Structure *14*, 1449–1458, September 2006 ©2006 Elsevier Ltd All rights reserved DOI 10.1016/j.str.2006.07.010

The Crystal Structure of the Venezuelan Equine Encephalitis Alphavirus nsP2 Protease

Andrew T. Russo,[1] Mark A. White,[1]
and Stanley J. Watowich[1,*]
[1] Department of Biochemistry and Molecular Biology and
Sealy Center for Structural Biology and Molecular
Biophysics
The University of Texas Medical Branch
Galveston, Texas 77555

Summary

Alphavirus replication and propagation is dependent on the protease activity of the viral nsP2 protein, which cleaves the nsP1234 polyprotein replication complex into functional components. Thus, nsP2 is an attractive target for drug discovery efforts to combat highly pathogenic alphaviruses. Unfortunately, antiviral development has been hampered by a lack of structural information for the nsP2 protease. Here, we report the crystal structure of the nsP2 protease (nsP2pro) from Venezuelan equine encephalitis alphavirus determined at 2.45 Å resolution. The protease structure consists of two distinct domains. The nsP2pro N-terminal domain contains the catalytic dyad cysteine and histidine residues organized in a protein fold that differs significantly from any known cysteine protease or protein folds. The nsP2pro C-terminal domain displays structural similarity to S-adenosyl-L-methionine-dependent RNA methyltransferases and provides essential elements that contribute to substrate recognition and may also regulate the structure of the substrate binding cleft.

Introduction

Venezuelan equine encephalitis alphavirus (VEEV) is a significant cause of human and livestock disease in Central and South America, with outbreaks occasionally reaching as far north as southern Texas (Weaver et al., 1996). Human epidemics reported in 1995 in Venezuela and Colombia caused widespread illness with mortality rates approaching 1% (Weaver et al., 1996). Several nations, including the United States and the former Soviet Union, have reportedly developed weapons-grade VEEV (Bronze et al., 2002), and stockpiles of such weapons may still exist. Consequently, VEEV is classified as a select agent and a National Institutes of Health Category B priority pathogen. Existing health care resources are poorly prepared to deal with an outbreak of VEEV. No antiviral drugs exist to treat VEEV infection. Moreover, the VEEV TC-83 vaccine strain provides only partial protection against infection and is not approved for general human immunization (Weaver et al., 1999).

VEEV is an enveloped, positive-sense ssRNA virus representative of the family *Togaviridae* and genus *Alphaviridae* (Strauss and Strauss, 1994). After infection, alphavirus RNA is directly translated to typically produce polyprotein nsP1234 containing nonstructural proteins nsP1–nsP4 (Strauss and Strauss, 1994). In VEEV, an opal codon between *nsp3* and *nsp4* results in the expression of polyprotein nsP123 containing nsP1, nsP2, and nsP3; polyprotein nsP1234 is produced by readthrough of the opal codon (Feng et al., 1990). These polyproteins form the viral replication complex and are processed by the proteolytic activity of nsP2. In the late stage of infection, a positive-sense 26S subgenomic RNA is synthesized. Translation of the 26S RNA produces a structural polyprotein that is subsequently processed into individual structural proteins by a combination of viral and host proteases in the endoplasmic reticulum.

Structural characterization of alphaviruses includes cryoelectron microscopy image reconstructions of infectious particles determined to 9 Å resolution for Sindbis virus (Mukhopadhyay et al., 2006) and to 8.5 Å resolution for VEEV (Z. Li, personal communication). Crystal structures of the C-terminal domain of the capsid protein of Sindbis (Choi et al., 1991, 1996; Tong et al., 1992), Semliki Forest (Choi et al., 1997), and VEEV (S.J.W., unpublished data; Protein Data Bank [PDB] code: 1EP5) and the soluble ectodomain of Semliki Forest virus (SFV) E1 glycoprotein (Lescar et al., 2001) have been solved to atomic resolution. No structures have been solved of the alphavirus nonstructural proteins, but sufficient structural similarity with other proteins exists to permit construction of homology models for the central third of nsP2, the N-terminal domain of nsP3, and the central region of nsP4. The lack of high-resolution structures for alphavirus replication complex proteins prevents the use of powerful structure-based drug discovery and design methodologies to combat VEEV and alphavirus infections.

The alphavirus nsP2 protein has multiple enzymatic activities. A 456 amino acid N-terminal region (Gly1–Ile456, VEEV numbering) has been shown to possess ATPase and GTPase activity (Rikkonen et al., 1994), RNA helicase activity (Gomez de Cedron et al., 1999), and RNA 5′-triphosphatase activity (Vasiljeva et al., 2000). The 338 amino acid C-terminal region of nsP2 (Met457–Cys794, VEEV numbering) has been associated with regulating the 26S subgenomic RNA synthesis (Suopanki et al., 1998), downregulating minus-strand RNA synthesis late in infection (Sawicki et al., 2006; Sawicki and Sawicki, 1993), targeting nsP2 for nuclear transport (Peranen et al., 1990), and proteolytic processing of the alphavirus nonstructural polyprotein replication complex (Vasiljeva et al., 2001, 2003). Sequence analysis suggests that alphavirus nsP2 proteases are cysteine proteases and members of peptidase family C9 of clan CA (Rawlings et al., 2006). The nsP2 protease is an attractive target for antiviral therapeutics since it cleaves substrates with defined recognition sequences (Asp/Glu-Ala-Gly-Ala or Glu-Ala-Gly-Cys in VEEV) (Strauss and Strauss, 1994) and is required for alphavirus replication. Structural information for the nsP2 protease would greatly facilitate drug discovery and development efforts for VEEV and related alphaviruses. In

