Molecular Mechanisms of Viral and Host Cell Substrate Recognition by Hepatitis C Virus NS3/4A Protease

Keith P. Romano, Jennifer M. Laine, Laura M. Deveau, Hong Cao, Francesca Massi, and Celia A. Schiffer*

University of Massachusetts Medical School, Department of Biochemistry and Molecular Pharmacology, Worcester, Massachusetts 01605

Received 23 February 2011/Accepted 4 April 2011

Hepatitis C NS3/4A protease is a prime therapeutic target that is responsible for cleaving the viral polyprotein at junctions 3-4A, 4A4B, 4B5A, and 5A5B and two host cell adaptor proteins of the innate immune response, TRIF and MAVS. In this study, NS3/4A crystal structures of both host cell cleavage sites were determined and compared to the crystal structures of viral substrates. Two distinct protease conformations were observed and correlated with substrate specificity: (i) 3-4A, 4A4B, 5A5B, and MAVS, which are processed more efficiently by the protease, form extensive electrostatic networks when in complex with the protease, and (ii) TRIF and 4B5A, which contain polyproline motifs in their full-length sequences, do not form electrostatic networks in their crystal complexes. These findings provide mechanistic insights into NS3/4A substrate recognition, which may assist in a more rational approach to inhibitor design in the face of the rapid acquisition of resistance.

Hepatitis C virus (HCV) is a genetically diverse member of the genus Hepacivirus of the Flaviviridae family and infects over 180 million people worldwide (33). HCV contains a positive, single-stranded RNA genome that is translated as a single polyprotein along the endoplasmic reticulum by host cell machinery. The viral polyprotein is subsequently processed by host cell and viral proteases into structural (C, E1, and E2) and nonstructural (p7, NS2, NS3, NS4A, NS4B, NS5A, and NS5B) components (23). The NS3/4A protein, a bifunctional protease/helicase enzyme formed by the noncovalent association of NS3 and NS4A, hydrolyzes four known sites along the viral polyprotein, thereby liberating nonstructural proteins essential for viral replication. Previous kinetic data suggest that the first cleavage event at junction 3-4A occurs in cis as a unimolecular process, while processing of the remaining junctions 4A4B, 4B5A, and 5A5B occurs bimolecularly in trans (2, 22). Interestingly, these data demonstrated that the NS4A sequence is essential for the cleavage of junction 4B5A. These viral substrates share little sequence similarity except for an acid at P6, cysteine or threonine at P1, and serine or alanine at P1’ (Table 1). Previous work by our group has revealed that the diverse set of NS3/4A substrate sequences are recognized in a conserved three-dimensional shape, defining a consensus van der Waals volume, or substrate envelope (29). This conserved mode of substrate recognition regulates polyprotein processing and thus the biology of HCV replication.

In addition to its essential role in processing the viral polyprotein, NS3/4A protease also confounds the innate immune response to viral infection by disrupting activation of the transcription factors interferon regulatory factor 3 (IRF-3) and nuclear factor xB (NF-xB) (13, 21). Upon the detection of viral RNA in host cells, these transcription factors are induced through two distinct pathways involving signaling through Toll-like receptor 3 (TLR3) or retinoic acid-inducible gene I (RIG-I) (1, 37). NS3/4A protease disrupts the TLR3 and RIG-I cascades by cleaving the essential adaptor proteins Toll-interleukin-1 receptor domain-containing adaptor-inducing beta interferon (TRIF) and mitochondrial antiviral signaling protein (MAVS), respectively (12, 21). The TRIF and MAVS cleavage sites share little sequence homology with other viral substrates: TRIF contains cysteine at P1 and serine at P1’, while MAVS contains glutamate at P6 and cysteine at P1 (Table 1). Notably, in place of an acid at position P6, TRIF consists of a track of eight proline residues spanning P13 to P6, which has been previously implicated as part of the substrate recognition motif for NS3/4A (11). The NS3-mediated processing of both viral and host cell targets is central to the interplay between HCV replication and the innate immune response, thus highlighting the importance of better elucidating the mechanisms of substrate recognition.

Despite great efforts devoted to the development of NS3/4A protease inhibitors, the rapid rise of drug resistance in human clinical trials has limited the efficacy of the most promising drug candidates. Drug resistance mutations in the protease emerge as molecular changes that prevent the binding of drugs but still permit the recognition and cleavage of substrates. A more detailed understanding of the molecular details underlying substrate recognition is therefore critical for explaining patterns of drug resistance and for designing novel drugs that are less susceptible to resistance. Here we analyze crystal structures of NS3/4A protease in complex with N-terminal products of viral substrates 3-4A, 4A4B, 4B5A, and 5A5B. TRIF and MAVS crystal complexes further reveal that these host cell products bind to the protease active site in a conserved three-dimensional manner similar to that of the viral products. No-
tably, extensive electrostatic networks involving protease residues D81, R155, D168, and R123 form in product complexes.

