



Structural basis for substrate specificity of alphavirus nsP2 proteases

Andrew T. Russo, Robert D. Malmstrom, Mark A. White, Stanley J. Watowich*

Department of Biochemistry and Molecular Biology and the Sealy Center for Structural Biology and Molecular Biophysics,
University of Texas Medical Branch, Galveston, TX 77555, USA

ARTICLE INFO

Article history:

Received 26 January 2010

Accepted 18 April 2010

Available online 24 April 2010

Keywords:

Alphavirus

nsP2 protease

Protease specificity

ABSTRACT

The alphavirus nsP2 protease is essential for correct processing of the alphavirus nonstructural polyprotein (nsP1234) and replication of the viral genome. We have combined molecular dynamics simulations with our structural studies to reveal features of the nsP2 protease catalytic site and S1'–S4 subsites that regulate the specificity of the protease. The catalytic mechanism of the nsP2 protease appears similar to the papain-like cysteine proteases, with the conserved catalytic dyad forming a thiolate-imidazolium ion pair in the nsP2-activated state. Substrate binding likely stabilizes this ion pair. Analysis of bimolecular complexes of Venezuelan equine encephalitis virus (VEEV) nsP2 protease with each of the nsP1234 cleavage sites identified protease residues His⁵¹⁰, Ser⁵¹¹, His⁵⁴⁶ and Lys⁷⁰⁶ as critical for cleavage site recognition. Homology modelling and molecular dynamics simulations of diverse alphaviruses and their cognate cleavage site sequences revealed general features of substrate recognition that operate across alphavirus strains as well as strain specific covariance between binding site and cleavage site residues. For instance, compensatory changes occurred in the P3 and S3 subsite residues to maintain energetically favourable complementary binding surfaces. These results help explain how alphavirus nsP2 proteases recognize different cleavage sites within the nonstructural polyprotein and discriminate between closely related cleavage targets.

© 2010 Elsevier Inc. All rights reserved.

1. Introduction

Alphaviruses are enveloped arthropod borne (+)-sense RNA viruses that cause a variety of human and animal diseases worldwide [1]. Many human diseases resulting from alphavirus infections are nonlethal and characterized by fever, rash and intense joint pain [1]. These diseases are typically associated with “Old World” alphaviruses, such as Ross River virus (RRV) in Africa, Barmah Forest virus (BFV) in Australia, Semliki Forest virus (SFV) in Asia and Sindbis virus (SINV) in Europe, Africa, Asia and Australia. The recent Chikungunya virus epidemic in India and the Indian Ocean islands infected more than a million people, with infection rates approaching 45% in some areas [2,3]. In contrast, “New World” alphaviruses such as Eastern, Western and Venezuelan equine encephalitis viruses, typically cause severe flu-like symptoms, encephalitis and death. Mortality rates range from ~1% for Venezuelan equine encephalitis virus (VEEV) to upwards of 80% for Eastern equine encephalitis virus (EEEV) [4]. No vaccines or therapeutics are approved to combat alphavirus infections in humans.

Following infection, the alphavirus genome is released into the host cell cytoplasm and, depending on the alphavirus strain, translated into either one or two nonstructural polyproteins (nsPs). Strains possessing an opal codon between nsP3 and nsP4 produce two nonstructural polyproteins; nsP123 is produced when translation terminates at the opal codon and nsP1234 is produced by opal codon readthrough [1]. Strains lacking the opal codon only produce the nsP1234 polyprotein. The alphavirus polyprotein forms the initial virus replication complex, which is cleaved into the individual nonstructural proteins by nsP2 to enable replication to proceed productively. The necessary role of nsP2 protease in alphavirus replication makes it an appealing target for antiviral drug design.

Biochemical studies of the nsP2 protease showed considerable differences in cleavage efficiency at the nsP12, nsP23 and nsP34 polyprotein cleavage sites [5–7] with the substrate preference of both full-length nsP2 and the isolated protease domain being nsP34 > nsP12 >> nsP23 [7]. *In vitro* protease activity was enhanced when the peptide substrate contained at least ten prime side residues and five nonprime side residues (Schechter and Berger nomenclature [8]). Even though 15 residues flanking the scissile bond were required for optimal activity, in controlled biochemical studies only substrate residues P1'–P4 specifically impacted substrate recognition by the nsP2 protease [5]. The P2 position is a conserved glycine residue in all alphavirus nsP2 substrates. The remaining P1'–P4 substrate residues diverge both within and

* Corresponding author at: Department of Biochemistry and Molecular Biology, University of Texas Medical Branch, 301 University Blvd, Galveston, TX 77555-0647, USA. Tel.: +1 409 747 4749; fax: +1 409 747 4745.

E-mail address: watowich@xray.utmb.edu (S.J. Watowich).

between alphavirus strains, although the P1 and P3 sites show less variability than the P1' and P4 sites.

The recently determined crystal structure of the VEEV nsP2 protease identified likely interactions between the protease binding site and substrate residues [9]. However, that preliminary analysis did not investigate the molecular basis for the observed substrate preferences and how each alphavirus protease recognizes at least three different substrates. Here we provide a detailed analysis of alphavirus nsP2 protease substrate specificity and how variable target substrates are recognized. In addition, amino acid sequence covariation between the protease and its substrates across different alphaviruses is identified. These studies define the nsP2 protease residues that participate in substrate recognition and explain the observed variation of polyprotein target sequences between alphavirus strains.

2. Methods

2.1. Protein/peptide models

All three-dimensional models were constructed with PyMOL using VEEV nsP2 protease coordinates (Protein Data Bank identifier 2HWK [9]). Modelled peptide substrates were positioned by aligning the C α and C β atoms of nsP2 protease residues Cys⁴⁷⁷, His⁵⁴⁶ and Trp⁵⁴⁷ with the corresponding residues in the Ulp1-SUMO complex (PDB identifier 1EUW [10]), respectively. Following alignment, Ulp1-SUMO substrate residues were converted to the corresponding nsP2 substrate residues using the PyMOL mutagenesis function. Appropriate residues were added to the peptide ends using PyMOL builder, and the substrate manually docked. These protease/peptide complexes were subsequently refined using energy minimization and molecular dynamics (MD) simulations within the program CHARMM [11].