*Correspondence: watowich@xray.utmb.edu

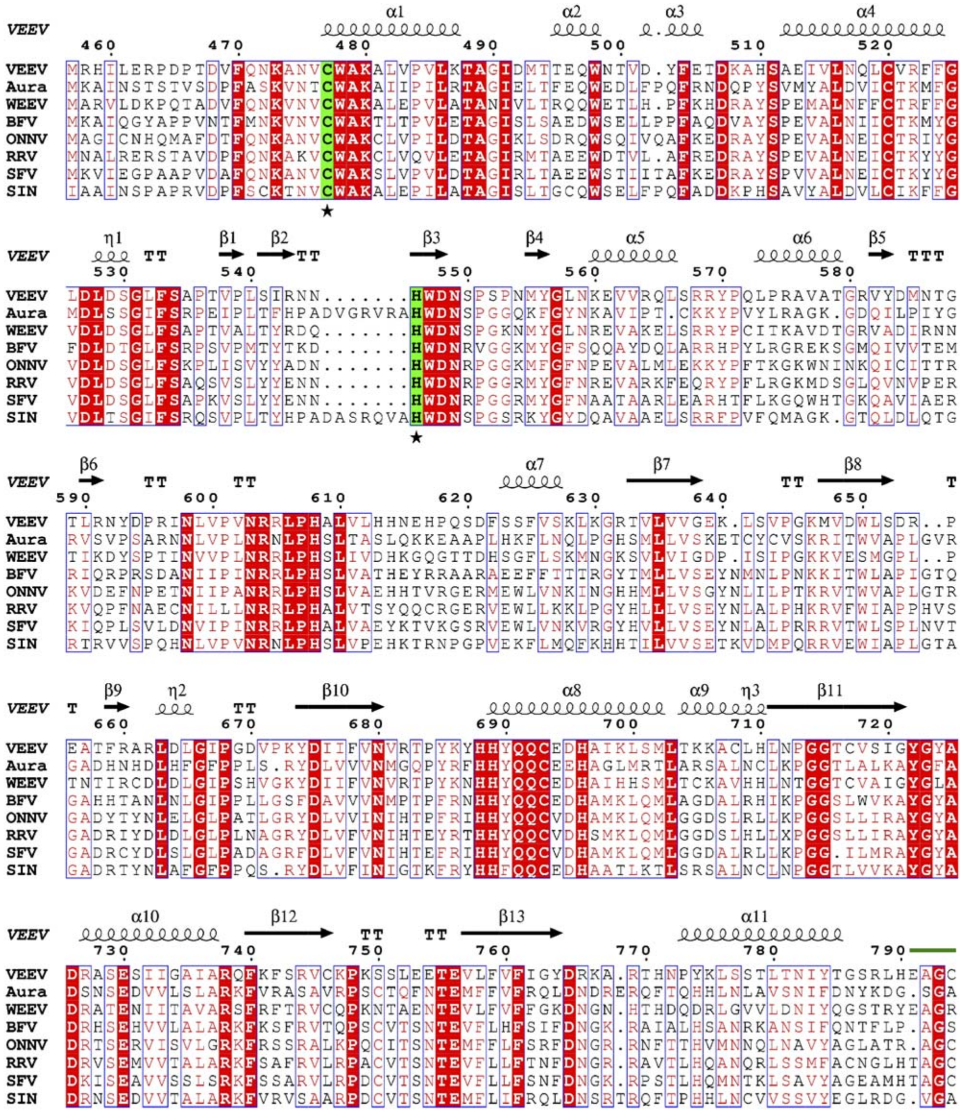

Figure 1. Sequence Alignment of nsP2pro from Representative Alphaviruses

Active site residues are highlighted in green and are marked with a star below the alignment. The nsP2/3 cleavage site residues (Glu-Ala-Gly-Cys) are indicated by a green bar above the alignment. Secondary structure annotation, based on the VEEV structure, is indicated above the alignment. Sequences used in this alignment are: VEEV (Venezuelan equine encephalitis virus; Kinney et al., 1989), Aura (Aura virus; Rumenapf et al., 1995), WEEV (Western equine encephalitis virus 5614; Uryvaev et al., 1994), BFV (Barmah Forest virus; Lee et al., 1997), ONNV (O'nyong-nyong virus SG650; Lanciotti et al., 1998), RRV (Ross River virus NB5092; Faragher et al., 1988), SFV (Semliki Forest virus; Salonen et al., 2003), and SINV (Sindbis virus MRE16; Myles et al., 2003). Sequences were aligned with Megalign (DNASTAR, Inc.). The figure was prepared with ESPript (Gouet et al., 1999).

pursuit of this goal, we report the crystal structure of the C-terminal region of VEEV nsP2 (nsP2pro). This structure consists of two domains: a novel cysteine protease domain, followed by a methyltransferase-like domain of unknown function. Both residues of the catalytic dyad, Cys477 and His546, are located in the N-terminal domain, which is largely helical. The active site is positioned adjacent to the interface between domains, and both domains contribute to substrate recognition. However, the majority of residues involved in substrate binding are from the N-terminal domain. The effects of temperature-sensitive mutants in highly conserved residues identified in Sindbis and Semliki Forest virus alphaviruses (Agapov et al., 1998; Hahn et al., 1989; Lulla et al., 2006a; Suopanki et al., 1998) are explained. Also, explanations for the role of nsP2 in alphavirus RNA replication are proposed based on the structural similarity of the C-terminal domain to known RNA binding methyltransferases and the location in the structure of temperature-sensitive mutations known to affect RNA synthesis and processing.

Results and Discussion

Overall Structure

The VEEV nsP2 region chosen for structural studies was based on sequence alignment (Figure 1) between VEE and SFVs and on previous functional mapping of SFV

Table 1. X-Ray Crystallographic Data Collection and Processing Statistics

	Native	KAu(CN)$_2$	Thimerosal	Terbium
Wavelength (Å)	1.3808	1.5418	1.5418	1.5418
Space group	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$	$P2_12_12_1$
Unit cell parameters				
a (Å)	47.5	47.3	47.5	47.3
b (Å)	72.6	72.7	72.9	72.8
c (Å)	106.7	105.6	105.6	105.9
Resolution limits (Å)	40–2.45	40.0–3.0	40.0–3.0	40.0–3.0
Completeness (%)	99.7 (100)	76.7 (81.0)	91.8 (95.7)	100 (100)
<I/σI>	14.2 (3.1)	7.8 (2.8)	6.1 (2.2)	10.8 (4.4)
R_{sym}[a] (%)	9.4 (34.0)	18.0 (44.4)	16.1 (44.4)	16.8 (41.4)
Phasing power[b] (acentric/centric)		1.27/1.13	0.96/0.78	1.41/1.06
R_{Cullis}[c] (acentric/centric)		0.45/0.45	0.67/0.70	0.35/0.40
FOM[d] (acentric/centric)	0.29/0.35			

Values for the highest-resolution shell are indicated by parentheses. All reflections were processed.

[a] $R_{sym} = \Sigma|I_h - \langle I\rangle_h|/\Sigma I_h$, where $\langle I\rangle_h$ is the average over symmetry equivalents, and h is Miller reflection index (hkl).

[b] Phasing power is the mean value of the heavy atom structure factor amplitude divided by the lack of closure for isomorphous/anomalous differences.

[c] R_{Cullis} is the lack of closure divided by the absolute value of the difference between FPH and FP for isomorphous differences of acentric/centric data.

[d] FOM is the figure of merit.

Table 2. Crystal Structure Refinement Statistics

Resolution limits (Å)	40.0–2.45
R factor (working/free)[a] (%)	19.3/24.7
Number of atoms (protein/water/formate)	2552/25/9
Average B factor (Å^2)	37.1/34.6/59.1
Rmsd: bonds/angles	0.012 Å/1.43°
Ramachandran analysis (Non-Gly, Non-Pro)	
Most favored	247
Allowed	33
Disallowed	1

[a] R factor = $\Sigma|F_o - F_c|/\Sigma F_o$.

nsP2 (Vasiljeva et al., 2001) that defined a soluble nsP2 region with protease activity (nsP2pro; VEEV residues Met457–Cys794). This protein was expressed in *E. coli*, purified, and crystallized. Analysis of diffraction data indicated that nsP2pro crystals belonged to space group $P2_12_12_1$. Based on the Matthews coefficient (Vm = 2.4 Å^3/Da) (Matthews, 1968), each asymmetric unit was predicted to contain a single monomer of nsP2pro, and the crystal solvent content was determined to be ~49% (Russo and Watowich, 2006). Crystals diffracted to at least 2.45 Å resolution at the CAMD PX beamline. The nsP2pro structure was solved by using multiple isomorphous replacement and anomalous scattering (MIRAS) methods. Data collection and MIRAS statistics are presented in Table 1. The final model refined to an R factor and R_{free} of 19.3% and 24.7%, respectively, and contained 320 residues, 25 water molecules, and 3 formate ions. In the final model, the first and last clearly visible residues were Asp468 and Ser787, respectively. Structure refinement statistics are summarized in Table 2.

