THE JOURNAL OF BIOLOGICAL CHEMISTRY Vol. 277, No. 52, Issue of December 27, pp. 50564–50572, 2002
Printed in U.S.A.

Structural Basis for the Substrate Specificity of Tobacco Etch Virus Protease*

Received for publication, July 18, 2002, and in revised form, September 19, 2002
Published, JBC Papers in Press, October 10, 2002, DOI 10.1074/jbc.M207224200

Jason Phan, Alexander Zdanov, Artem G. Evdokimov‡, Joseph E. Tropea, Howard K. Peters III, Rachel B. Kapust, Mi Li§, Alexander Wlodawer, and David S. Waugh¶

From the Macromolecular Crystallography Laboratory, Center for Cancer Research, NCI-Frederick, National Institutes of Health, Frederick, Maryland 21702-1201

Because of its stringent sequence specificity, the 3C-type protease from tobacco etch virus (TEV) is frequently used to remove affinity tags from recombinant proteins. It is unclear, however, exactly how TEV protease recognizes its substrates with such high selectivity. The crystal structures of two TEV protease mutants, inactive C151A and autolysis-resistant S219D, have now been solved at 2.2- and 1.8-Å resolution as complexes with a substrate and product peptide, respectively. The enzyme does not appear to have been perturbed by the mutations in either structure, and the modes of binding of the product and substrate are virtually identical. Analysis of the protein-ligand interactions helps to delineate the structural determinants of substrate specificity and provides guidance for reengineering the enzyme to further improve its utility for biotechnological applications.

The Picornaviridae are a large superfamily of (+)-strand RNA viruses that are responsible for a variety of plant and animal pathologies (1). Their RNA genomes are translated into polyprotein precursors that are co-translationally cleaved by viral proteases to generate the mature proteins (2). The majority of these processing events are mediated by the picornavirus 3C-type proteases, which are structurally similar to serine proteases like trypsin and chymotrypsin, but utilize a cysteine thiol instead of a serine hydroxyl as the active-site nucleophile (1, 3). Because they play an essential role in viral replication, 3C proteases are viewed as attractive molecular targets for antiviral therapeutics (4).

The stringent sequence specificity of rhinovirus 3C protease and the 3C-like nuclear inclusion protease encoded by TEV[1] has also led to their widespread application in the biotechnology sector as reagents for endoproteolytic removal of affinity tags from recombinant proteins (5). In contrast to Factor Xa, enterokinase, and thrombin, neither of these viral proteases has ever been reported to cleave genetically engineered fusion proteins at unintended locations. All 3C-type proteases exhibit a strong preference for glutamine in the P1 position of their substrates and for small aliphatic residues in the P1′ subsite, but these are clearly not the only specificity determinants (3, 6). Studies with oligopeptide substrates have established that the P6 and P3 subsites are also important specificity determinants for TEV protease (7), whereas it is the P4 and P2′ positions that appear to make the greatest contribution to the unique specificity of rhinovirus 3C protease (8).

Despite the fact that 3C-type proteases have been the subject of considerable interest, the structural basis of their substrate specificity remains obscure. Although the crystal structures of 3C proteases from hepatitis A virus (9), rhinovirus-14 (10), and poliovirus (11, 12) have been determined, none of them have cognate peptides in the active site. Consequently, efforts to explain the substrate specificity of these enzymes have relied on modeling or, in a few cases, on the structures of enzyme-inhibitor complexes (13, 14). Here, we show that the crystal structures of the catalytically inactive and catalytically active TEV proteases in complex with a peptide substrate and product, respectively, are extremely similar and that the mutation of the catalytic cysteine does not affect the conformation of the active site. The two structures reveal a wealth of information about how picornavirus 3C-type proteases selectively recognize their substrates.