2.2. Simulations

To perform MD simulations of each nsP2 protease/peptide substrate complex, noncarbon hydrogens were added to each complex using the HBUILD [12] module in CHARMM. Histidine residues, including the catalytic His⁵⁴⁶, were uncharged. The nsP2 protease/peptide substrate complex was parameterized using the CHARMM 22 force field [13]. Simulations were performed *in vacuo*. Energy minimizations were initially performed with the protease fixed and the substrate unconstrained, followed by minimization with both the protease and substrate unconstrained. MD simulations were performed with 1 fs steps for 1 ns with no atomic positional constraints. The final 500 ps of each MD simulations were used to calculate average structures, which were subjected to a final energy minimization and used to analyze protease–substrate contacts.

Two-dimensional contact diagrams of the nsP2pro/substrate complex were generated using LIGPLOT [14] with a 4 Å cut-off distance for displayed contacts. Protease substrate contacts were also evaluated using CSU software [15] at the LPC CSU server (bip.weizmann.ac.il/oca-bin/lpccsu). Active site pK_a values were calculated using the PROPKA web interface (<http://nova.colombo58.unimi.it/protpka.htm> [16]).

2.3. Sequences

Representative alphavirus sequences were obtained from the National Center for Biotechnology Information (NCBI) protein sequence database as follows: Aura virus (AURAV), NCBI ID: NP_819011; Barmah Forest virus (BFV), NCBI ID: NP_054023; O'nyong-nyong virus (ONNV), NCBI ID: AAC97204; Ross River virus

(RRV), NCBI ID: NP_740679; salmon pancreas disease virus (SPDV), NCBI ID: Q8JJX1; Semliki Forest virus (SFV) NCBI ID: NP_740666; Sindbis virus (SINV), NCBI ID: AAM10974; sleeping disease virus (SDV), NCBI ID: NP_598184; Venezuelan equine encephalitis virus (VEEV), NCBI ID: AAB02516 and Western equine encephalitis virus (WEEV), NCBI ID: CAA52868.

3. Results and discussion

3.1. Overview of the VEEV nsP2 protease structure and catalytic mechanism

The protease domain of VEEV nsP2 is organized into two domains (Fig. 1). The N-terminal domain is composed of residues 468–603 and contains the catalytic dyad residues Cys⁴⁷⁷ and His⁵⁴⁶ [9]. The N-terminal domain also contains the majority of the residues involved in substrate binding. The C-terminal domain, composed of residues 604–793, contains several residues involved in substrate binding. However, these residues do not form part of the central catalytic site. Asn⁵⁴⁴ and Asn⁵⁴⁵ form the turn at the end of a highly dynamic β -hairpin that extends towards the C-terminal domain, possibly restricting access of substrate to the binding groove and contributing to substrate binding. The substrate binding groove of the protease can accommodate a five residue peptide in an extended conformation. In modelling studies, longer peptides extended beyond the binding groove.

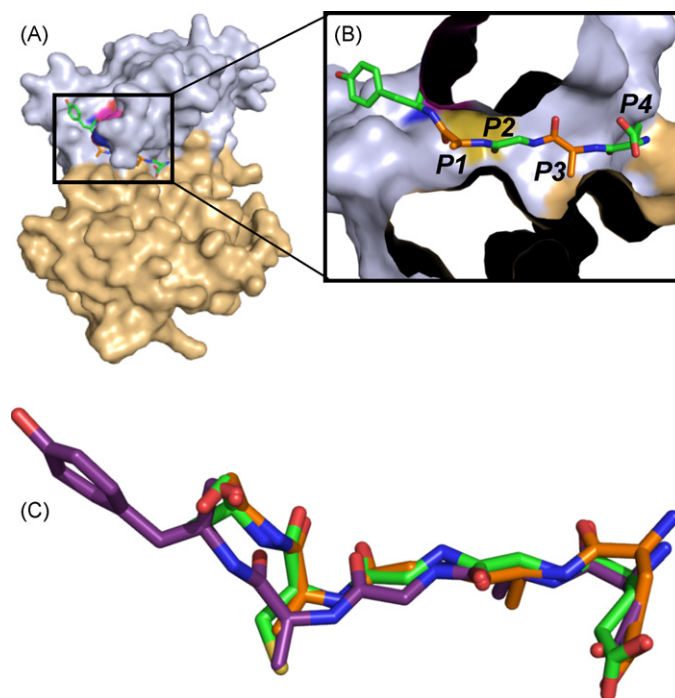


Fig. 1. nsP2pro with modelled nsP34 substrate peptide. (a) The N-terminal domain is colored light blue and the C-terminal domain is colored light brown. Substrate peptide, derived from the cleavage site between nsP1 and nsP2, is colored by atom type, with carbon, oxygen and nitrogen atoms colored green, red and blue, respectively. (b) Boxed region in panel A is magnified, rotated slightly and the surface of nsP2 protease cropped to provide a clearer image of the relationship between the substrate residues P1'–P4 and the protease binding subsites S1'–S4. The protease is colored as in panel A and the coloring of the substrate residues is alternated. P1', P2 and P4 are colored by atom as in panel A with carbon atoms in green. P1 and P3 are colored the same as in panel A with carbon in orange. (c) Overlay of models of peptides representing VEEV nonstructural polyprotein cleavage substrates nsP12, nsP23, and nsP34. Peptides are colored with nsP12 carbon atoms in orange, nsP23 carbon atoms in green and nsP34 carbon atoms in purple. Peptides extend left to right from P1' to P5.