The polypeptide chain folds into two distinct compact domains of approximately equal size with multiple water molecules positioned at the interface between the domains (Figure 2). This interface consists of regions of helix and random coil. The N-terminal proteolytic domain from residue Asp468 to residue Asn603 contains the conserved cysteine protease catalytic dyad formed by Cys477 and His546. It is organized around a central cluster of helices that are flanked by two short β hairpins. The orientation of the catalytic dyad residues in nsP2 is similar to the catalytic dyad conformation observed in papain (Figure 3A) (Drenth et al., 1968; Smith, 1957). The catalytic cysteine is positioned at the N-terminal end of an α helix, and the catalytic histidine is part of a β strand. However, organization of the tertiary structure of the nsP2pro protease domain is different from that observed in papain and any other known protein structure, thus clearly indicating that the nsP2pro domain adopts a novel fold.

The C-terminal domain of nsP2pro extends from Arg604 to Ser793. This domain contains approximately equal amounts of helix and strand secondary structural elements arranged in three layers, with the faces of a central β sheet flanked by α helices. The function of the C-terminal domain is unclear, but 3D structural comparisons with DALI (Holm and Sander, 1995), the NCBI Vector Alignment Search Tool (VAST) (Gibrat et al., 1996), and ProFunc (Laskowski et al., 2005) all indicate that the tertiary structure of the C-terminal domain of nsP2pro is similar to that of the methyltransferase family of enzymes (Figure 5). The structure of the nsP2pro C-terminal domain is similar to proteins belonging to the S-adenosyl-L-methionine (SAM)-dependent methyltransferases superfamily as defined by the SCOP (Structural Classification of Proteins) taxonomy (Murzin et al., 1995). This superfamily includes methyltransferases from a wide variety of organisms, including the flavivirus RNA cap (nucleoside-2′-O-)-methyltransferase domain of RNA polymerase NS5 and the *E. coli* heat shock protein FtsJ RNA methyltransferase. Sequence identities are 21% between nsP2pro and FtsJ and 19% between nsP2pro and dengue virus NS5 methyltransferase. Sequence similarity between dengue virus NS5 methyltransferase and nsP2pro, which is believed to be enzymatically inactive, has been noted previously (Sawicki et al., 2006).

Structure and Sequence Conservation

Sequence alignment across representative alphavirus nsP2pro sequences reveals only limited sequence identity and moderate similarity amongst the related sequences (Figure 1). The catalytic residues Cys477 and His546 are invariant in all alphavirus nsP2 sequences. The 3 residues immediately following each catalytic residue are also completely conserved, and at least 1, Trp547, has been shown to be necessary for proteolytic activity (Strauss et al., 1992). Overall sequence conservation in nsP2pro is 18%, with only 62 residues being invariant over all alphavirus strains. Interestingly, secondary structure elements (determined from the coordinates of the VEEV nsP2pro structure) in many regions of the protein correspond with regions of poor sequence conservation.

Proteolytic Domain and Catalytic Site

The core of the nsP2pro proteolytic domain consists of six helices flanked by two regions of short β hairpins

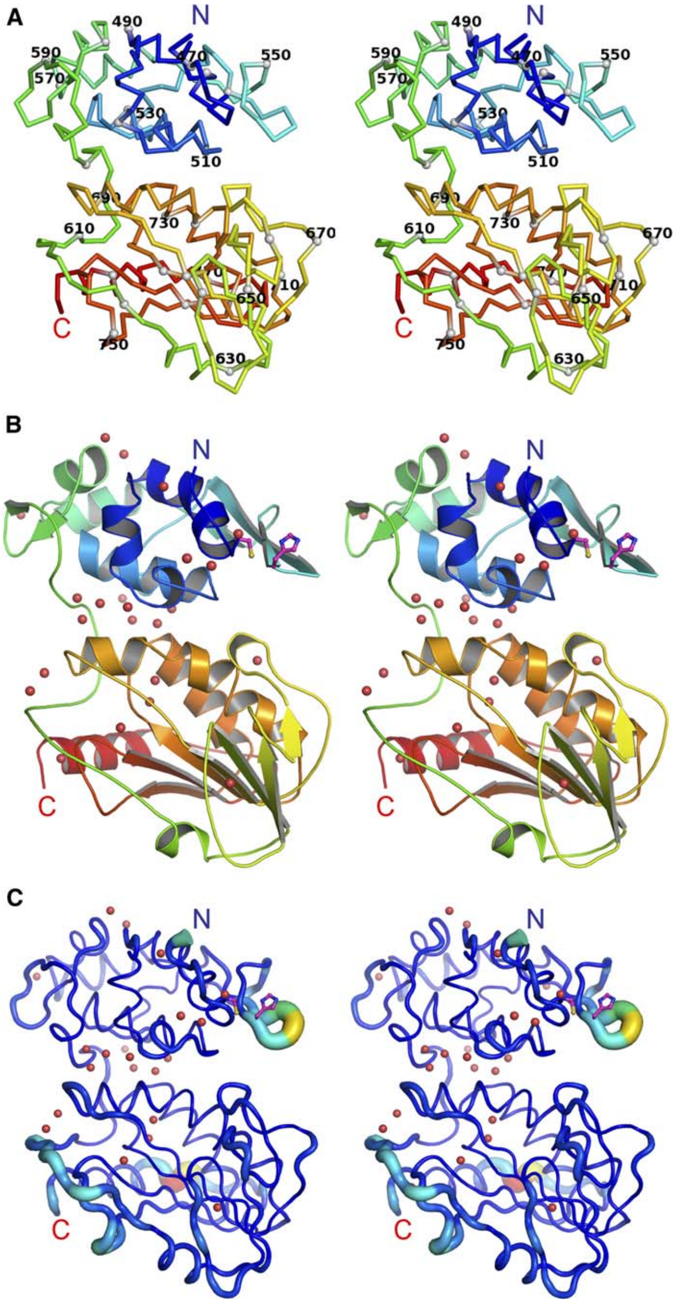

Figure 2. Structure of VEEV nsP2pro

(A) Trace of nsP2pro C_{α} atoms colored from blue (N terminus) to red (C terminus) with every 10th residue displayed as a light-gray ball. Every 20th residue position is labeled.

(B) Ribbon diagram of nsP2pro colored from blue (N terminus) to red (C terminus) and including bound water (red spheres).

(C) B factor "worm" representation of nsP2pro. Color and tube diameter reflect relative main chain B factors. Red color and a large diameter indicate high B values. The catalytic residues Cys477 and His547 are shown in ball-and-stick, and waters are shown as red spheres. All wall-eye stereographic images were made with Pymol.

and a single-turn 3_{10} helix. Structure similarity searches with the NCBI VAST search (Gibrat et al., 1996), DALI (Holm and Sander, 1995), and the protein structure comparison service SSM (Krissinel and Henrick, 2004) revealed very low similarity to several cysteine proteases, including papain, several cathepsins, and the FMDV leader peptidase. However, the results obtained with different structure comparison utilities were inconsistent. The majority of proteins that exhibited low similarity to the nsP2pro proteolytic domain were cysteine proteases, but their identity and statistical significance varied when analyzed by the different utilities. The DALI program identified the FMDV leader peptidase (PDB code: 1QMY) as having a low similarity to the nsP2pro domain, with a z score of 3.7 and 9% identity over 79 aligned residues. The VAST program identified papain (PDB code: 1PPN) and human cathepsin X (PDB code: 1EF7) as having little structural similarity to the nsP2pro domain. Cathepsin X was a slightly better structural match to the nsP2pro domain than papain since it was calculated to have a lower VAST p value and a higher residue percent identity within the superimposed regions. The regions of structural alignment between the VEEV proteolytic domain and cysteine proteases are very limited, and they include only a few residues in the immediate vicinity of the catalytic dyad. Other than nsP2pro, no known cysteine protease has been found to derive both residues of the catalytic dyad from the same domain. This suggests that the nsp2pro N-terminal domain is a novel cysteine protease fold.