EXPERIMENTAL PROCEDURES

Protein Expression and Purification—The His_7-TEV(S219D) protease expression vector (pRK529) is identical to pKM607 (15), except that the protease produced by pRK529 does not have a C-terminal polyarginine tag. The His_6-TEV(C151A) mutant was constructed by overlap extension PCR (16) using pRK508 (17) as the template. The His_6-TEV(C151A) protease was produced as an MBP fusion protein, which was subsequently cleaved by tobacco vein mottling virus protease at a designed site in the linker to remove the MBP domain. Both forms of TEV protease were produced in *Escherichia coli* BL21(DE3) cells that also contained an accessory plasmid encoding the *argU* and *ileX* tRNAs. A third plasmid (pRK1037) was used to produce the tobacco vein mottling virus protease for intracellular processing of the catalytically inactive MBP-His_6-TEV(C151A) fusion protein. The cells were grown at 37 °C in shake flasks containing Luria broth supplemented with the

* This work was supported in part by NCI Contract NO1-CO-56000 from the National Institutes of Health. The Industrial Macromolecular Crystallography Association Collaborative Access Team is supported by the companies of the Industrial Macromolecular Crystallography Association through a contract with the Illinois Institute of Technology, executed through the Illinois Institute of Technology Center for Synchrotron Radiation Research and Instrumentation. Use of the Advanced Photon Source was supported by the United States Department of Energy, Basic Energy Sciences, Office of Science, under Contract W-31-109-Eng-38. The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

The atomic coordinates and structure factors (code 1lvb and 1lvm) have been deposited in the Protein Data Bank, Research Collaboratory for Structural Bioinformatics, Rutgers University, New Brunswick, NJ (http://www.rcsb.org/).

‡ Present address: Procter & Gamble Pharmaceuticals, Health Care Research Center-Discovery, 8700 Mason-Montgomery Rd., Mason, OH 45040-9462.

§ Intramural Research Support Program, SAIC-Frederick, P. O. Box B, Frederick, MD 21702-1201.

¶ To whom correspondence should be addressed: Macromolecular Crystallography Lab., NCI-Frederick, NIH, P. O. Box B, Frederick, MD 21702-1201. Fax: 301-846-7148; E-mail: waughd@ncifcrf.gov.

[1] The abbreviations used are: TEV, tobacco etch virus; MBP, maltose-binding protein.

appropriate antibiotics until they reached mid-log phase, at which point isopropyl-β-D-thiogalactopyranoside was added to a final concentration of 1 mM, and the temperature was reduced to 30 °C for 4 h. Selenomethionine-substituted C151A protease was produced in the same way using the media formulation described by Doublié (18). The S219D protease was purified by immobilized metal ion affinity chromatography on Ni^{2+}-nitrilotriacetic acid resin, followed by cation exchange chromatography. The C151A protease was purified by immobilized metal ion affinity chromatography and gel filtration chromatography. An amylose column was used to absorb the residual undigested MBP-His_6-TEV(C151A) fusion protein prior to gel filtration.

Crystallization of the C151A Protease-Peptide Substrate Complex and Data Collection—The inactive enzyme-substrate complex was prepared by mixing the protein solution (13.5 mg/ml) with the peptide substrate in a 5-fold molar excess. Crystals were grown using the vapor diffusion method in a hanging drop setup at 18 °C (19) by mixing equal volumes of the enzyme-substrate complex with 8% (w/v) polyethylene glycol 6000 and 100 mM Tris-HCl (pH 8.0). Rectangular crystals appeared after 1 week and continued to grow to 0.35 × 0.35 × 0.20 mm within 3 weeks. The selenomethionine-substituted protein-substrate complex was prepared and co-crystallized similarly. The crystals belong to space group $P4_22_12$ with unit cell dimensions $a = b = 125.32$ Å and $c = 127.93$ Å. The asymmetric unit contains a dimer with the specific volume V_m = 4.35 Å^3/Da (20) and ~71% solvent.

All crystallographic diffraction experiments were carried out at the Industrial Macromolecular Crystallography Association insertion device on beamline 17ID (Advanced Photon Source, Argonne National Laboratory) with an Area Detector Systems Corp. Quantum 210 CCD detector. A single native or selenomethionine-substituted crystal was transferred to a cryoprotectant solution (20% (w/v) sucrose, 20% (w/v) glycerol, and 10% 3-(*N*-phenylmethyl-*N*,*N*-dimethylammonio)propanesulfonate) in mother liquor and flash-cooled at 100 K for data collection. Data for selenomethionine-substituted TEV protease were collected at three wavelengths near the selenium absorption edge: the high-energy remote (λ_1), the inflection point (λ_2), and the *f″* peak (λ_3). The exposure time was 10 s/image at an oscillation angle of 0.25°. All data were processed with the HKL2000 suite of programs (21).