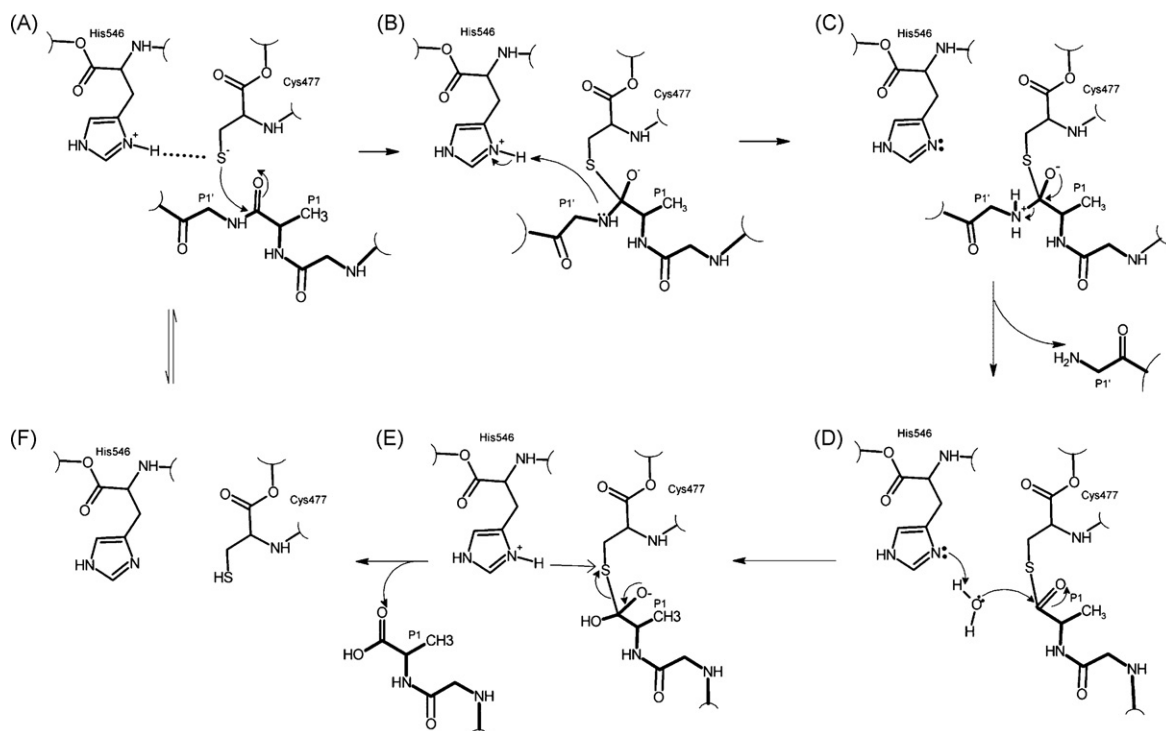


Fig. 2. Proposed catalytic cycle of VEEV nsP2 protease. (a) Catalytic dyad is in the ionized state represented by a thiolate/imidazolium ion pair. Substrate (indicated by boldface bonds) binds to the active site and the P1 carbonyl carbon is attacked by the deprotonated thiol. (b) The negatively charged tetrahedral intermediate is formed and the P1' amine is protonated by the imidazolium ion. (c) The unstable intermediate breaks down, releasing the C-terminal product and leaving an acylated thiol. (d) and (e) The acyl thiol undergoes hydrolysis releasing the N-terminal product and free enzyme with active ionized catalytic dyad or the inactive uncharged dyad (f).

The VEEV protease structure, together with the reaction mechanism proposed for related papain and papain-like proteases [17,18] provided insights into the catalytic mechanism of the VEEV nsP2 protease (Fig. 2). In its activated substrate-bound state, the nsP2 protease catalytic dyad likely exists as a Cys⁴⁷⁷ thiolate–His⁵⁴⁶ imidazolium ion pair [18]. In the apo nsP2 protease crystal structure, the His⁵⁴⁶ rotamer was not positioned to participate in the ion pair necessary for catalysis. However, electron density for the His⁵⁴⁶ side chain is weak and poorly defined, indicating this residue is conformationally flexible. A simple torsion of His⁵⁴⁶ positioned a rotamer to form the catalytically active thiolate-imidazolium ion pair with Cys⁴⁷⁷ (Fig. 2f). This His⁵⁴⁶ rotamer did not affect the backbone conformation and was energetically favourable. It is probable that binding of substrate (Fig. 2a) stabilizes His⁵⁴⁶ in the active form of the catalytic dyad, thus allowing the thiolate anion to attack the scissile carbonyl carbon and the reaction to proceed as presented in Fig. 2.

Structure-based predictions of side chain pK_a values for the catalytic dyad residues provide additional support for a thiolate-imidazolium ion pair catalytic mechanism for VEEV nsP2 protease. In papain, the pK_a values for the catalytic dyad cysteine and histidine residues were measured to be 3.3 and 8.5, respectively [19]. These values are substantially shifted from free amino acid pK_a values of 8.4 for cysteine and 6.0 for histidine [20]. The pK_a values for Cys⁴⁷⁷ and His⁵⁴⁶ within the nsP2 protease active conformation structure were calculated (PROPKA program) to be 5.1 and 8.0, respectively. These pK_a values are consistent with the existence of a thiolate-imidazolium ion pair at physiological pH in VEEV nsP2 protease.

3.2. VEEV S1' subsite

Simulated annealing of the protease–substrate complexes was performed with molecular dynamic simulations and energy mini-

mization to improve understanding of the molecular requirements for polyprotein processing and identify interactions between nsP2pro and its peptide substrates. In the following discussion we use the convention of Schechter and Berger [8], where S and P prefixes refer to complementary protease binding sites and peptide substrate residues, respectively. In this convention, substrate cleavage occurs between the P1' and P1 residues.

Limited intermolecular interactions between the S1' and P1' subsites are observed in the VEEV nsP2pro/peptide complex (Fig. 3, Table 1). The P1' residue of the VEEV nsP12, nsP23 and nsP34 polyprotein cleavage substrates is glycine, alanine and tyrosine, respectively. The S1' subsite within VEEV nsP2pro is defined by residues Ala⁴⁷⁴, Asn⁴⁷⁵, Cys⁴⁷⁷ and His⁵⁴⁶. Interactions between nsP2pro and the nsP12 and nsP23 cleavage sites are similar, and consist of van der Waals contacts between Asn⁴⁷⁵ and Cys⁴⁷⁷ and the P1' residue, and a hydrogen bond between His⁵⁴⁶ and the P1' amide nitrogen (Fig. 3). In addition, Ala⁴⁷⁴ interacts with the alanine C_β methyl group of the nsP23 P1' (Fig. 3, Table 1). Although these contacts are also observed between nsP2pro and the P1' tyrosine of the nsP34 cleavage site, an additional interaction between Lys⁴⁸⁰ and the P1' tyrosine of nsP34 occurs. This interaction can be considered as either a hydrogen bond between the lysine ε-amino and tyrosine hydroxyl groups or a cation–π interaction between lysine and the tyrosine ring. This additional interaction between the nsP34 cleavage site and the S1' is likely one of the dominant factors influencing the preference of the nsP2 protease for nsP34-based substrate peptides [6,7]. Other than the side chain interaction with Lys⁴⁸⁰, the observed contacts between nsP2pro and the P1' tyrosine residue are consistent with the other two nonstructural polyprotein cleavage sites. This is largely because the S1' subsite of VEEV nsP2pro is shallow and solvent exposed. Thus, contacts between protease and the substrate P1' residue are mostly through substrate backbone atoms.