The cysteine proteases identified as having low similarity to VEEV nsP2pro all belonged to the cysteine

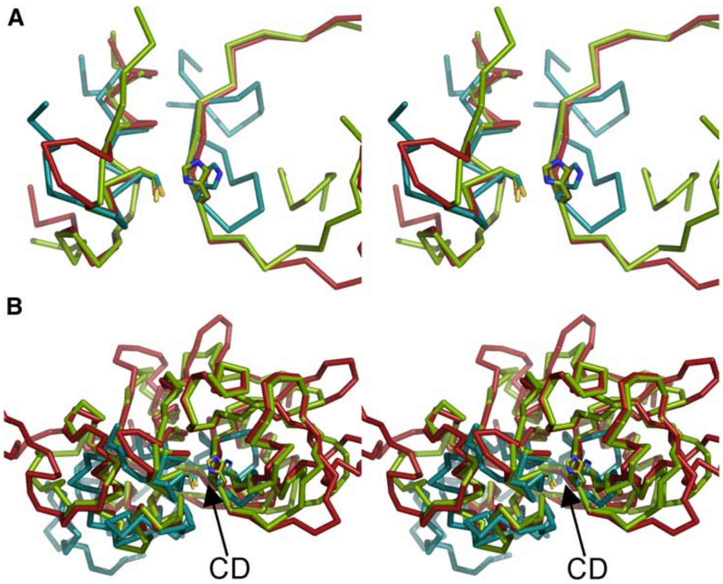

Figure 3. Superposition of the nsP2pro Catalytic Dyad with Those of Papain and Human Cathepsin X

(A) A close-up view of the catalytic dyad of cathepsin X (green) and papain (red) showing strong similarity between the two and clear structural differences from VEEV nsP2pro (light blue). The divergence of nsP2pro from the papain and cathepsin X structures increases with increasing distance from catalytic dyad.

(B) An expanded view of the superposition of cysteine protease structures shows that cathepsin X (green) and papain (red) have similar two-domain tertiary structures, and that they form distinct tertiary structures relative to VEEV nsP2pro (light blue).

proteinase superfamily (SCOP). These cysteine proteinases are characterized by a common catalytic core made of one α helix and three strands of β sheet (Murzin et al., 1995). In each of the cysteine protease structures, nsP2pro, papain, and human cathepsin X, the catalytic cysteine is situated at the N terminus of an α helix, and the catalytic histidine is located on a β strand (Figure 3). In papain and cathepsin X, the cysteine and histidine residues of the catalytic dyad are located in two separate domains, and the active site occupies the interface between these domains (Drenth et al., 1968; Smith, 1957). In contrast, in nsP2pro, the β strand containing the catalytic histidine is part of a short β hairpin within the N-terminal proteolytic domain and is not provided by a separate domain. Moreover, the nsP2pro secondary structure composition and topological arrangement differ significantly from cysteine protease structures deposited in the PDB. Notably, tertiary structure comparisons between nsP2pro and other proteins within the PDB indicate that, to our knowledge, the nsP2pro proteolytic domain represents a unique cysteine protease fold and a novel protein fold.

The nsP2pro structure confirms the hypothesis, generated from mutational studies (Strauss et al., 1992), that Cys477 and His546 come together to form a catalytic dyad within the protease active site (Figures 2A, 3A, and 4C). Although a conserved asparagine (Asn549) is located near these residues, and could potentially function as the third element of a catalytic triad, this asparagine residue is not oriented to interact with the histidine residue. Moreover, mutational studies of the related Sindbis virus (SINV) show that this residue is not essential for activity (Strauss et al., 1992).

A deep and pronounced groove transects the active site, suggesting the locations of S1, S2, and S3 sites in the protein that orient a peptide substrate relative to the catalytic dyad (Figure 4). A consensus peptide substrate Glu-Ala-Gly-Ala, corresponding to the nsP1-nsP2 cleavage motif, was modeled into the protease active site based on coordinates obtained by aligning Cys477, His546, and Trp547 with the corresponding residues in the Ulp1-SUMO complex (root-mean-square deviation [rmsd] of 1.1 Å) (PDB code: 1EUV). The backbone coordinates of the Ulp1 substrate bound to SUMO were used to position the P1–P4 residues of the nsP2 substrate consensus peptide. The P4 glutamic acid rotamer is also derived directly from this alignment. The position of the peptide substrate within the binding groove was based exclusively from the alignment of the active site residues of nsP2pro and Ulp1 and was not manually adjusted or refined by using molecular dynamic calculations. The fit of the Glu-Ala-Gly-Ala substrate within the nsP2 active site clearly delineates locations for the S1, S2, and S3 binding sites on the protease. These sites appear as shallow depressions on the protein surface, consistent with a surface that interacts with a peptide substrate containing either glycine or small side chain consensus residues. The S1, S2, and S3 sites line a long, deep groove formed at the interface between the nsP2Pro N- and C-terminal domains. Residues Asn544, Asn545, and His546 form a thumb that may regulate access into and out of this binding groove. The elevated temperature factors for residues within this region (Figure 2B) suggest that this protein segment is flexible. This thumb corresponds to a region of the protein where the SINV antigenic complex has 7 residues inserted in the aligned alphavirus sequences.

The S1 pocket is a small depression that is located ~6 Å from the catalytic cysteine and formed by residues Val476, Asn475, and Ala509. These S1 residues are either highly conserved or have only conservative substitutions (Figure 1). The backbone amides from Cys477 and Val476 likely form the oxyanion hole that is observed in many other cysteine and serine proteases.

The most significant contribution to defining the S2 binding site is made by Trp547; this site is conserved across all known alphavirus nsP2 sequences. The structures of several cysteine proteases that require substrates with a P2 glycine motif contain bulky aromatic residues immediately after the catalytic histidine, and these aromatic residues define the protease S2 site. Golubtsov and coworkers (2006) have called this the

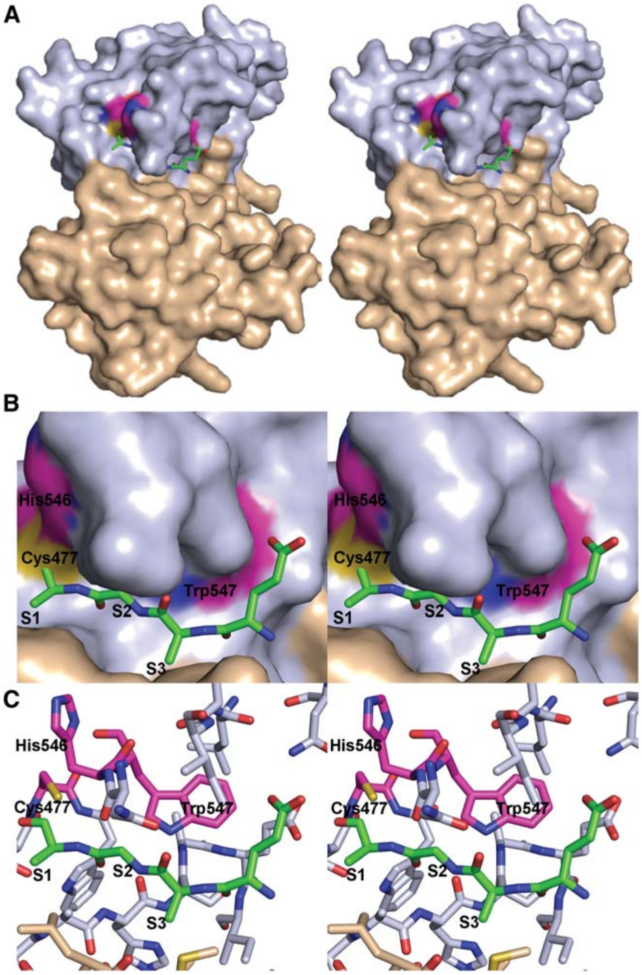

Figure 4. Model of the EAGA Peptide Bound at the nsP2pro Active Site

(A) Surface rendering of nsP2pro and a stick model of the EAGA substrate. The N-terminal domain is colored pale blue, and the C-terminal domain is colored tan. The catalytic residues plus Trp547 are colored by atom type, with carbon colored magenta, nitrogen colored blue, oxygen colored red, and sulfur colored yellow. The peptide substrate is colored similarly by atom type, except that carbon atoms are green.