Viral substrate and peptide product purchase and storage. Thirty milligrams of each substrate, with the same N-terminal sequence (20 to 26% polyethylene glycol [PEG] 3350, 0.1 M sodium MES buffer at pH 6.5, and 4% ammonium sulfate) in 24-well VDX hanging-drop trays.

Data collection and structure solution. Crystals large enough for data collection were flash-frozen in liquid nitrogen for storage. The apoenzyme, TRIF, and MAVS crystals were mounted under a constant cryostream, and X-ray diffraction data were collected at Advanced Photon Source LS-CAT 21-ID-F, BioCARS 14-BMC, and our in-house RAXIS IV X-ray system, respectively. Diffraction intensities of product complexes were indexed, integrated, and scaled using the program HKL2000 (27). All structure solutions were generated using simple isomorphous molecular replacement with PHASER (24). The B-chain model of viral substrate product 4A4B (3M5M) (29) was used as the starting model for all structure solutions. Initial refinement was carried out in the absence of modeled ligand, which was subsequently built in during later stages of refinement. Upon obtaining the correct molecular replacement solutions, the phases were improved by building solvent molecules using ARP/WARP (25). Subsequent crystallographic refinement was carried out within the CCP4 program suite with iterative rounds of TLS and restrained refinement until convergence was achieved (6). The final structures were evaluated with MolProbity (7) prior to deposition in the Protein Data Bank. Five percent of the data was reserved for the free R-value calculation to limit the possibility of model bias throughout the refinement process (4). Interactive model building and electron density viewing were carried out using the program COOT (10).

Double-difference plots and global analysis. Double-difference plots were computed as described previously (28). Briefly, the atomic distances between each Ca of a given protease molecule and every other Ca in the same molecule were calculated. The differences of these Ca-Ca distances between each pair of protease molecules were then calculated and plotted as a contour graph for visualization. These analyses allowed for effective structural comparisons without the biases associated with superimpositions and space group differences. Double-difference plots were used to determine the structurally invariant regions of the protease, consisting of residues 32 to 36, 42 to 47, 50 to 54, 84 to 86, and 140 to 143. Structural superimpositions were carried out in MoPro (9) using the Ca atoms of these residues for all protease complexes. The apoenzyme structure was used as the reference structure for the alignments of the TRIF and MAVS product complexes. The Ca root mean square deviation (RMSD) for each residue was subsequently calculated to assess the degree of structural variation throughout the protein. The B-factor column of a representative structure was replaced with these values and used to generate the rainbow color spectrum to visualize these variations.

Viral substrate envelope. All active-site alignments were performed with PyMOL using the Ca atoms of protease residues 137 to 139 and 154 to 160. For each alignment, the B chain of complex 4A4B was used as the reference structure. The NS3/4A viral substrate envelope, representing the consensus van der Waals volume shared by any three of the four viral substrate products, was computed as described previously using the full-length NS3/4A structure (1CU1) (35) and product complexes 4A4B (3M5M), 4B5A (3M5N), and 5A5B (3M5O) (29). Active-site comparisons of the four NS3/4A viral product complexes were performed by superimposition of the Ca atoms of residues 137 to 139 and 154 to 160, revealing that both the active-site residues and substrate products spanning P6 to P1 align closely, with average Ca RMSDs of 0.24 Å and 0.35 Å, respectively.

NMR spectroscopy and data processing. Nuclear magnetic resonance (NMR) data were recorded using 650 μl of 390 μM [U-15N]NS3/4A protease [95% H2O/5% D2O, 25 mM sodium phosphate at pH 7.2, 150 mM KCl, 5 μM zinc chloride, and 1 mM tris(carboxymethyl)phosphine (TCEP)]. Backbone 1H and 15N resonance assignments for NS3/4A protease were kindly provided by Herbert Klei of Bristol-Myers Squibb and were confirmed using nuclear Overhauser enhancement spectroscopy (NOESY) experiments. Backbone 1H and 15N resonance assignments of NS3/4A protease bound to the N-terminal cleavage prod-

