, S. Haldar, J. E. Penner-Hahn, M. P. DeLisa, G. Georgiou, J. C. A. Bardwell, and J.-F. Collet. 2004. An engineered pathway for the formation of protein disulfide bonds. *Science* 303:1185-1189.
84. Mengele, R., and M. Sumper. 1992. Drastic differences in glycosylation of related S-layer glycoproteins from moderate and extreme halophiles. *J. Biol. Chem.* 267:8182-8185.
85. Moult, J., and E. Melamud. 2000. From fold to function. *Current Opinion in Structural Biology* 10:384-389.
86. Murshudov, G. N., A. A. Vagin, and E. J. Dodson. 1997. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Cryst.* D53:240-255.
87. Nandhagopal, N., A. A. Simpson, J. R. Gurnon, X. Yan, T. S. Baker, M. V. Graves, J. L. Van Etten, and M. G. Rossmann. 2002. The structure and evolution of the major capsid protein of a large, lipid-containing DNA virus. *PNAS* 99:14758-14763.
88. O'Connor, B. D., and T. O. Yeates. 2004. GDAP: a web tool for genome-wide protein disulfide bond prediction. *Nucl. Acids Res.* 32:W360-364.

89. Olsen, G. J., and C. R. Woese. 1997. Archaeal genomics: an overview. *Cell* 89:991-994.
90. Ortmann, A. C., B. Wiedenheft, T. Douglas, and M. Young. 2006. Hot crenarchaeal viruses reveal deep evolutionary connections. *Nat Rev Microbiol* 4:520-8.
91. Otwinowski, Z., and W. Minor. 1997. Processing of X-ray diffraction data collected in oscillation mode, p. 307-326. *In* C. Carter and R. Sweet (ed.), *Macromolecular Crystallography, Part A*, vol. 276. Academic Press.
92. Painter, J., and E. A. Merritt. 2005. A molecular viewer for the analysis of TLS rigid-body motion in macromolecules. *Acta Cryst.* D61:465-471.
93. Painter, J., and E. A. Merritt. 2006. Optimal description of a protein structure in terms of multiple groups undergoing TLS motion. *Acta Crystallographica Section D* 62:439-450.
94. Painter, J., and E. A. Merritt. 2006. TLSMD web server for the generation of multi-group TLS models. *Journal of Applied Crystallography* 39:109-111.
95. Palm, P., C. Schleper, B. Grampp, S. Yeats, P. McWilliam, W. D. Reiter, and W. Zillig. 1991. Complete nucleotide sequence of the virus SSV1 of the archaeobacterium *Sulfolobus shibatae*. *Virology* 185:242-50.
96. Pedersen, L. C., T. A. Darden, and M. Negishi. 2002. Crystal structure of β 1,3-glucuronyltransferase I in complex with active donor substrate UDP-GlcUA. *J. Biol. Chem.* 277:21869-21873.
97. Pedersen, L. C., K. Tsuchida, H. Kitagawa, K. Sugahara, T. A. Darden, and M. Negishi. 2000. Heparan/chondroitin sulfate biosynthesis. Structure and mechanism of human glucuronyltransferase I. *J. Biol. Chem.* 275:34580-34585.
98. Peng, X., H. Blum, Q. She, S. Mallok, K. Brugger, R. A. Garrett, W. Zillig, and D. Prangishvili. 2001. Sequences and replication of genomes of the archaeal Rudiviruses SIRV1 and SIRV2: Relationships to the archaeal Lipothrixvirus SIFV and some eukaryal viruses. *Virology* 291:226-234.
99. Peng, X., A. Kessler, H. Phan, R. A. Garrett, and D. Prangishvili. 2004. Multiple variants of the archaeal DNA rudivirus SIRV1 in a single host and a novel mechanism of genomic variation. *Molecular Microbiology* 54:366-375.
100. Poirot, O., K. Suhre, C. Abergel, E. O'Toole, and C. Notredame. 2004. 3DCoffee@igs: a web server for combining sequences and structures into a multiple sequence alignment. *Nucl. Acids Res.* 32:W37-40.