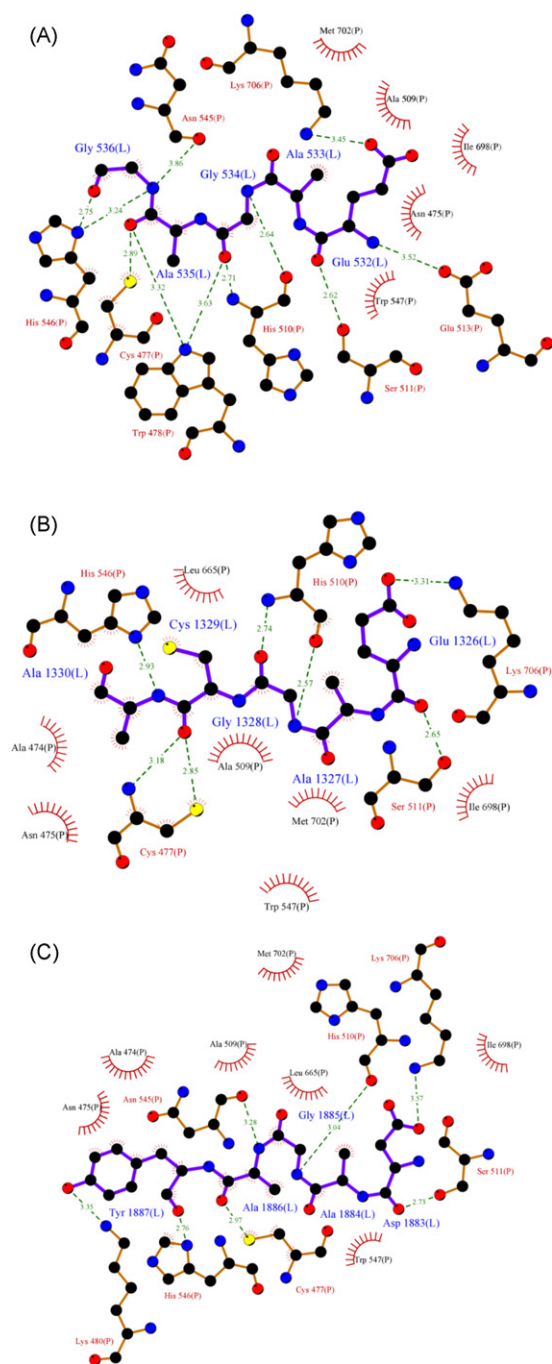


Fig. 3. Schematic of the nsP2 protease binding site with modelled substrates (a) nsP12 substrate, (b) nsP23 substrate and (c) nsP34 substrate. Substrate peptide bonds are purple and nsP2 protease bonds are brown. Substrate residues labels are blue. Numbering for peptide substrates is according to the sequence of VEEV nonstructural polyprotein. Schematics were produced with LIGPLOT [14].

3.3. VEEV S1 subsite

The P1 residue of VEEV polyprotein substrates binds to the S1 subsite of nsP2pro. For each polyprotein substrate, the catalytic thiol of Cys⁴⁷⁷ and the carbonyl carbon of the substrate scissile bond are correctly positioned for nucleophilic attack to occur (Figs. 2 and 3). To facilitate this reaction, Cys⁴⁷⁷ is deprotonated and stabilized by a thiolate-imidazolium ion pair with His⁵⁴⁶. For all cleavage sites within the VEEV nonstructural polyprotein, the P1 residue is either alanine or cysteine (Table 2). The nsP2 protease S1 subsite contains residues that make polar and nonpolar

Table 1

Interactions between VEEV nsP2 protease and peptide substrates.

	Substrate			Enzyme	Contact		
	nsP12	nsP23	nsP34		nsP12	nsP23	nsP34
P1'	Gly ⁵³⁶	Ala ¹³³⁰	Tyr ¹⁸⁶⁷	Ala ⁴⁷⁴		+	+
				Asn ⁴⁷⁵	+	+	+
				Cys ⁴⁷⁷	+	+	+
				Lys ⁴⁸⁰			± [†]
				His ⁵⁴⁶	±	±	±
P1	Ala ⁵³⁵	Cys ¹³²⁹	Ala ¹⁸⁶⁶	Asn ⁴⁷⁵	+	+	+
				Cys ⁴⁷⁷	+	±	+
				Trp ⁴⁷⁸	±		
				Ala ⁵⁰⁹	+	+	+
				Asn ⁵⁴⁵	±		±
				His ⁵⁴⁶	+	+	+
				Leu ⁶⁶⁵	+	+	+
P2	Gly ⁵³⁴	Gly ¹³²⁸	Gly ¹⁸⁶⁵	Cys ⁴⁷⁷	+		+
				Trp ⁴⁷⁸		+	
				Ala ⁵⁰⁹	+	+	+
				His ⁵¹⁰	±	±	±
				Trp ⁵⁴⁷	+	+	+
P3	Ala ⁵³³	Ala ¹³²⁷	Ala ¹⁸⁶⁴	His ⁵¹⁰	+	+	+
				Ile ⁵⁴²	+		+
				Trp ⁵⁴⁷		+	+
				Ile ⁶⁹⁸	+	+	+
				Met ⁷⁰²	+	+	+
P4	Glu ⁵³²	Glu ¹³²⁶	Asp ¹⁸⁶³	Ser ⁵¹¹	±	±	±
				Glu ⁵¹³	±		
				Trp ⁵⁴⁷	+	+	+
				Met ⁷⁰²			+
				Lys ⁷⁰⁶	±	±	±

+, VDW contact; ±, hydrogen bond; ±[†], indicates probable cation- π interaction. Contacts were determined using the contact of structural units (CSU) server [15] and LIGPLOT [14] with a 4.0 Å cut-off distance.

contacts to the P1 substrate residue (Table 1, Fig. 3). In the nsP23 substrate, the P1 carbonyl oxygen orientation relative to the Cys⁴⁷⁷ backbone amide is consistent with hydrogen bonding between these residues; in this case the Cys⁴⁷⁷ amide nitrogen serves as an oxyanion hole for the hydrolysis reaction. In contrast, molecular modelling of the nsP12 substrate suggested that the indole nitrogen of Trp⁴⁷⁸ might also function as part of the oxyanion hole. Molecular modelling of the nsP34 substrate revealed the P1 carbonyl oxygen is oriented such that conformational rearrangements expected to occur when the tetrahedral reaction intermediate is formed would position Cys⁴⁷⁷ and/or Trp⁴⁷⁸ to serve as the oxyanion hole during the hydrolysis reaction.