### Table 1. Primary cleavage sequences of NS3/4A substrates (genotype 1a)

<table>
<thead>
<tr>
<th>Substrate</th>
<th>P6</th>
<th>P5</th>
<th>P4</th>
<th>P3</th>
<th>P2</th>
<th>P1</th>
<th>P1'</th>
<th>P2'</th>
<th>P3'</th>
<th>P4'</th>
</tr>
</thead>
<tbody>
<tr>
<td>TRIF</td>
<td>P3</td>
<td>S</td>
<td>S</td>
<td>T</td>
<td>P</td>
<td>C</td>
<td>S</td>
<td>A</td>
<td>H</td>
<td>L</td>
</tr>
<tr>
<td>MAVS</td>
<td>E</td>
<td>R</td>
<td>E</td>
<td>V</td>
<td>P</td>
<td>C</td>
<td>H</td>
<td>R</td>
<td>P</td>
<td>S</td>
</tr>
<tr>
<td>3-4A</td>
<td>D</td>
<td>L</td>
<td>E</td>
<td>V</td>
<td>V</td>
<td>T</td>
<td>S</td>
<td>T</td>
<td>W</td>
<td>V</td>
</tr>
<tr>
<td>4A4B</td>
<td>D</td>
<td>E</td>
<td>M</td>
<td>E</td>
<td>E</td>
<td>C</td>
<td>S</td>
<td>Q</td>
<td>H</td>
<td>L</td>
</tr>
<tr>
<td>4B5A</td>
<td>E</td>
<td>C</td>
<td>T</td>
<td>T</td>
<td>P</td>
<td>C</td>
<td>S</td>
<td>G</td>
<td>S</td>
<td>W</td>
</tr>
<tr>
<td>5A5B</td>
<td>E</td>
<td>D</td>
<td>V</td>
<td>V</td>
<td>C</td>
<td>D</td>
<td>S</td>
<td>M</td>
<td>S</td>
<td>Y</td>
</tr>
</tbody>
</table>

**Materials and Methods**

Mutagenesis and gene information. The HCV genotype 1a NS3/4A protease gene described in a Bristol-Meyers Squibb patent (34) was synthesized by GenScript and cloned into the pET28a expression vector (Novagen). The gene encodes a highly soluble form of the NS3/4A protease domain as a single chain, with 11 core amino acids of NS4A located at the N terminus. The inactive S139A protease variant and S139A/K36A double mutant were subsequently constructed using the QuiChange site-directed mutagenesis kit from Stratagene and sequenced by Davis Sequencing for confirmation.

Expression and purification of NS3/4A protease constructs. The NS3/4A enzyme was expressed in Escherichia coli BL21(DE3) (31) and purified using nickel-affinity chromatography.

NMR spectroscopy and data processing. Nuclear magnetic resonance (NMR) data were recorded using 650 μl of 390 μM [U-15N]NS3/4A protease [95% H2O/5% D2O, 25 mM sodium phosphate at pH 7.2, 150 mM KCl, 5 μM zinc chloride, and 1 mM tris(carboxymethyl)phosphine (TCEP)]. Backbone 1H and 15N resonance assignments for NS3/4A protease were kindly provided by Herbert Klei of Bristol-Myers Squibb and were confirmed using nuclear Overhauser enhancement spectroscopy (NOESY) experiments. Backbone 1H and 15N resonance assignments of NS3/4A protease bound to the N-terminal cleavage prod-

Downloaded from http://jvi.asm.org on October 17, 2017 by guest
uct of substrate 4A4B were obtained from the assignment of the free protein by following the chemical shift changes upon titration of the ligand. All experiments were performed at 298 K using a Varian Inova spectrometer operating at 600 MHz (14.1 T). Spectra were processed using nmrPipe (8) and Sparky (15).

Binding of the unlabeled peptide corresponding to the N-terminal cleavage product of substrate 4A4B to [U-15N]NS3/4A protease was monitored using a series of two-dimensional 15N-H heteronuclear single-quantum coherence (HSQC) spectra collected at increasing concentrations of the peptide to a final concentration of 2.5 mM. The change of the cross-peak positions for NS3/4A residues was recorded as a function of titrated peptide concentration. The normalized change in the chemical shift was calculated for each protease residue using the equation: 

\[
\Delta = \frac{H_{\text{free}} - H_{\text{bound}}}{\Delta H}
\]

where \(\Delta H\) is the chemical shift change observed between the free and bound states, \(\gamma\) is the gyromagnetic ratio, and the subscript indicates either H or 15N nuclei. The b factors for each residue of the apoenzyme crystal structure were then replaced by the maximal chemical shift changes from the titration data. The crystal structure was colored in PyMOL (9) according to the chemical shift magnitudes to graphically depict the locations of the shifting residues.