(B) Close-up view of the nsP2pro substrate binding groove with the model EAGA consensus substrate bound. Atom coloring is as in Figure 3A. Placement of the substrate peptide clearly indicates the S1, S2, and S3 substrate binding sites.

(C) Close-up view of the residues in the nsP2pro binding groove and the bound substrate.

glycine specificity motif (GSM). The nsP2pro structure shows that the conserved Trp547 indole nitrogen is positioned between two shallow surface depressions near the active site (Figure 4C). The interaction between the S2 tryptophan and the substrate P2 glycine suggests the use of the GSM selection strategy in alphavirus nsP2pros.

Residues Ile698 and Met702, located in the C-terminal domain, form the S3 binding site (Figure 3). Met702 is highly conserved (~75% identity) within the alphavirus strains listed in Figure 1. Although residue Ile698 is not highly conserved within the alphavirus strains, amino acid substitutions at this position are all hydrophobic, with methionine and isoleucine predominating. Additionally, the shallow S3 pocket is flanked by residues Ala509 and His510, both located at sites of high sequence conservation (~75%).

Three nsP2 cleavage sites are present in the nsP1234 polyprotein. All cleavage intermediates containing nsP2 have been shown to be proteolytically active (Vasiljeva et al., 2003). Protease activity in the polyprotein suggests the possibility of *cis* cleavage events contributing to processing as well as bimolecular *trans* cleavage. Available evidence is strongly suggestive that the processing of the nsP23 cleavage site occurs in *trans* (Vasiljeva et al., 2003). This suggestion is supported by examination of the nsP2pro structure. The measured distance from the last visible residue in the nsP2pro structure, Ser787, to the P1 alanine of the model substrate (representing the nsP23 cleavage site) is ~42 Å. A minimum of 12 residues in extended conformation would be necessary to span this distance. However, only 7 residues are missing at the C terminus of the nsP2pro structure, clearly indicating that the nsP23 cleavage site is not accessible to the protease active site in this conformation. This structural insight is in agreement with the proposed *trans* cleavage mechanism for nsP23 (Vasiljeva et al., 2003).

C-Terminal Domain

The fold of the SAM-dependent methyltransferase superfamily is described in SCOP (Murzin et al., 1995) as three layers, termed a/b/a, with a mixed β sheet of seven strands arranged in the order 3-2-1-4-5-7-6 and sandwiched between helices with strand 7 oriented antiparallel to the other strands. This is an appropriate description of the fold of the nsP2pro C-terminal domain, although one of the α helix layers consists of a single helix that is 5 residues long. This small helix results in many residues of β strands 6 and 7 being exposed to solvent in the nsP2pro structure. The nsP2pro C-terminal domain shows significant tertiary structure similarity to known methyltransferase structures (e.g., FtsJ, dengue virus NS5). An alignment between nsP2pro and FtsJ was

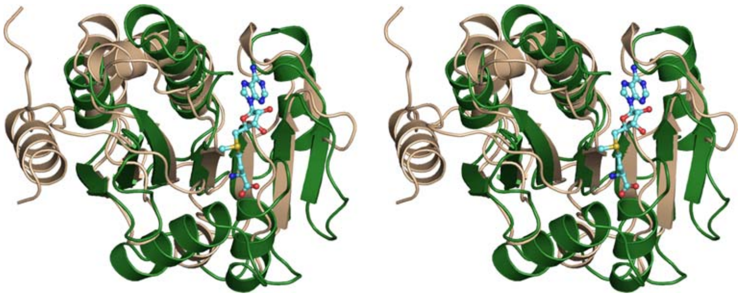

Figure 5. Superposition of the nsP2 C-Terminal Domain with *E. coli* FtsJ RNA Methyltransferase

FtsJ methyltransferase is shown in green, and the C-terminal domain of VEEV nsP2pro is colored tan. SAM substrate bound to FtsJ is colored light blue. Superposition was obtained by maximizing the spatial overlap of the six longest β strands in the core region of each protein. Although the core β sheets are conserved, the remaining secondary structural elements are not. The secondary structure surrounding the SAM binding site is not conserved, indicating that this domain in nsP2pro is likely not a functional methyltransferase.

constructed by using only the six longest strands of the β sheet and the location of SAM in the FtsJ structure (Figure 5). All of the β strands align very well, and the long helix on the upper face of the β sheet is brought into close alignment. However, the backbone alignment in proximity to SAM is poor, and residues in the region of nsP2pro that correspond to the FtsJ methyltransferase SAM substrate binding site show no significant similarity to each other. Moreover, little sequence identity is observed in alphaviruses for residues aligned to the SAM substrate binding site. These observations are consistent with the proposal that the nsP2pro C-terminal domain lacks methyltransferase enzymatic activity despite having structural similarity to the FtsJ methyltransferase and low sequence similarity to dengue virus NS5 methyltransferase. However, the nsP2pro methyltransferase fold could be used as a scaffold to bind RNA elements that may regulate protease activity and virus replication.

Alphavirus nsP2pro Functional Mutants

Temperature-sensitive (ts) mutants in the C-terminal region of nsP2 that affect RNA synthesis and protease activity differently have been identified in related Sindbis and Semliki Forest viruses (Table 3) (Agapov et al., 1998; Hahn et al., 1989; Lulla et al., 2006b; Suopanki et al., 1998). We have examined the role of four temperature-sensitive mutations in the context of the VEEV nsP2pro structure (Figure 6). These temperature-sensitive mutations are of special interest because they occur at residues highly conserved across alphavirus strains and disrupt only a subset of nsP2pro functions (Hahn et al., 1989; Strauss et al., 1992). Additional mutants identified in Sindbis and Semliki Forest viruses (Sawicki et al., 2006) have not been included in this study due to low sequence identity at the site of mutation. Mutants are referred to by using VEEV sequence numbering.

Table 3. Alphavirus nsP2 Temperature-Sensitive Mutants

		Temperature-Sensitive Phenotype		
Mutant Name	Mutation (VEEV Residue)	(−) Strand Synthesis	Protease Activity	26S RNA Synthesis
ts7	D522N (N517)	D[a]	wt[b]	wt
ts18	F509L (F504)	wt	D	D
ts24	G736S (G723)	wt	D	D
S2	P726T (P713)	D	wt	D

[a] "D" indicates a defect in normal function at the nonpermissive temperature.

[b] "wt" indicates normal wild-type behavior at the nonpermissive temperature.