3.4. VEEV S2 subsite

Molecular modelling and molecular dynamic simulations of the binding between nsP2 protease and its varied polyprotein substrates identified crucial interactions between Trp⁵⁴⁷ in the S2 subsite and the P2 residue. In all examined alphavirus sequences nsP2 Trp⁵⁴⁷ is completely conserved and the P2 glycine is invariant in the nonstructural polyprotein cleavage targets. The orientation of Trp⁵⁴⁷ in the model restricts the volume of the S2 subsite such that it can only accommodate a P2 glycine residue. Modelling other residues at the P2 position, including alanine, resulted in energetically unfavourable steric clashes between the P2 residues and Trp⁵⁴⁷. These results provide a structure-based explanation of the correlation noted by Golubtsov et al. that cysteine proteases that cleaved substrates containing a P2 glycine had bulky aromatic residues immediately following the active site histidine [21]. This glycine specificity motif (GSM) was speculated to be involved in alphavirus nsP2 substrate recognition [21].

Table 2
Alphavirus sequence covariance for protease subsite residues S1'–S4 and polyprotein cleavage substrate residues P1'–P4.

Virus	S1' subsite residues						P1' residue		
							nsP12	nsP23	nsP34
VEEV	A ⁴⁷⁴	N ⁴⁷⁵	C ⁴⁷⁷	K ⁴⁸⁰	H ⁵⁴⁶	G	A	Y	
WEEV	V	–	–	–	–	–	–	–	
EEEV	–	–	–	–	–	–	–	–	
RRV	–	K	–	–	–	–	–	–	
SFV	–	–	–	–	–	–	–	–	
ONNV	–	–	–	–	–	–	–	–	
BFV	V	–	–	–	–	–	–	–	
SINV	T	–	–	–	–	A	–	–	
AURAV	V	–	–	–	–	A	–	–	
SPDV	R	–	–	V	–	–	S	–	
SDV	R	–	–	V	–	–	S	–	

Virus	S1 subsite residues							P1 residue		
								nsP12	nsP23	nsP34
VEEV	N ⁴⁷⁵	C ⁴⁷⁷	W ⁴⁷⁸	A ⁵⁰⁹	N ⁵⁴⁵	H ⁵⁴⁶	L ⁶⁶⁵	A	C	A
WEEV	–	–	–	–	Q	–	–	–	R	–
EEEV	–	–	–	–	E	–	–	–	R	–
RRV	K	–	–	–	–	–	–	–	–	–
SFV	–	–	–	–	–	–	–	–	–	–
ONNV	–	–	–	–	–	–	–	–	–	G
BFV	–	–	–	–	D	–	–	E	S	G
SINV	–	–	–	P	A	–	F	–	A	G
AURAV	–	–	–	P	A	–	F	–	A	G
SPDV	–	–	–	P	S	–	–	–	A	G
SDV	–	–	–	P	S	–	–	S	A	G

Virus	S2 subsite residues					P2 residue		
						nsP12	nsP23	nsP34
VEEV	W ⁴⁷⁸	C ⁴⁷⁷	A ⁵⁰⁹	H ⁵¹⁰	W ⁵⁴⁷	G	G	G
WEEV	–	–	–	Y	–	–	–	–
EEEV	–	–	–	F	–	–	–	–
RRV	–	–	–	Y	–	–	–	–
SFV	–	–	–	Y	–	–	–	–
ONNV	–	–	–	Y	–	–	–	–
BFV	–	–	–	Y	–	–	–	–
SINV	–	–	P	–	–	–	–	–
AURAV	–	–	P	Y	–	–	–	–
SPDV	–	–	P	–	–	–	–	–
SDV	–	–	P	–	–	–	–	–

Virus	S3 subsite residues					P3 residue		
						nsP12	nsP23	nsP34
VEEV	H ⁵¹⁰	I ⁵⁴²	W ⁵⁴⁷	I ⁶⁹⁸	M ⁷⁰²	A	A	A
WEEV	Y	Y	–	–	–	–	–	–
EEEV	F	Y	–	–	–	–	–	–
RRV	Y	Y	–	M	–	–	–	–
SFV	Y	Y	–	M	–	–	–	–
ONNV	Y	Y	–	M	–	–	–	–
BFV	Y	Y	–	M	–	–	–	–
SINV	–	Y	–	A	T	I	V	V
AURAV	Y	F	–	G	T	–	S	S
SPDV	–	L	–	V	A	–	V	V
SDV	–	L	–	V	A	V	V	V

Virus	S4 subsite residues					P4 residue		
						nsP12	nsP23	nsP34
VEEV	S ⁵¹¹	E ⁵¹³	W ⁵⁴⁷	M ⁷⁰²	K ⁷⁰⁶	E	E	D
WEEV	–	–	–	–	–	–	–	E
EEEV	–	–	–	–	–	–	–	E
RRV	–	–	–	–	D	Y	T	R
SFV	–	V	–	–	D	Y	T	R
ONNV	–	–	–	–	D	D	R	R
BFV	–	–	–	–	D	–	P	R
SINV	–	V	–	T	S	A	G	G
AURAV	–	M	–	T	S	D	G	G
SPDV	–	L	–	A	G	D	M	G
SDV	–	L	–	A	G	D	M	G

The S2 subsite is not defined exclusively by steric constraints imposed by Trp⁵⁴⁷. Several additional residues (e.g., Cys⁴⁷⁷, Ala⁵⁰⁹) contribute van der Waals interactions and the backbone carbonyl of His⁵¹⁰ forms a hydrogen bond to the amine of P2 glycine (Table 1). Although a nonglycine amino acid at the P2 position could undergo 180° backbone rotation to direct its side chain out of the S2 subsite and into the solvent, such a rotation would be energetically unfavourable since it would disrupt the hydrogen bonding between the cleavage substrate and His⁵¹⁰.

3.5. VEEV S3 and S4 subsites

The S3 subsite accommodates a small nonpolar residue. In VEEV, each cleavage substrate contains an alanine residue at the P3 position (Table 1). Contacts between S3 residues and the P3 amino acid are predominately hydrophobic (Fig. 3, Table 1). With the exception of Trp⁵⁴⁷, residues that form the S3 subsite are poorly conserved (Table 2).