Crystallization and Data Collection for the S219D Mutant Protein—Crystals of the TEV(S219D) mutant complexed with the peptide Ac-ENLYFQG were grown at pH 8.5 at a protein concentration of 4.5 mg/ml with a 5-fold molar excess of the peptide in 0.1 M Tris-HCl, 0.2 M $MgCl_2$, 10% glycerol, and 2.0 M ammonium sulfate as precipitant. These crystals belong to the tetragonal system, space group $P4_32_12$, with unit cell parameters $a = b = 75.50$ Å and $c = 183.17$ Å and with two molecules/asymmetric unit (V_m = 2.25 Å^3/Da and 44% solvent content). Diffraction data were collected at 100 K using a single crystal with dimensions of 0.6 × 0.6 × 0.1 mm on beamline X9B at the National Synchrotron Light Source (Brookhaven, NY). The data set consisted of 135 frames exposed for 15 s each with oscillations of 0.5°. The program suite HKL2000 (21) was used for processing of the diffraction intensities.

Structure Solution and Refinement of the C151A Mutant—The positions of 12 selenium atoms were determined, and the phases were calculated and refined by the automated crystallographic structure solution package SOLVE (22). The resulting electron density was improved by 2-fold non-crystallographic symmetry averaging and the maximum likelihood method of phase recombination using RESOLVE (23). The phases were extended to 2.2 Å, and a polyalanine model for the dimeric TEV protease complex was built automatically to 75% completeness. Sequence assignment and manual building of the structure were carried out with the graphics program O (24). The multi-wavelength anomalous diffraction-phased electron density map was readily interpretable and contiguous for residues 8–221 of the protease, as well as for the peptide residues TENLYFQSGT. The structure was subjected to a torsion angle simulated annealing protocol and positional and temperature factor refinements, with 2-fold non-crystallographic symmetry restraints against the data set collected at selenium λ_1 = 0.9755 Å using the program CNS (25). In addition to 182 water molecules, there are two glycerol molecules, derived from the cryoprotectant, located at the interface between the protomers. The first threonine of the bound peptide is missing from subunit B. The hexahistidine affinity tag and the 7 N-terminal residues of TEV protease are disordered, as are the 15 C-terminal residues. Because subsequent refinement with native data did not improve the model, only the structure of the selenomethionine-substituted protein was fully refined.

Structure Determination of the S219D Mutant Protein—The structure of the S219D mutant of TEV protease was solved by molecular replacement with the program package AMoRe (26) using the structure of the monomer of the inactive C151A mutant as the initial model. The solution of the molecular replacement was obtained with a correlation coefficient of 0.495 and an *R* factor of 0.443. Structure refinement was carried out with the program package CNS utilizing non-crystallographic symmetry restraints. The final model included a dimer of S219D molecules (molecules A and B), each complexed with a hexapeptide product in the active site, as well as 577 water molecules. Only the positions of residues 3–221 were traced in molecule B, whereas the electron density map for molecule A was sufficient to locate not just the N-terminal residues starting from position 1, but even an affinity tag (Gly-His_7). Thus, the refined model of molecule A contains residues −8 to 221. In addition, we located a heptapeptide in molecule A that could be unambiguously assigned as consisting of C-terminal amino acid residues 230–236.

RESULTS AND DISCUSSION

Crystallization and Structure Determination—Both forms of the TEV protease catalytic domain that are discussed here consist of amino acid residues 189–424 of the mature (49 kDa) NIa (nuclear inclusion a) protease (7). Hence, residue 1 in our numbering scheme corresponds to residue 189 of the mature NIa protein, and the catalytic triad residues His^{46}, Asp^{81}, and Cys^{151} (in our numbering scheme) correspond to residues 234, 269, and 339 in the full-length protein. The catalytically inactive C151A mutant has an N-terminal polyhistidine tag (Ser-His_6, residues −7 to −1). The catalytically active S219D mutant has a slightly different N-terminal tag (Gly-His_7, residues −8 to −1).