101. Porter, K., P. Kukkaro, J. K. H. Bamford, C. Bath, H. M. KivelÃ¶, M. L. Dyll-Smith, and D. H. Bamford. 2005. SH1: A novel, spherical halovirus isolated from an Australian hypersaline lake. *Virology* 335:22-33.
102. Prangishvili, D., H. P. Arnold, D. Gotz, U. Ziese, I. Holz, J. K. Kristjansson, and W. Zillig. 1999. A novel virus family, the Rudiviridae: structure, virus-vost interactions and genome variability of the *Sulfolobus* viruses SIRV1 and SIRV2. *Genetics* 152:1387-1396.
103. Prangishvili, D., and R. A. Garrett. 2004. Exceptionally diverse morphotypes and genomes of crenarchaeal hyperthermophilic viruses. *Biochem Soc Trans* 32:204-8.
104. Prangishvili, D., and R. A. Garrett. 2005. Viruses of hyperthermophilic Crenarchaea. *Trends in Microbiology* 13:535-542.
105. Prangishvili, D., R. A. Garrett, and E. V. Koonin. 2006. Evolutionary genomics of archaeal viruses: Unique viral genomes in the third domain of life. *Virus Research* 117:52-67.
106. Prangishvili, D., K. Stedman, and W. Zillig. 2001. Viruses of the extremely thermophilic archaeon *Sulfolobus*. *Trends in Microbiology* 9:39-43.
107. Prieur, D., G. Erauso, C. Geslin, S. Lucas, M. Gaillard, A. Bidault, A.-C. Mattenet, K. Rouault, D. Flament, P. Forterre, and M. L. Romancer. 2004. Genetic elements of Thermococcales. *Biochemical Society Transactions* 32:184-187.
108. Prinz, W. A., F. Aslund, A. Holmgren, and J. Beckwith. 1997. The role of the thioredoxin and glutaredoxin pathways in reducing protein disulfide bonds in the *Escherichia coli* cytoplasm. *J. Biol. Chem.* 272:15661-15667.
109. R & D Laboratory - New Brunswick Scientific. 1996. Fundamentals of Fermentation: Techniques For Benchtop Fermentors Part I - *E. coli*, Edison, NJ.
110. Rachel, R., M. Bettstetter, B. P. Hedlund, M. Haring, A. Kessler, K. O. Stetter, and D. Prangishvili. 2002. Remarkable morphological diversity of viruses and virus-like particles in hot terrestrial environments. *Archives of Virology* 147:2419-2429.
111. Rice, G., K. Stedman, J. Snyder, B. Wiedenheft, D. Willits, S. Brumfield, T. McDermott, and M. J. Young. 2001. Viruses from extreme thermal environments. *PNAS* 98:13341-13345.

112. Rice, G., L. Tang, K. Stedman, F. Roberto, J. Spuhler, E. Gillitzer, J. E. Johnson, T. Douglas, and M. Young. 2004. The structure of a thermophilic archaeal virus shows a double-stranded DNA viral capsid type that spans all domains of life. *PNAS* 101:7716-7720.
113. Richardson, J. S. 1981. The anatomy and taxonomy of protein structure. *Adv Protein Chem* 34:167-339.
114. Romero, P., Z. Obradovic, and A. K. Dunker. 1997. Sequence data analysis for long disordered regions prediction in the calcineurin family. *Genome Informatics* 8:110-124.
115. Romero, P., Z. Obradovic, X. Li, E. Garner, C. Brown, and A. K. Dunker. 2001. Sequence complexity of disordered protein. *Proteins: Struct. Funct. Gen.* 42:38-48.
116. Rossmann, M. G., A. D. B. Malcolm, and W. F. Bodmer. 1981. Evolution of glycolytic enzymes [and discussion]. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 293:191-203.
117. Rux, J. J., P. R. Kuser, and R. M. Burnett. 2003. Structural and phylogenetic analysis of adenovirus hexons by use of high-resolution x-ray crystallographic, molecular modeling, and sequence-based methods. *J. Virol.* 77:9553-9566.
118. Safo, M. K., Q. Zhao, T.-P. Ko, F. N. Musayev, H. Robinson, N. Scarsdale, A. H. J. Wang, and G. L. Archer. 2005. Crystal structures of the *BlaI* repressor from *Staphylococcus aureus* and its complex with DNA: Insights into transcriptional regulation of the *bla* and *mec* operons. *J. Bacteriol.* 187:1833-1844.
119. Schneider, K. L., K. S. Pollard, R. Baertsch, A. Pohl, and T. M. Lowe. 2006. The UCSC archaeal genome browser. *Nucl. Acids Res.* 34:D407-410.
120. Schulz, G. E. 1979. Nucleotide Binding Proteins, p. 79-94. *In* M. Balaban (ed.), *Molecular Mechanism of Biological Recognition*. Elsevier/North-Holland, New York.
121. She, Q., R. K. Singh, F. Confalonieri, Y. Zivanovic, G. Allard, M. J. Awayez, C. C. Y. Chan-Weiher, I. G. Clausen, B. A. Curtis, A. De Moors, G. Erauso, C. Fletcher, P. M. K. Gordon, I. Heikamp-de Jong, A. C. Jeffries, C. J. Kozera, N. Medina, X. Peng, H. P. Thi-Ngoc, P. Redder, M. E. Schenk, C. Theriault, N. Tolstrup, R. L. Charlebois, W. F. Doolittle, M. Duguet, T. Gaasterland, R. A. Garrett, M. A. Ragan, C. W. Sensen, and J. Van der Oost. 2001. The complete genome of the crenarchaeon *Sulfolobus solfataricus* P2. *PNAS* 98:7835-7840.