VEEV cleavage substrates contain either aspartate or glutamate at the P4 position. The S4 subsite is formed from polar and nonpolar residues, although polar contacts dominate the interactions between P4 and S4 residues. In all cleavage substrate models, the side chain hydroxyl of Ser⁵¹¹ makes a hydrogen bond to the P4 backbone carbonyl, and appears to be important for correct positioning and orientation of the substrate peptide. In addition, Lys⁷⁰⁶ forms a salt bridge with the acidic P4 sidechain, and this contact is probably a significant determinant of substrate recognition.

3.6. Covariance of cleavage site (Pn) and binding subsite (Sn) sequences in alphaviruses

3.6.1. S1' subsite

The majority of P1' residues in alphavirus nsP12 and nsP23 cleavage substrates are glycine or alanine, although serine occurs at the nsP23 P1' position in the distantly related salmonid viruses (Table 2). Tyrosine is conserved at the nsP34 cleavage site P1' position in all examined alphavirus sequences.

Five residues, located at positions 474, 475, 477, 480 and 546, contribute to the S1 subsite. Cys⁴⁷⁷ and His⁵⁴⁶ are completely conserved in all alphaviruses. Asn⁴⁷⁵ is highly conserved among alphaviruses (Table 2). Alanine is the predominant residue at position 474, although several differences are observed at this position in other alphavirus strains. Only backbone atoms of Ala⁴⁷⁴ interacted with the nsP23 and nsP34 P1' residue. This lack of sidechain involvement in S1' specificity removes constraints on sidechain identity and enables substitutions at residue 474 to be accommodated. Lys⁴⁸⁰ is conserved in the nonsalmonid alphaviruses and valine replaces Lys⁴⁸⁰ in the salmonid alphaviruses. Lys⁴⁸⁰ interacts with the P1' tyrosine of nsP34 cleavage substrate through either a cation– π interaction with the tyrosine ring or hydrogen bond with the tyrosine hydroxyl and provides a strong interaction that dictates nsP34 specificity. This interaction is absent in the salmonid alphaviruses, since Lys⁴⁸⁰ is replaced by valine. However, in the salmonid viruses a compensatory change of Ala⁴⁷⁴ to Arg⁴⁷⁴ in the S1' subsite preserves the specific cation– π interaction (Table 2).

3.6.2. S1 subsite

Within and between alphavirus strains, the S1 subsite accommodates diverse P1 residues. In Western equine encephalitis virus (WEEV), the P1 residue is either alanine or arginine; in Barmah Forest virus (BFV) the P1 residue is glutamate, serine or glycine. The S1 subsite primarily interacts with the C α and carbonyl carbon atoms of the P1 residue. The P1 backbone orientation enables long flexible sidechains (e.g., arginine) to adopt conformations that position the P1 sidechain beyond the S1 subsite and in solvent. Branched or aromatic sidechains at the P1 position do not have the necessary

flexibility to circumvent the steric constraints of the S1 subsite, and thus are not observed at the P1 position of cleavage substrates.

Seven residues, located at positions 475, 477, 478, 509, 545, 546 and 665, form the alphavirus S1 subsite (Table 2), although residue 478 does not interact with nsP23 and nsP34 cleavage substrates. Residues that delineate the S1 subsite are largely conserved among alphaviruses, with substitutions generally restricted to chemically similar residues (Table 2). Asn⁵⁴⁵ interacts with the P1 residue through hydrogen bonding between their backbone amides; this S1 residue can tolerate diverse substitutions because its sidechain does not participate in substrate binding. Covariant substitutions between the P1 and S1 subsite residues are not readily apparent, implying this site is relatively tolerant of amino acid substitutions.

Lulla et al. [5] observed that the SFV nsP2 protease cleaved synthetic substrates containing alanine, serine and glycine (and presumably cysteine) at the P1 position, whereas substrates containing P1 glutamate and arginine residues were not cleaved. Glutamate is observed at P1 in BFV cleavage substrates, while arginine is observed at P1 in WEEV and EEEV cleavage substrates. Nearly all S1 residues are conserved among these four alphavirus strains. The notable exception is Asn⁵⁴⁵ whose contributions to the S1 pocket are through backbone atoms as discussed previously. Interestingly, our molecular modelling studies do not provide a structure-based explanation as to why the SFV nsP2 protease was observed to be inactive against substrates with P1 arginine or glutamate, whereas other alphavirus strains with >95% sequence identity in the S1 subsite are active against these substrates. Future work will investigate the thermodynamic factors that may be responsible for the observed substrate specificities.

3.6.3. S2 subsite

The P2 residue in all alphavirus polyprotein cleavage substrates is glycine, which interacts with the S2 subsite formed in part by conserved residues Cys⁴⁷⁷, Trp⁴⁷⁸ and Trp⁵⁴⁷. Residues 509 and 510 also contribute to the S2 subsite, with alanine or proline residues found at position 509 and histidine, tyrosine or phenylalanine residues found at position 510. The conservative substitutions observed at positions 509 and 510 do not significantly affect the binding characteristics or geometry of the S2 subsite. The C α atom of Ala⁵⁰⁹ makes hydrophobic contacts with the C α atom of the P2 glycine; this interaction is preserved when proline is substituted at this position. Residue 510 interacts with the P2 glycine through backbone hydrogen bonds, with the planar ring sidechain buried within the interface between nsP2pro domains. Thus, substitution of His⁵¹⁰ with tyrosine or phenylalanine maintains the local structure of the S2 subsite, does not disrupt backbone interactions with the P2 residue, and preserves substrate specificity.

3.6.4. S3 subsite

Within each alphavirus strain, the P3 residue is almost completely conserved among the nsP12, nsP23 and nsP34 polyprotein cleavage substrates. Alanine occupies the P3 position of the polyprotein cleavage substrates in the majority of alphavirus strains examined. However, in Sindbis, Aura, and salmonid alphaviruses, valine, isoleucine and serine are found at the P3 position. These substitutions increase sidechain volume, and result in steric clashes between the P3 and S3 residues unless there are compensatory substitutions within the S3 subsite (Fig. 4).