van der Waals contact energy, van der Waals contact energies between protease residues and peptide products were computed using a simplified Lennard-Jones potential as described previously (26). Briefly, the Lennard-Jones potential \(V_r\) was calculated for each protease-product atom pair, where \(\sigma\) and \(\epsilon\) represent the interatomic distance, van der Waals well depth, and atomic diameter, respectively: 

\[
V_r = 4\epsilon[(\sigma/r)^{12} - (\sigma/r)^{6}]
\]

was computed for all possible protease-protein atom pairs within 5 Å, and potentials for nonbonded pairs separated by less than the distance at the minimum potential were equated to zero. Using this simplified potential value for each nonbonded protease-protein atom pair, the total van der Waals contact energy \(\Sigma V_r\) was computed for each peptide residue. For graphical convenience, van der Waals energy indexes were then calculated by multiplying the raw values by a factor of 10.

Fluorescence polarization. For fluorescence polarization experiments, the NS3/4A protease domain was purified in buffer (25 mM HEPES [pH 7.5], 150 mM NaCl, 20% glycerol, 4 mM DTT) as described above and subsequently concentrated to 200 to 400 μM. The concentrated stocks were then two-fold diluted in 384-well plates (Corning) in reaction buffer using the Victor-3 plate reader (Perkin-Elmer). Five sets of binding data were recorded as a function of titrated peptide concentration. The total NS3/4A concentration, \(K_d\) represents the postcleavage product 3-4A. The seven structures constituting a total of 12 NS3/4A protease monomers, which all adopt the same chymotrypsin-like tertiary fold defined by the labeling scheme for trypsin (3). The N-terminal distorted β-barrel subdomain contains two α helices (α0 and α1) and seven β strands (A0 to F1), while the C-terminal β-barrel subdomain comprises two α helices (α2 and α3) and six β strands (A2 to F2). The cofactor NS4A contributes a single β strand to the N-terminal distorted β barrel, which is essential for efficient catalytic function (22). The catalytic triad is located in the cleft between these subdomains, with the N-terminal β barrel contributing residues H57 and D81 and the C-terminal β barrel contributing the nucleophile S139. The active-site residues in all crystal structures share a similar architecture, defined by the catalytic triad residues (S139A, H57, and D81) and backbone nitrogens of the oxoanion hole (G137, S138, and S139A).

Certain global differences are observed between these struc-

### RESULTS

Structure determination of apo-NS3/4A and product complexes. The apo-NS3/4A protease domain and host cell product complexes TRIF and MAVS all crystallized in the space group P2_12_12, with one molecule in the asymmetric unit (Table 2). For all structures, we utilized the highly soluble NS3/4A protease domain described previously, containing the essential residues of cofactor NS4A covalently linked at the N terminus (34). This NS3/4A construct also contains the inactivating mutation S139A, which is designed to further enhance protein stability by minimizing autoproteolysis during the crystallization process. The partially inactivated variant still exhibits residual proteolytic activity, as observed for other serine proteases (5, 18), likely facilitated by the nucleophilic attack of water. Thus, the complete characterization of full-length substrate peptides was not possible, and short peptides corresponding to the N-terminal cleavage products of authentic substrates were used for all crystallization trials. Peptide products 4A4B, 4B5A, 5A5B, and MAVS spanned P7 to P1, while the TRIF sequence spanned P13 to P1 with a track of eight proline residues at the N terminus (Table 1). The entire peptide sequence spanning P7 to P1 could be modeled in each structure except for the TRIF complex, which revealed electron density for the residues spanning P6 to P1 but not the proline track.