Mutants ts18 (F504L) and ts24 (G723S) impair protease activity and cause an increase in 26S subgenomic RNA synthesis relative to 42S genomic RNA (Hahn et al., 1989; Suopanki et al., 1998). Interestingly, these mutations are found in different domains of nsP2pro; ts18 is in the N-terminal domain, and ts24 is in the C-terminal domain. Both mutations occur at residues buried deep in the hydrophobic cores of their respective domains. Phe504 interacts with Trp478, Phe470, Trp498, and Lys473, all conserved residues across alphavirus strains. Phe504 also interacts with Val477, which is not completely conserved, but is replaced by threonine in Aura virus. The location of Val477 near the protein surface suggests that threonine could be tolerated in this position by directing the hydroxyl group toward the surface and still presenting a nonpolar methyl group to Phe504. Mutation of Phe504 to leucine removes three nonpolar carbons from the interior of the N-terminal domain, leaving a void, which likely perturbs the hydrophobic core, resulting in instability at the nonpermissive temperature and loss of function.

The mutation of glycine to serine in ts24 could have multiple effects. The Gly723 main chain is completely buried and surrounded by the side chains of conserved residues Tyr784, Tyr724, His608, and Ala725. The backbone dihedral angles at this residue map to a region of the Ramachandran plot that is disallowed for nonglycine residues (lower-right quadrant of the phi-psi plot) (Ramachandran et al., 1963). Substitution of any residue here will strain the backbone and likely destabilize the protein. In addition, depending on the side chain rotamer, the surrounding residues would have steric clashes with a serine substitution at this position. These clashes and the energetic cost of burying a polar hydroxyl in the hydrophobic core would likely destabilize the hydrophobic interactions in the core and cause instability at the nonpermissive temperature and loss of function.

Mutants ts7 (N517D) and S2 (P713T) occur at solvent-exposed residues. Position 517 is aspartic acid in SINV and Aura viruses and asparagines in other alphaviruses. In SINV, mutation from aspartic acid to the alphavirus consensus residue asparagine impairs downregulation of (−) strand RNA synthesis at the nonpermissive temperature (Suopanki et al., 1998). Protease activity and 26S RNA synthesis are unaffected (Suopanki et al., 1998), implying that this mutation does not destabilize the protein. Residue 517 is within a highly polar region

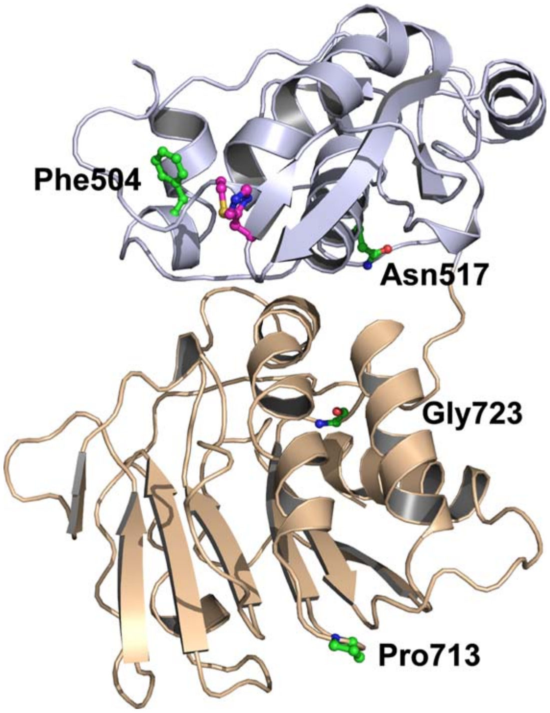

Figure 6. Locations of Temperature-Sensitive Mutants Mapped onto the VEEV nsP2pro Structure

The nsP2pro cartoon is colored by domain as in Figure 2. Stick representations of residues are colored by atom type, with carbon atoms in the mutation sites (labeled) colored green and the catalytic dyad (unlabeled) colored magenta.

of nsP2pro and has several side chain-water hydrogen bonds. The residue does not appear to make any critical interactions with other side chains in the region, but it does contact the surface in a cavity that may be involved in interactions with other proteins or RNA.

The S2 mutation disrupts nsP2-mediated viral RNA synthesis, but it does not disrupt nsP2pro activity (Sawicki et al., 2006). Mutations at this position in Sindbis virus and Sindbis replicons (Agapov et al., 1998; Frolov et al., 1999, Frolova et al., 2002) are also known to reduce the viral cytopathatic effect and replication in cell culture in a cell-line-dependent manner. Pro713 is found in a tight turn at the end of a helix and participates in a parallel special β bulge with residues Asp675, Asn712, Gly714, and Gly715. This β bulge is located on the face distal to the N-terminal protease domain. The sequence (XPGG) and structural motif observed at residue 713 is found in several RNA methyltransferases, including the FtsJ and reovirus core methyltransferases. It is possible that this unusual interaction is necessary for local structural stability of the C-terminal domain and recognition of RNA substrates. Although the C-terminal domain of nsP2pro may not exhibit methyltransferase activity, it could utilize the methyltransferase structural motif as a scaffold for binding RNA and regulating virus replication.

Conclusions

We have described an atomic resolution model of the protease domain of VEEV nsP2, representing the first, to our knowledge, structure of an alphavirus nsP2pro. This model shows that the proteolytically active region of nsP2 is organized into two discrete folded domains: an N-terminal domain that encompasses the protease catalytic dyad and active site, and a C-terminal domain with structural similarity to methyltransferases but with no known enzymatic activity. To our knowledge, the tertiary structure of the N-terminal domain has not been observed previously and represents a novel cysteine protease and protein fold. The substrate binding site and catalytic dyad are well defined. Placement of bound ligand from related cysteine proteases into the nsP2pro active site clearly delineates the S1, S2, and S3 binding sites and suggests substrate recognition and binding mechanisms. The C-terminal domain forms part of the active site and may be involved in substrate recognition, regulation of protease activity, and regulation of RNA replication. This structure will significantly aid drug discovery and development efforts to combat VEEV and related viruses.

Experimental Procedures

Cloning, Expression, Purification, and Crystallization of the Protease Domain of VEEV nsP2, nsP2pro

Cloning, expression, purification, and crystallization of nsP2pro have been described previously (Russo and Watowich, 2006). In brief, DNA coding for nsP2pro (residues Met457–Cys794 of VEEV nsP2) was amplified by polymerase chain reaction (PCR) and cloned into the pETBlue1 T7 expression vector (Novagen). Tuner DE3 (pLacI) *E. coli* cells were transformed with this vector, grown in flask cultures at 37°C, and induced by the addition of IPTG. Cells were pelleted and lysed, and the supernatant was processed through SP-Sepharose, Ni-Sepharose, and Superdex-200 chromatography columns to obtain >99% pure nsP2pro (data not shown). Mass spectroscopy and N-terminal sequencing were used to verify the identity of the purified protein (data not shown). Purified nsP2pro was concentrated to ~6.0 mg/ml and crystallized with 3.0 M ammonium formate (pH unadjusted), 2% 2-methyl-2, 4-pentanediol (MPD), 1% glycerol, and 0.2 mM zinc acetate.

Preparation of Isomorphous Heavy Atom Derivatives

Crystals were transferred from crystallization drops; washed in stabilizing solution containing 3.0 M lithium formate, 2% MPD, and 1% glycerol; and transferred to stabilizing solution containing 1–20 mM of the appropriate heavy atom compound. Crystals were soaked for a variety of times varying from 2 to 7 days in the heavy atom solution and were then backsoaked into 2.5 M lithium formate, 2% MPD, and 40% glycerol. Typical crystals used for data collection were 400–500 μm long, 20 μm wide, and less than 5 μm thick.