Residues 510, 542, 547, 698 and 702 form the S3 subsite. Trp⁵⁴⁷ is invariant, with conservative substitutions occurring for His⁵¹⁰ and Ile⁶⁹⁸ (Table 2). Variations in S3 residues are highly correlated with changes at P3 residue. In Sindbis, Aura, and salmonid viruses, the P3 valine or serine residues are >50 Å³ larger than the canonical Ala residue. In Sindbis and Aura virus nsP2 proteases, threonine replaces the canonical Met⁷⁰², resulting in ~50 Å³ increase in the volume of the S3 subsite relative to the VEEV S3 subsite (Fig. 4b). In

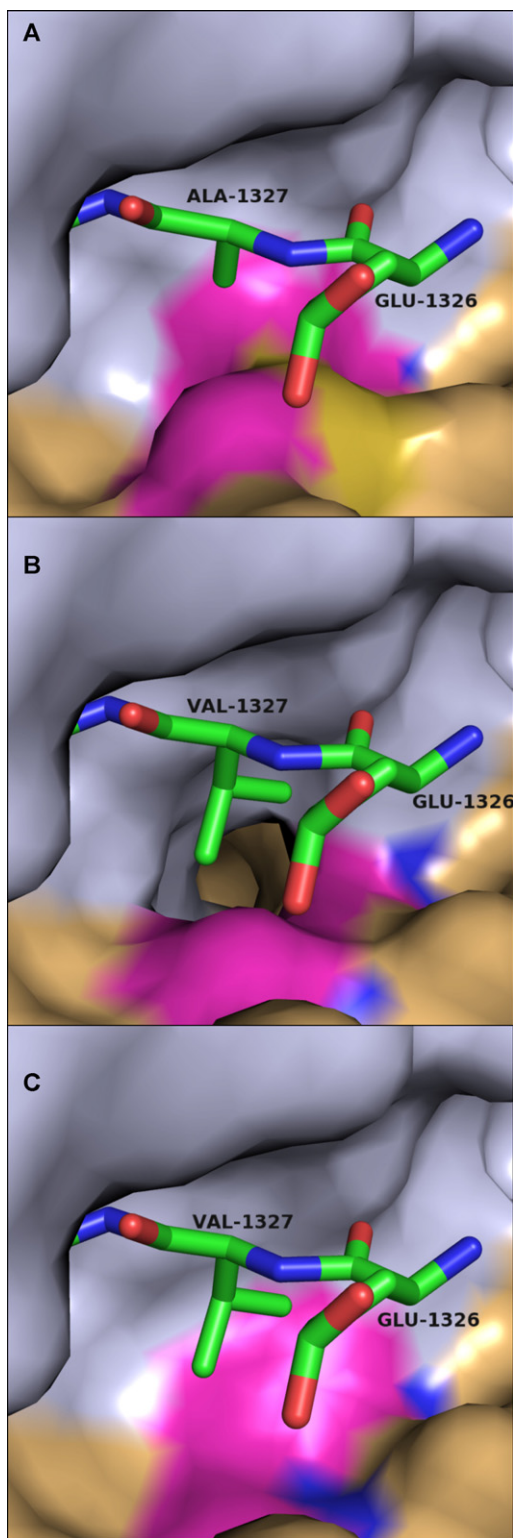


Fig. 4. Covariance of amino acid substitutions in different alphaviruses strains at the S3 subsite and the P3 cleavage substrate. (a) VEEV nsP2 protease/nsP23 substrate complex with P3 alanine (Ala¹³²⁷). S3 residues Met⁷⁰² and Ile⁶⁹⁸ are highlighted. (b) Complex as modelled in the Aura virus with a P3 valine, threonine at position 702 and glycine at position 698. (c) Complex as modelled in the salmonid alphaviruses (salmon pancreas disease virus and sleeping disease virus) with P3 valine, alanine at position 702 and valine at position 698. Substrate peptide is shown as sticks. Carbon atoms are green, nitrogen blue and oxygen red. The nsP2 protease surface is shown with the N-terminal domain colored gray and the C-terminal domain colored tan. S3 subsite residues exhibiting significant variation are highlighted in magenta. All residue numbers correspond to VEEV numbering convention.

addition, Ile⁶⁹⁸ is replaced by small hydrophobic sidechains in the Sindbis and Aura viruses, resulting in S3 volume increases of $>80 \text{ \AA}^3$ relative to Ile⁶⁹⁸ (Fig. 4b). In the salmonid alphaviruses, increases in the volume of the S3 subsite are achieved through substitution of alanine for Met⁷⁰² and valine for Ile⁶⁹⁸ (Table 2, Fig. 4c). In each case, the increased volume of the S3 subsite exceeds the volume increase of the P3 sidechain. Thus, residue substitutions that increase the size of the P3 sidechain are complemented with covariant S3 substitutions that increase the volume of the S3 subsite.

3.6.5. S4 subsite

Diverse residues occur at the P4 position of the polyprotein cleavage substrates both within and between alphaviruses. Residues 511, 513, 547, 702 and 706 form the S4 subsite. Interestingly, subsite residues Ser⁵¹¹ and Trp⁵⁴⁷ are conserved, although the P4 residue is highly variable. Little correlation is observed between most S4 residues and the P4 residue. However, variations in S4 subsite residue 706 and the P4 substrate residue are highly correlated, implying this interaction is a key determinant of substrate specificity. In the equine encephalitis viruses, residue 706 is lysine and the nsP34 P4 residue is acidic, allowing these residues to interact through a salt bridge (Fig. 3). In Ross River, Semliki forest, O'nyong-nyong and Barmah Forest viruses, residue 706 is aspartic acid and the nsP34 P4 residue is arginine, which maintains the salt bridge between these residues. In the nsP12 and nsP23 cleavage substrates, covariance between the P4 residue and residue 706 largely preserves the salt bridge and hydrogen bond interactions between these residues.

4. Conclusion

Proteolytic processing of alphavirus nonstructural polyproteins requires the nsP2 protease to differentiate between three similar cleavage sites within the nonstructural polyprotein so that polyprotein cleavage occurs in the correct order and at the correct time during viral replication cycle. This work elucidates structural features of the nsP2 protease domain/substrate interaction that contribute to the catalytic mechanism, substrate specificity and discrimination between the nonstructural polyprotein cleavage sites and indicates how specific mutations to the cleavage sites and substrates might affect the protease–substrate interaction. Work by Lulla et al. [5] suggested that major determinants of specificity between cleavage sites within the SFV nonstructural polyprotein reside in substrate residues P1' through P4, with the influence of the nonprime residues dominating specificity. They implicated P4 as a major determinant of substrate specificity in SFV. The covariance between S4 subsite residue 706 and the P4 substrate residue, which maintains a salt bridge in the majority of alphavirus strains, provides a structural basis for the observed specificity. In addition, the P1' residue contributes to substrate specificity due to cation– π interactions between the invariant P1' tyrosine of the nsP34 cleavage site and conserved basic residues in the S1' subsite; this interaction likely accounts for the preferential cleavage of this substrate.