Tertiary structure analysis. Structural analyses of the NS3/4A apoenzyme were carried out in conjunction with (i) the host cell product complexes TRIF and MAVS, (ii) the viral product complexes 4A4B, 4B5A, and 5A5B (29), and (iii) the full-length NS3/4A structure (35), in which the C terminus represents the postcleavage product 3-4A. The seven structures constituting a total of 12 NS3/4A protease monomers, which all adopt the same chymotrypsin-like tertiary fold defined by the labeling scheme for trypsin (3). The N-terminal distorted β-barrel subdomain contains two α helices (α0 and α1) and seven β strands (A0 to F1), while the C-terminal β-barrel subdomain comprises two α helices (α2 and α3) and six β strands (A2 to F2). The cofactor NS4A contributes a single β strand to the N-terminal distorted β barrel, which is essential for efficient catalytic function (22). The catalytic triad is located in the cleft between these subdomains, with the N-terminal β barrel contributing residues H57 and D81 and the C-terminal β barrel contributing the nucleophile S139. The active-site residues in all crystal structures share a similar architecture, defined by the catalytic triad residues (S139A, H57, and D81) and backbone nitrogens of the oxoanion hole (G137, S138, and S139A).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value for:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Protein Data</td>
<td>TRIF</td>
</tr>
<tr>
<td>Bank no.</td>
<td>3RC4</td>
</tr>
<tr>
<td>Modeled ligand</td>
<td>PSSTPC</td>
</tr>
<tr>
<td>Resolution (Å)</td>
<td>1.50 (1.50–1.55)</td>
</tr>
<tr>
<td>Space group</td>
<td>P2_12_12</td>
</tr>
<tr>
<td>Molecules in asymmetric unit</td>
<td>1</td>
</tr>
<tr>
<td>Cell dimensions (Å)</td>
<td></td>
</tr>
<tr>
<td>a</td>
<td>53.9</td>
</tr>
<tr>
<td>b</td>
<td>58.1</td>
</tr>
<tr>
<td>c</td>
<td>61.3</td>
</tr>
<tr>
<td>Completeness (%)</td>
<td>87.4 (93.1)</td>
</tr>
<tr>
<td>Total reflections</td>
<td>136929</td>
</tr>
<tr>
<td>Unique reflection</td>
<td>27770</td>
</tr>
<tr>
<td>Avg I/σ</td>
<td>13.5 (4.4)</td>
</tr>
<tr>
<td>Redundancy</td>
<td>4.9 (4.7)</td>
</tr>
<tr>
<td>Rsym (%)</td>
<td>4.8 (30.1)</td>
</tr>
<tr>
<td>RMSD</td>
<td></td>
</tr>
<tr>
<td>Bonds (Å)</td>
<td>0.009</td>
</tr>
<tr>
<td>Angles (°)</td>
<td>1.28</td>
</tr>
<tr>
<td>Rfree (%)</td>
<td>18.7</td>
</tr>
<tr>
<td>Rfree (%)</td>
<td>21.5</td>
</tr>
</tbody>
</table>

### Notes

- *Values in parentheses are for the highest-resolution shell.
- *Values in parentheses are for the highest-resolution shell.
tures when superpositions are performed. Double-difference plots between each cocomplex and the apoenzyme were therefore generated to determine the most invariant regions (Fig. 1). Product 3-4A varied most extensively from the apo state, which likely reflects differences in genotype, protein size, and crystal packing of the full-length construct. The remaining product complexes derive from the same NS3/4A protease domain construct and, in general, vary less extensively from the apoenzyme. Notably, the host cell product complexes are most similar to the apoenzyme, while viral product complexes 4A4B, 4B5A, and 5A5B vary more extensively. Taken together, these findings suggest that structural differences likely reflect the inherent flexibility in certain regions of the protease.

The set of double-difference plots was further analyzed, and the most invariant regions of the protease were determined to contain residues 32 to 36, 42 to 47, 50 to 54, 84 to 86, and 140 to 143 (Fig. 1). Product complexes 4A4B, 4B5A, 5A5B, TRIF, and MAVS consist of 10 protease molecules, which were subsequently superposed onto the apoenzyme using the Cα atoms of the structurally invariant residues. The Cα average root mean square deviation (RMSD) was calculated for each residue, and the seven most variable regions of the protease (Fig. 2A) were determined to be (i) the linker connecting cofactor 4A at the N terminus, (ii) the loop containing residues 65 to 70, (iii) the zinc-binding site containing residues 95 to 105, (iv) the 310 helix region spanning residues 128 to 136, (v) the zinc-binding site containing residues 145 to 148, (vi) the active-site antiparallel β sheet containing residues 156 to 168, and (vii) the C-terminal 3 helix. These regions are solvent exposed and likely influenced by both crystal packing effects and inherent flexibility.

Extensive structural differences are observed near the active site as indicated by large RMSDs for residues 156 to 168. These differences are most pronounced for the β strands E2 and F2, which form the antiparallel β sheet constituting the majority of the active site. This region is least variable near the catalytic triad, while the average RMSDs increase significantly toward the loop connecting these β strands (Fig. 2B). Though
the architecture of the protease catalytic triad is conserved, these observations suggest a potential dynamic interaction between the antiparallel β sheet of the protease active site and substrate products. Further studies are necessary to probe the nature and extent of such dynamic interplay. Nevertheless, though the C$_\alpha$ atoms of active-site residues shift relative to the protease core, these residues superpose well onto the apoenzyme with a RMSD range of 0.3 to 0.5Å. Moreover, the residue side chains adopt similar rotamer conformations and interact with the same surrounding residues, suggesting that potential flexibility in the protease active site would not disrupt its tertiary structure.