122. Sigrist, C. J., L. Cerutti, N. Hulo, A. Gattiker, L. Falquet, M. Pagni, A. Bairoch, and P. Bucher. 2002. PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief Bioinform* 3:265-74.
123. Snyder, J. C., J. Spuhler, B. Wiedenheft, F. F. Roberto, T. Douglas, and M. J. Young. 2004. Effects of culturing on the population structure of a hyperthermophilic virus. *Microbial Ecology* 48:561-566.
124. Snyder, J. C., K. Stedman, G. Rice, B. Wiedenheft, J. Spuhler, and M. J. Young. 2003. Viruses of hyperthermophilic Archaea. *Research in Microbiology* 154:474-482.
125. Sonnhammer, E. L., G. von Heijne, and A. Krogh. 1998. A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc Int Conf Intell Syst Mol Biol* 6:175-82.
126. Stedman, K. M., C. Schleper, E. Rumpf, and W. Zillig. 1999. Genetic requirements for the function of the archaeal virus SSV1 in *Sulfolobus solfataricus*: construction and testing of viral shuttle vectors. *Genetics* 152:1397-1405.
127. Terwilliger, T. C. 2002. Automated main-chain model-building by template-matching and iterative fragment extension. *Acta Cryst D* 59:34-44.
128. Terwilliger, T. C. 2000. Maximum likelihood density modification. *Acta Cryst D* 56:965-972.
129. Terwilliger, T. C., and J. Berendzen. 1999. Automated MAD and MIR structure solution. *Acta Crystallogr. D* 55:849-861.
130. Thoma, R., G. Obmolova, D. A. Lang, M. Schwander, P. Jenö, R. Sterner, and M. Wilmanns. 1999. Efficient expression, purification and crystallisation of two hyperthermostable enzymes of histidine biosynthesis. *FEBS Letters* 454:1-6.
131. Thompson, J. D., Gibson, T.J., Plewniak, F., Jeanmougin, F. and Higgins, D.G. 1997. The ClustalX windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research* 24:4876-4882.
132. Thompson, J. D., Higgins, D.G. and Gibson, T.J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucl. Acids Res.* 22:4673-4680.

133. Toth, E. A., C. Worby, J. E. Dixon, E. R. Goedken, S. Marqusee, and T. O. Yeates. 2000. The crystal structure of adenylosuccinate lyase from *Pyrobaculum aerophilum* reveals an intracellular protein with three disulfide bonds. *J Mol Biol* 301:433-50.
134. Tuerk, C., and L. Gold. 1990. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* 249:505-510.
135. Unligil, U. M., and J. M. Rini. 2000. Glycosyltransferase structure and mechanism. *Current Opinion in Structural Biology* 10:510-517.
136. Unligil, U. M., S. Zhou, S. Yuwaraj, M. Sarkar, H. Schachter, and J. M. Rini. 2000. X-ray crystal structure of rabbit N-acetylglucosaminyltransferase I: catalytic mechanism and a new protein superfamily. *EMBO J.* 19:5269-5280.
137. Vestergaard, G., M. Haring, X. Peng, R. Rachel, R. A. Garrett, and D. Prangishvili. 2005. A novel rudivirus, ARV1, of the hyperthermophilic archaeal genus *Acidianus*. *Virology* 336:83-92.
138. Wang, C., M. Eufemi, C. Turano, and A. Giartosio. 1996. Influence of the carbohydrate moiety on the stability of glycoproteins. *Biochemistry* 35:7299-7307.
139. Wiedenheft, B., K. Stedman, F. Roberto, D. Willits, A.-K. Gleske, L. Zoeller, J. Snyder, T. Douglas, and M. Young. 2004. Comparative genomic analysis of hyperthermophilic archaeal Fuselloviridae viruses. *J. Virol.* 78:1954-1961.
140. Wilce, J. A., J. P. Vivian, A. F. Hastings, G. Otting, R. H. A. Folmer, I. G. Duggin, R. G. Wake, and M. C. J. Wilce. 2001. Structure of the RTP-DNA complex and the mechanism of polar replication fork arrest. *Nat Struct Mol Biol* 8:206-210.
141. Woese, C. R., and G. E. Fox. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *PNAS* 74:5088-5090.
142. Woese, C. R., O. Kandler, and M. L. Wheelis. 1990. Towards a Natural System of Organisms: Proposal for the Domains Archaea, Bacteria, and Eucarya. *PNAS* 87:4576-4579.
143. Xiang, X., L. Chen, X. Huang, Y. Luo, Q. She, and L. Huang. 2005. *Sulfolobus tengchongensis* Spindle-Shaped Virus STSV1: virus-host interactions and genomic features. *J. Virol.* 79:8677-8686.

144. Zahringer, U., H. Moll, T. Hettmann, Y. A. Knirel, and G. Schafer. 2000. Cytochrome $b_{558/566}$ from the archaeon *Sulfolobus acidocaldarius* has a unique Asn-linked highly branched hexasaccharide chain containing 6-sulfoquinovose. *Eur J Biochem* 267:4144-4149.
145. Zillig, W., D. Prangishvilli, C. Schleper, M. Elferink, I. Holz, S. Albers, D. Janekovic, and D. Goetz. 1996. Viruses, plasmids and other genetic elements of thermophilic and hyperthermophilic Archaea. *FEMS Microbiology Reviews* 18:225-236.