We identified several structural determinants of substrate specificity that recognize features common to all three cleavage sites within the viral polyprotein. These determinants are strongly conserved among alphaviruses. Examination of contacts between S2/P2 and S3/P3 residues illustrate preserved binding motifs among nonstructural polyprotein cleavage sites such as the ubiquitous P2 glycine interaction with Trp⁵⁴⁷ and hydrogen bonding between P2 and His⁵¹⁰. Important substrate binding interactions involving P3 are preserved by covariance of residues that form the S3 subsite; compensatory residue changes maintain similar volumes for the P3 sidechain and S3 subsite. These observations benefit drug discov-

ery efforts by identifying target sites and interactions on the nsP2 protease that might lead to either broad-spectrum anti-alphaviral drugs or highly specific treatments for individual alphavirus diseases.

Acknowledgements

We thank Dr. M. Karplus for kindly providing the CHARMM Molecular Dynamic Software. This work was supported in part by NIAID/NIH AI53551 (SJW). MAW was supported by a grant from the Sealy and Smith Foundation to the Sealy Center for Structural Biology and Molecular Biophysics.

References

- [1] J.H. Strauss, E.G. Strauss, The alphaviruses: gene expression, replication, and evolution, *Microbiol. Rev.* 58 (1994) 491–562.
- [2] C. Lahariya, S.K. Pradhan, Emergence of chikungunya virus in Indian subcontinent after 32 years: a review, *J. Vector Borne Dis.* 43 (2006) 151–160.
- [3] N/A, Outbreak news. Chikungunya. India, *Wkly. Epidemiol. Rec.* 81 (2006) 409–410.
- [4] P.V. Aguilar, S. Paessler, A.S. Carrara, S. Baron, J. Poast, E. Wang, A.C. Moncayo, M. Anishchenko, D. Watts, R.B. Tesh, S.C. Weaver, Variation in interferon sensitivity and induction among strains of eastern equine encephalitis virus, *J. Virol.* 79 (2005) 11300–11310.
- [5] A. Lulla, V. Lulla, K. Tints, T. Ahola, A. Merits, Molecular determinants of substrate specificity for Semliki Forest virus nonstructural protease, *J. Virol.* 80 (2006) 5413–5422.
- [6] L. Vasiljeva, A. Merits, A. Golubtsov, V. Sizemskaja, L. Kaariainen, T. Ahola, Regulation of the sequential processing of Semliki Forest virus replicase polyprotein, *J. Biol. Chem.* 278 (2003) 41636–41645.
- [7] L. Vasiljeva, L. Valmu, L. Kaariainen, A. Merits, Site-specific protease activity of the carboxyl-terminal domain of Semliki Forest virus replicase protein nsP2, *J. Biol. Chem.* 276 (2001) 30786–30793.
- [8] I. Schechter, A. Berger, On the size of the active site in proteases. I. Papain, *Biochem. Biophys. Res. Commun.* 27 (1967) 157–162.
- [9] A.T. Russo, M.A. White, S.J. Watowich, The crystal structure of the Venezuelan equine encephalitis alphavirus nsP2 protease, *Structure* 14 (2006) 1449–1458.
- [10] E. Mossessova, C.D. Lima, Ulp1-SUMO crystal structure and genetic analysis reveal conserved interactions and a regulatory element essential for cell growth in yeast, *Mol. Cell.* 5 (2000) 865–876.
- [11] B.R. Brooks, C.L. Brooks III, A.D. Mackerell Jr., L. Nilsson, R.J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caffisch, L. Caves, Q. Cui, A.R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoseck, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R.W. Pastor, C.B. Post, J.Z. Pu, M. Schaefer, B. Tidor, R.M. Venable, H.L. Woodcock, X. Wu, W. Yang, D.M. York, M. Karplus, CHARMM: the biomolecular simulation program, *J. Comput. Chem.* 30 (2009) 1545–1614.
- [12] A.T. Brunger, M. Karplus, Polar hydrogen positions in proteins: empirical energy placement and neutron diffraction comparison, *Proteins* 4 (1988) 148–156.
- [13] A.D. Mackerell, D. Bashford, Bellott, R.L. Dunbrack, J.D. Evanseck, M.J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F.T.K. Lau, C. Mattos, S. Michnick, T. Ngo, D.T. Nguyen, B. Prodhom, W.E. Reiher, B. Roux, M. Schlenkrich, J.C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, M. Karplus, All-atom empirical potential for molecular modeling and dynamics studies of proteins, *J. Phys. Chem. B* 102 (1998) 3586–3616.
- [14] A.C. Wallace, R.A. Laskowski, J.M. Thornton, LIGPLOT: a program to generate schematic diagrams of protein–ligand interactions, *Protein Eng.* 8 (1995) 127–134.
- [15] V. Sobolev, A. Sorokine, J. Prilusky, E.E. Abola, M. Edelman, Automated analysis of interatomic contacts in proteins, *Bioinformatics* 15 (1999) 327–332.
- [16] H. Li, A.D. Robertson, J.H. Jensen, Very fast empirical prediction and rationalization of protein pK_a values, *Proteins* 61 (2005) 704–721.
- [17] A.J. Beveridge, A theoretical study of the active sites of papain and S195C rat trypsin: implications for the low reactivity of mutant serine proteinases, *Protein Sci.* 5 (1996) 1355–1365.
- [18] L. Polgar, Mercaptide–imidazolium ion-pair: the reactive nucleophile in papain catalysis, *FEBS Lett.* 47 (1974) 15–18.
- [19] S.D. Lewis, F.A. Johnson, J.A. Shafer, Potentiometric determination of ionizations at the active site of papain, *Biochemistry* 15 (1976) 5009–5017.
- [20] R.M.C. Dawson, D.C. Elliot, W.H. Elliot, K.M. Jones, *Data for Biochemical Research*, 3rd ed., Clarendon Press, Oxford, 1989.
- [21] A. Golubtsov, L. Kaariainen, J. Caldentey, Characterization of the cysteine protease domain of Semliki Forest virus replicase protein nsP2 by in vitro mutagenesis, *FEBS Lett.* 580 (2006) 1502–1508.