Analysis of viral product binding. Fluorescence polarization experiments reveal that the viral products bind with different affinities to the protease, with the $K_d$ for 4B5A over 10-fold weaker than those for 4A4B and 5A5B (Fig. 3). Crystal structures reveal that these viral products bind in a conserved manner, forming an antiparallel β sheet with protease residues 154 to 160 and burying 500 to 600 Å$^2$ of solvent-accessible surface area (19). The peptide product backbone torsion angles are very similar, with positions P1 to P4 being the most similar and residues P5 to P7 deviating progressively toward the C terminus. A constrained P2 $\phi$ torsion angle of about −60° is observed in product complexes 3-4A, 4A4B, and 5A5B. Interestingly, these P2 residues could sterically tolerate the substitution of proline, which is found at the P2 position in substrates 4B5A, TRIF, and MAVS. The ability of the P2 residue to adopt this constrained backbone torsion is a likely determinant in the recognition process, allowing for the proper positioning of the P1 cysteine for catalysis in the active site.

There are many conserved features in viral product binding, involving both backbone and side chain interactions (Fig. 4). For example, eight hydrogen bonds between backbone amide and carbonyl groups are completely conserved in the product complexes, involving protease residues G137, S138, S139A, R155, A157, and S159. S159 (C159 in product complex 3-4A) and A157 each contribute two hydrogen bonds with bound products at positions P5 and P3, respectively. The P1 residue, which is cysteine in all substrates but 3-4A, interacts favorably with the $\pi$ system of electrons of F154. All P1 terminal carboxyl groups sit in the oxyanion hole, hydrogen bonding with the Ne nitrogen of H57 and the amide nitrogens of residues 137 to 139. Though the coordinates of the P1 terminal oxygen atom are not included in the full-length NS3/4A structure, geometric restraints would position it similarly to the other peptide products. Thus, the same set of protease residues contacts all peptide sequences, although the precise nature of these interactions varies depending on the particular residue involved in each contact.

Despite these similarities, there are also unique features that likely underlie the particular specificity of NS3/4A for each substrate (Fig. 4). For example, the four acidic residues in...
product 4A4B lead to a highly charged peptide in solution. In the bound state, however, the atomic geometry suggests that the
P5 glutamate is protonated and hydrogen bonding with the car-
boxyl group of the P3 glutamate, which itself forms an ionic
interaction with the terminal nitrogen of K136. In fact, K136
interacts differently with all four viral products, forming (i) a
hydrogen bond with the P2 carbonyl oxygen of 3-4A, (ii) a salt
bridge with the P3 glutamate of 4A4B, and (iii) an extended
conformation that does not interact considerably in product com-
exes 4B5A and 5A5B. Fluorescence polarization data confirm a
more significant loss in binding affinity of product 4A4B for the
K136A protease variant than of products 4B5A and 5A5B (Fig.
3). Thus, the affinity of a substrate product likely arises from the
side chain interactions unique to that particular product.

FIG. 4. Stereo view of viral product binding to NS3/4A protease. N-terminal NS3/4A cleavage products 3-4A (A), 4A4B (B), 4B5A (C), and
5A5B (D) are shown binding to the protease active site, with the electrostatic interactions of backbone and side chain atoms depicted in black and
red, respectively.
solution studies recapitulate our structural observations and suggest that our crystal structure analyses are indeed representative of the interactions occurring in solution.

Analysis of host cell product complexes TRIF and MAVS. Active-site superpositions (residues 137 to 139 and 154 to 160) reveal that TRIF and MAVS peptide products both bind to the protease active site in a conserved three-dimensional shape (Fig. 6). Both substrate products form antiparallel β sheets with protease residues 154 to 160. There is no clear electron density in the TRIF complex for the proline residues spanning P13 to P7. Nevertheless, the residues from P6 to P1 overlap closely with the corresponding residues in the MAVS complex. The P1 cysteine residues of both substrate products interact with the aromatic ring of F154. Eight hydrogen bonds are observed in both structures, involving the amide nitrogens or carboxyl oxygens of residues G137, S138, S139A, R155, A157, and S159. A157 and S159 each contribute two hydrogen bonds with the P3 and P5 residues, respectively. In both structures, the carbonyl groups at position P1 interact with the protease oxyanion hole, defined by the backbone amide nitrogens of residues 137 to 139. Thus, the postcleavage products of the cellular substrates TRIF and MAVS bind to the protease active site in a conserved manner despite their large variations in primary sequence.