Data Collection and Analysis

Crystals were soaked in a cryoprotectant solution containing 2.5 M lithium formate, 2% MPD, and 40% glycerol for 5–10 min before being flash cooled in a 100 K nitrogen gas stream. Diffraction data were collected at 100 K by using 1° wide frames on a DIP2030 imaging plate detector mounted on a MacScience M06HF rotating anode X-ray generator equipped with a 100 μm CuK$_\alpha$ source and Rigaku confocal optics. A high-resolution native data set was collected on a MAR CCD at a wavelength of 1.3808 Å at the Center for Advanced Microstructures and Devices (CAMD) Gulf Coast Consortium Protein Crystallography PX1 beamline. Diffraction data were indexed, integrated, and scaled by using HKL2000 (Otwinowski and Minor, 1997). Identification of heavy atom sites, determination of phases, and density modification with DM (Cowtan, 1994) were performed by using the SHARP software package (Buster Development Group) (Bricogne et al., 2003) to give an initial experimental electron density map at 3.0 Å resolution. The initial Figure of Merit for acentric (FOMacen) and centric (FOMcen) reflections was 0.29 and 0.35, respectively. After density modification in DM, the mean FOM increased to 0.82. The DM-improved map was used for initial model building with TEXTAL (Romo et al., 2005); this placed polypeptide into >80% of the electron density map. Approximately half of the residues placed were identified correctly. The model was improved with iterative rounds of manual model building in Xtalview (McRee, 1999), by using composite omit maps, and PMB/CNS (Brunger et al., 1998) refinement against the 2.45 Å data set. The stereochemical bond rmsd target was set to 0.012 Å, as determined by PMB (Singh et al., 2006) based on the ratio of observed:free parameters. Final refinement steps used the improved set of CNS bond length and angle parameters from the latest version of PMB. Sequence alignments based on secondary structure predictions were determined with SSM (Krissinel and Henrick, 2004) accessed through the ProFunc web server (Laskowski et al., 2005).

Acknowledgments

We thank Dr. Henry Bellamy for assistance with X-ray diffraction data collection at the Center for Advanced Microstructures and Devices (CAMD) PX beamline. M.A.W. was supported by a grant from the Sealy and Smith Foundation to the Sealy Center for Structural Biology and Molecular Biophysics.

Received: June 27, 2006
Revised: July 29, 2006
Accepted: July 31, 2006
Published: September 12, 2006

References

Agapov, E.V., Frolov, I., Lindenbach, B.D., Pragai, B.M., Schlesinger, S., and Rice, C.M. (1998). Noncytopathic Sindbis virus RNA vectors for heterologous gene expression. Proc. Natl. Acad. Sci. USA *95*, 12989–12994.

Bricogne, G., Vonrhein, C., Flensburg, C., Schiltz, M., and Paciorek, W. (2003). Generation, representation and flow of phase information in structure determination: recent developments in and around SHARP 2.0. Acta Crystallogr. D Biol. Crystallogr. *59*, 2023–2030.

Bronze, M.S., Huycke, M.M., Machado, L.J., Voskuhl, G.W., and Greenfield, R.A. (2002). Viral agents as biological weapons and agents of bioterrorism. Am. J. Med. Sci. *323*, 316–325.

Brunger, A.T., Adams, P.D., Clore, G.M., DeLano, W.L., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.S., Kuszewski, J., Nilges, M., Pannu, N.S., et al. (1998). Crystallography & NMR system: a new software suite for macromolecular structure determination. Acta Crystallogr. D Biol. Crystallogr. *54*, 905–921.

Choi, H.K., Tong, L., Minor, W., Dumas, P., Boege, U., Rossmann, M.G., and Wengler, G. (1991). Structure of Sindbis virus core protein reveals a chymotrypsin-like serine proteinase and the organization of the virion. Nature *354*, 37–43.

Choi, H.K., Lee, S., Zhang, Y.P., McKinney, B.R., Wengler, G., Rossmann, M.G., and Kuhn, R.J. (1996). Structural analysis of Sindbis virus capsid mutants involving assembly and catalysis. J. Mol. Biol. *262*, 151–167.

Choi, H.K., Lu, G., Lee, S., Wengler, G., and Rossmann, M.G. (1997). Structure of Semliki Forest virus core protein. Proteins *27*, 345–359.

Cowtan, K.D. (1994). dm: an automated procedure for phase improvement by density modification. Joint CCP4 and ESF-EACBM Newsletter on Protein Crystallography *31*, 34–38.

Drenth, J., Jansonius, J.N., Koekoek, R., Swen, H.M., and Wolthers, B.G. (1968). Structure of papain. Nature *218*, 929–932.

Faragher, S.G., Meek, A.D., Rice, C.M., and Dalgarno, L. (1988). Genome sequences of a mouse-avirulent and a mouse-virulent strain of Ross River virus. Virology *163*, 509–526.

Feng, Y.X., Copeland, T.D., Oroszlan, S., Rein, A., and Levin, J.G. (1990). Identification of amino-acids inserted during suppression of Uaa and Uga termination codons at the Gag-Pol junction of moloney murine leukemia-virus. Proc. Natl. Acad. Sci. USA *87*, 8860–8863.

Frolov, I., Agapov, E., Hoffman, T.A., Jr., Pragai, B.M., Lippa, M., Schlesinger, S., and Rice, C.M. (1999). Selection of RNA replicons capable of persistent noncytopathic replication in mammalian cells. J. Virol. *73*, 3854–3865.

Frolova, E.I., Fayzulin, R.Z., Cook, S.H., Griffin, D.E., Rice, C.M., and Frolov, I. (2002). Roles of nonstructural protein nsP2 and α/β interferons in determining the outcome of Sindbis virus infection. J. Virol. *76*, 11254–11264.

Gibrat, J.F., Madej, T., and Bryant, S.H. (1996). Surprising similarities in structure comparison. Curr. Opin. Struct. Biol. *6*, 377–385.

Golubtsov, A., Kaariainen, L., and Caldentey, J. (2006). Characterization of the cysteine protease domain of Semliki Forest virus replicase protein nsP2 by in vitro mutagenesis. FEBS Lett. *580*, 1502–1508.

Gomez de Cedron, M., Ehsani, N., Mikkola, M.L., Garcia, J.A., and Kaariainen, L. (1999). RNA helicase activity of Semliki Forest virus replicase protein NSP2. FEBS Lett. *448*, 19–22.

Gouet, P., Courcelle, E., Stuart, D.I., and Metoz, F. (1999). ESPript: analysis of multiple sequence alignments in PostScript. Bioinformatics *15*, 305–308.

Hahn, Y.S., Strauss, E.G., and Strauss, J.H. (1989). Mapping of RNA-temperature-sensitive mutants of Sindbis virus: assignment of complementation groups A, B, and G to nonstructural proteins. J. Virol. *63*, 3142–3150.

Holm, L., and Sander, C. (1995). Dali: a network tool for protein structure comparison. Trends Biochem. Sci. *20*, 478–480.

Kinney, R.M., Johnson, B.J., Welch, J.B., Tsuchiya, K.R., and Trent, D.W. (1989). The full-length nucleotide sequences of the virulent Trinidad donkey strain of Venezuelan equine encephalitis virus and its attenuated vaccine derivative, strain TC-83. Virology *170*, 19–30.

Krissinel, E., and Henrick, K. (2004). Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. Acta Crystallogr. D Biol. Crystallogr. *60*, 2256–2268.

Lanciotti, R.S., Ludwig, M.L., Rwaguma, E.B., Lutwama, J.J., Kram, T.M., Karabatsos, N., Cropp, B.C., and Miller, B.R. (1998). Emergence of epidemic O'nyong-nyong fever in Uganda after a 35-year absence: genetic characterization of the virus. Virology *252*, 258–268.