There are also many differences in the binding of TRIF and MAVS involving mainly side chain interactions with the protease (Fig. 6). For example, MAVS interacts closely with the protease electrostatic network formed by residues D81, R155, D168, and R123. The P4 glutamate of MAVS interacts with R155 and R123 in this network, while the P6 glutamate forms a salt bridge with R123. The TRIF peptide product, however, lacks such extended residues on this surface of the molecule, and the electrostatic network is notably absent, with the participating residues adopting conformations observed in the apoenzymatic structure. Previous studies demonstrate that a large fraction of full-length TRIF exists as polyproline II helices and that the interaction of NS3/4A with a polyproline II helix facilitates TRIF cleavage (11). The absence of clear electron density for the proline residues suggests that full-length TRIF may be necessary to stabilize the polyproline track in a conformation capable of specific interaction with NS3/4A.

Comparison of host cell and viral product binding. Host cell product binding was analyzed on a broader basis by comparison with the binding of viral substrates, previously reported by our group (29). Both viral (3-4A, 4A4B, 4B5A, and 5A5B) and host cell (TRIF and MAVS) substrate products bind to the protease active site in a conserved three-dimensional shape. The peptide backbone torsional angles are very similar, being most conserved at position P1 and deviating slightly toward position P6. All peptide products adopt constrained P2 ψ torsion angles, even those containing nonproline residues at this site. However, van der Waals analyses of substrate products indicate large variations in side chain interactions with the protease. All of the NS3/4A substrates contain either cysteine or threonine at position P1, while five of the six contain an acid at position P6. The P1 and P6 substrate residues each contribute the same amount of van der Waals energies in all product complexes (Fig. 7). The amino acid makeup of viral cleavage sequences is much more diverse at positions P5 to P2, and in general, the van der Waals energies at each position correlate with amino acid size. For example, the P4 glutamates in sub-
substrates 3-4A and MAVS, and to a lesser extent the P4 methionine in substrate 4A4B, are associated with larger van der Waals energies than the substrates with smaller P4 residues. Likewise, the larger glutamate residues at both P3 and P2 of substrate 4A4B also correlate with greater contact energies than for the other substrates, which contain smaller amino acids at these positions.

Though similar in shape, protease substrate binding can be further categorized into two groups: (i) product complexes 3-4A, 4A4B, 5A5B, and MAVS bind with an intact electrostatic network involving residues D81, R155, D168, and R123, while (ii) product complexes 4B5A and TRIF bind without this network such that R155, D168, and R123 maintain conformations observed in the apoenzyme. Notably, the 4B5A and TRIF cleavage sites contain fewer charged residues than the other substrates, which may underlie their inability to form the electrostatic network. Binding studies demonstrate that both 4B5A and TRIF have relatively weaker affinities for NS3/4A than the other substrates (11, 31). However, most biochemical studies have been conducted on small peptides corresponding to the immediate cleavage sequences of TRIF and 4B5A. Indeed,
kinetic studies revealed that full-length TRIF is processed more efficiently than peptides corresponding to the cleavage sequence (11). Additional molecular interactions by the full-length proteins or by adaptor proteins in the authentic cellular environment may better facilitate substrate binding. Nevertheless, the current structural analyses suggest that NS3/4A substrates vary in specificity by their ability to form and stabilize the protease electrostatic network; the sequence specificity is particularly influenced by amino acid variation at positions P6, P4, and P2.

Analysis using the viral substrate envelope. The binding of NS3/4A cellular substrates was analyzed in terms of the viral substrate envelope (Fig. 8), which was previously defined as the van der Waals volume shared by any three of the four viral products (29). This shape could not be predicted by the primary sequences alone, which highlights the conserved mode of viral substrate recognition despite their high sequence diversity. The backbone chains of both TRIF and MAVS fit entirely within the substrate envelope, as well as the side chains of TRIF spanning P5 to P1. The side chains of MAVS are also mostly confined within the substrate envelope, except for the longer side chains of the P6 glutamate, P5 arginine, and P4 glutamate. The carboxylic acids of these glutamate residues interact extensively with the protease electrostatic network, while the P5 arginine packs against loop residues 159 to 162. As these interactions occur outside the viral substrate envelope, we speculate that mutations that disrupt the electrostatic network, such as R155K and D168A, would preferentially reduce the proteolytic processing of MAVS compared to TRIF.

FIG. 8. Host cell product binding and the NS3/4A substrate envelope. (A) The NS3/4A substrate envelope is calculated from the consensus van der Waals volume shared by any three of the four viral products. (B) The TRIF cleavage product is largely confined within the substrate envelope. (C) MAVS is located mostly within the substrate envelope except for the side chain atoms spanning P6 to P4.

DISCUSSION

The recognition and proteolysis of the viral polyprotein and host cell adaptor proteins by NS3/4A protease play an integral role in the ability of HCV to replicate and evade the innate immune response to viral infection (12, 21). In this study, crystal structures of the NS3/4A protease domain revealed that viral and host cell products bind to the protease active site in
similar three-dimensional shapes, defined by the viral substrate envelope reported previously (29). The MAYS product complex reveals the formation of an extensive electrostatic network involving protease residues D81, R155, D168, and R123, which also form in viral product complexes 3-4A, 4A4B, and 5A5B. No such networks form in the TRIF and 4B5A complexes, and residues in this region of the protease adopt the same conformations observed in the apo state. The absence or presence of electrostatic networks also correlates with the affinities of product binding, with the $K_d$ of 4B5A being 10-fold weaker than those of products 4A4B and 5A5B. The greater catalytic efficiencies of NS3/4A for substrates 4A4B and 5A5B relative to 4B5A (31) may also derive from the formation of electrostatic networks. However, a short peptide may only partially mimic how the viral cleavage sequences are processed along the viral polyprotein in the natural cellular environment. Additional molecular features may further modulate the binding of TRIF and 4B5A, perhaps facilitated by the proline-rich regions contained in both proteins (11). Thus, the specificity of substrate processing by NS3/4A protease seems to arise from at least two distinct molecular interaction patterns, which likely influence the order and kinetics of polyprotein processing during the HCV life cycle.

In fact, these structural observations can be further linked to the known biology of NS3/4A processing during viral replication. Previous NS3-mediated cleavage assays of HCV polyprotein substrates revealed that NS4A is essential for the *trans* cleavage of junction 4B5A but is not required for the processing of junctions 4A4B and 5A5B (2, 22). Our structural analyses provide further insight into the molecular interactions underlying these previous findings. The NS4A cofactor likely stabilizes the tertiary protease fold required for the binding of NS3/4A substrates, and the binding of substrate 4B5A may absolutely depend on this particular protease conformation. Substrates 4A4B and 5A5B, however, may be able to induce these conformational changes through charge interactions, even in the absence of cofactor NS4A. Thus, our findings support the previous published data for NS3 HCV polyprotein processing, and future research is warranted to better ascertain the dynamic mechanisms of substrate recognition.

The ability of HCV to establish chronic human infections is highly dependent on the ability of the virus to effectively replicate while simultaneously evading the host cell immune response. The virally encoded NS3/4A protein plays an integral role in this process by mediating the cleavage of essential viral proteins and antiviral host cell adaptors. NS3/4A protease is thus a prime therapeutic target, and great efforts have been devoted to the development of protease inhibitors, which have demonstrated efficacy in late phases of human clinical trials. Nevertheless, the high rate and error-prone nature of HCV replication have led to the emergence of resistance against the most promising protease inhibitors to date, such as boceprevir, telaprevir, and ITMN-191 (16, 17, 20, 30, 32, 36). Inhibitor potency often derives from molecular interactions that are not essential for substrate recognition and cleavage. Mutations in these regions of the protease can selectively prevent drug binding while still allowing for the recognition and cleavage of viral and host cell substrates. Thus, identification of the protease residues that are important for substrate binding is crucial and will ultimately facilitate the design of drugs that target these particular residues. A more detailed understanding of the mechanisms underlying viral and host cell substrate recognition is therefore essential in facilitating a more rational approach to the design of more robust NS3/4A protease inhibitors.

**ACKNOWLEDGMENTS**

NMR chemical assignments were kindly provided by Herbert Klei of Bristol-Myers Squibb. We thank David Smith of the LS-CAT beamline at Argonne National Laboratory for data collection for the apo-enzyme; we also thank Shivender Sandilaya and Vukica Šrajer for data collection for the TRIF complex at BioCARDS, Madhavi Nalam and Rajintha Bandaranayake for assistance with structural refinement, and Seema Mittal, and Madhavi Kolli for their computational support.

Use of the Advanced Photon Source was supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences, under contract DE-AC02-06CH11357. Use of the BioCARDS Sector 14 was supported by the National Institutes of Health, National Center for Research Resources, under grant RR007707. Use of the LS-CAT Sector 21 was supported by the Michigan Economic Development Corporation and the Michigan Technology Tri-Corridor for the support of this research program (grant 08SP1000817). National Institutes of Health grants R01-AI085051 and 2R01-GM4347 supported this work.

**REFERENCES**