Laskowski, R.A., Watson, J.D., and Thornton, J.M. (2005). ProFunc: a server for predicting protein function from 3D structure. Nucleic Acids Res. *33*, W89–W93.

Lee, E., Stocks, C., Lobigs, P., Hislop, A., Straub, J., Marshall, I., Weir, R., and Dalgarno, L. (1997). Nucleotide sequence of the Barmah Forest virus genome. Virology *227*, 509–514.

Lescar, J., Roussel, A., Wien, M.W., Navaza, J., Fuller, S.D., Wengler, G., Wengler, G., and Rey, F.A. (2001). The Fusion glycoprotein shell of Semliki Forest virus: an icosahedral assembly primed for fusogenic activation at endosomal pH. Cell *105*, 137–148.

Lulla, A., Lulla, V., Tints, K., Ahola, T., and Merits, A. (2006a). Molecular determinants of substrate specificity for semliki forest virus nonstructural protease. J. Virol. *80*, 5413–5422.

Lulla, V., Merits, A., Sarin, P., Kaariainen, L., Keranen, S., and Ahola, T. (2006b). Identification of mutations causing temperature-sensitive defects in Semliki Forest virus RNA synthesis. J. Virol. *80*, 3108–3111.

Matthews, B.W. (1968). Solvent content in protein crystals. J. Mol. Biol. *33*, 491–497.

McRee, D.E. (1999). XtalView Xfit—a versatile program for manipulating atomic coordinates and electron density. J. Struct. Biol. *125*, 156–165.

Mukhopadhyay, S., Zhang, W., Gabler, S., Chipman, P.R., Strauss, E.G., Strauss, J.H., Baker, T.S., Kuhn, R.J., and Rossmann, M.G. (2006). Mapping the structure and function of the E1 and E2 glycoproteins in alphaviruses. Structure *14*, 63–73.

Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. *247*, 536–540.

Myles, K.M., Pierro, D.J., and Olson, K.E. (2003). Deletions in the putative cell receptor-binding domain of Sindbis virus strain MRE16 E2 glycoprotein reduce midgut infectivity in *Aedes aegypti*. J. Virol. *77*, 8872–8881.

Otwinowski, Z., and Minor, W. (1997). Processing of X-ray diffraction data collected in oscillation mode. Macromol. Crystallogr. A *276*, 307–326.

Peranen, J., Rikkonen, M., Liljestrom, P., and Kaariainen, L. (1990). Nuclear localization of Semliki Forest virus-specific nonstructural protein nsP2. J. Virol. *64*, 1888–1896.

Ramachandran, G.N., Ramakrishnan, C., and Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. J. Mol. Biol. *7*, 95–99.

Rawlings, N.D., Morton, F.R., and Barrett, A.J. (2006). MEROPS: the peptidase database. Nucleic Acids Res. *34*, D270–D272.

Rikkonen, M., Peranen, J., and Kaariainen, L. (1994). ATPase and GTPase activities associated with Semliki Forest virus nonstructural protein nsP2. J. Virol. *68*, 5804–5810.

Romo, T., Gopal, K., McKee, E., Kanbi, L., Pai, R., Smith, J., Sacchettini, J., and Ioerger, T. (2005). TEXTAL: AI-based structural determination for X-ray protein crystallography. IEEE Intell. Syst. *20*, 59–63.

Rumenapf, T., Strauss, E.G., and Strauss, J.H. (1995). Aura virus is a New World representative of Sindbis-like viruses. Virology *208*, 621–633.

Russo, A.T., and Watowich, S.J. (2006). Purification, crystallization and X-ray diffraction analysis of the C-terminal protease domain of Venezuelan equine encephalitis virus nsP2. Acta Crystallograph. Sect. F –Struct. Biol. Cryst. Comm. *62*, 514–517.

Salonen, A., Vasiljeva, L., Merits, A., Magden, J., Jokitalo, E., and Kaariainen, L. (2003). Properly folded nonstructural polyprotein directs the Semliki Forest virus replication complex to the endosomal compartment. J. Virol. *77*, 1691–1702.

Sawicki, D.L., and Sawicki, S.G. (1993). A second nonstructural protein functions in the regulation of alphavirus negative-strand RNA synthesis. J. Virol. *67*, 3605–3610.

Sawicki, D.L., Perri, S., Polo, J.M., and Sawicki, S.G. (2006). Role for nsP2 proteins in the cessation of alphavirus minus-strand synthesis by host cells. J. Virol. *80*, 360–371.

Singh, R., White, M.A., Ramana, K.V., Petrash, J.M., Watowich, S.J., Bhatnagar, A., and Srivastava, S.K. (2006). Structure of a glutathione conjugate bound to the active site of aldose reductase. Proteins *64*, 101–110.

Smith, E.L. (1957). Active site and structure of crystalline papain. Fed. Proc. *16*, 801–809.

Strauss, J.H., and Strauss, E.G. (1994). The alphaviruses: gene expression, replication, and evolution. Microbiol. Rev. *58*, 491–562.

Strauss, E.G., De Groot, R.J., Levinson, R., and Strauss, J.H. (1992). Identification of the active site residues in the nsP2 proteinase of Sindbis virus. Virology *191*, 932–940.

Suopanki, J., Sawicki, D.L., Sawicki, S.G., and Kaariainen, L. (1998). Regulation of alphavirus 26S mRNA transcription by replicase component nsP2. J. Gen. Virol. *79*, 309–319.

Tong, L., Choi, H.K., Minor, W., and Rossmann, M.G. (1992). The structure determination of Sindbis virus core protein using isomorphous replacement and molecular replacement averaging between two crystal forms. Acta Crystallogr. A *48*, 430–442.

Uryvaev, L.V., Volckhov, V.E., Iuferov, V.P., Samokhvalov, E.I., Lebedev, A., Safronov, P.F., and Netesov, S.V. (1994). Primary structure of proteins of the nsP2 and nsP3 polymerase complex confirm the recombinant nature of western encephalitis virus. Dokl. Akad. Nauk *335*, 813–818.

Vasiljeva, L., Merits, A., Auvinen, P., and Kaariainen, L. (2000). Identification of a novel function of the alphavirus capping apparatus. RNA 5′-triphosphatase activity of Nsp2. J. Biol. Chem. *275*, 17281–17287.

Vasiljeva, L., Valmu, L., Kaariainen, L., and Merits, A. (2001). Site-specific protease activity of the carboxyl-terminal domain of Semliki Forest virus replicase protein nsP2. J. Biol. Chem. *276*, 30786–30793.

Vasiljeva, L., Merits, A., Golubtsov, A., Sizemskaja, V., Kaariainen, L., and Ahola, T. (2003). Regulation of the sequential processing of Semliki Forest virus replicase polyprotein. J. Biol. Chem. *278*, 41636–41645.

Weaver, S.C., Salas, R., Rico-Hesse, R., Ludwig, G.V., Oberste, M.S., Boshell, J., and Tesh, R.B. (1996). Re-emergence of epidemic Venezuelan equine encephalomyelitis in South America. VEE Study Group. Lancet *348*, 436–440.

Weaver, S.C., Tesh, R.B., and Shope, R.E. (1999). Alphavirus infections. In Tropical Infectious Diseases: Principles, Pathogens, & Practice, R.E. Guerrant, D.H. Walker, and P.F. Weller, eds. (New York, NY: Churchill Livingstone), pp. 1281–1287.

Accession Numbers

Coordinates have been deposited in the Protein Data Bank (PDB) with accession code 2HWK.