Crystals of catalytically active TEV protease were initially obtained in the presence of an oligopeptide substrate (Ac-ENLYFQG) using a mutant form of the enzyme (S219D) that is resistant to autolysis (15). A complete data set was collected at 1.8-Å resolution from a single crystal (Table I). However, these crystals were difficult to grow and proved to be impossible to reproduce, and the unsuccessful efforts to obtain heavy atom derivatives exhausted their supply. Subsequently, it was discovered that a catalytically inactive mutant of TEV protease (C151A) in which the active-site cysteine had been replaced by alanine could be crystallized in a reproducible fashion. The TEV(C151A) protease mutant was crystallized in the presence of a slightly longer oligopeptide substrate (TTENLYFQSGT), and the structure of the complex was determined by multi-wavelength anomalous diffraction methods using selenomethionine-substituted protein (Table I). The resulting model was then used to solve the structure of the S219D mutant by molecular replacement.

Description of the Overall Structure—As anticipated, TEV protease adopts the characteristic two-domain antiparallel β-barrel fold that is the hallmark of trypsin-like serine proteases, with the catalytic triad residues His^{46}, Asp^{81}, and Cys^{151} located at the interface between domains (Fig. 1*A*). In the inactive TEV(C151A) protease mutant, the oligopeptide substrate is bound in an extended conformation in the active site, with well defined electron density for all but the N-terminal Thr residue. We found that the co-crystallized oligopeptide substrate had been cleaved by the catalytically active S219D mutant and that the larger of the two products (Ac-ENLYFQ) was still bound in the enzyme active site. All 6 residues are visible in the electron density map, although the side chain of Tyr assumes different conformations in molecules A and B. It is not entirely clear why the product remained associated with the enzyme in the crystal. The conformations of the peptide substrate in the C151A structure and of the product in the S219D structure are very similar between the P6 and P1 positions.

Crystals of both the C151A and S219D proteases contain two molecules (molecules A and B) in each of their respective asymmetric units, and the interactions within these dimers are quite extensive (Fig. 1, *B* and *C*). However, the two dimers are not

TABLE I
Data collection, phasing, and refinement statistics
BNL, Brookhaven National Laboratory; APS, Advanced Photon Source; r.m.s.d., root mean square deviation.

	S219D	C151A		
Data collection				
X-ray source	BNL X9B	APS 17ID		
Wavelength	0.92	0.9755 (Seλ_1)	0.9794 (Seλ_2)	0.9796 (Seλ_3)
Space group	$P4_32_12$	$P4_22_12$		
Resolution (Å)	30.0 to 1.8	50.0 to 2.2	50.0 to 2.6	50.0 to 2.5
Unique reflections	49,988	51,139	31,988	35,972
Completeness (%)	99.8	97.8	99.9	99.9
Redundancy	6.2	7.6	8.0	7.9
R_{merge} (%)	5.8	8.0	7.5	8.3
Phasing				
Anomalous differences (%)		4.6	5.2	5.9
Correlation of anomalous differences/dispersive differences (%)				
Seλ_2-Seλ_1		0.84/3.7		
Seλ_3-Seλ_1		0.85/5.6		
Seλ_3-Seλ_2		0.89/2.8		
No. of sites		12		
Figure of merit (20 to 2.2)				
Centric		0.72		
Acentric		0.80		
Refinement				
No. of reflections				
Working set/test set	43,567/4,885	45,949/2,411		
R_{cryst} (%)	17.1	23.6		
R_{free} (%)	22.9	26.3		
r.m.s.d from ideal geometry				
Lengths (Å)	0.02	0.02		
Angles	1.9°	1.40°		
No. of molecules				
Peptide	2	2		
Water	577	182		
Glycerol		2		
Ramachandran analysis				
Most favored/allowed (%)	88.6/11.1	88.1/11.3		

the same, and neither of them is likely to have any biological relevance because dynamic light scattering and gel filtration experiments clearly indicate that TEV protease is monomeric in solution (data not shown). Interpretable electron density for residues 8–221 was observed for both molecules in the C151A crystal. Residues 3–221 of molecule B are visible in the S219D crystal, but there is clear electron density for residues 1–221, as well as for all 8 residues that compose the N-terminal His tag in molecule A (Fig. 1*A*). Although not unprecedented, complete His tags are very rarely observed in crystal structures. The His tag is wedged on the surface of the crevice between monomers A and B of the S219D dimer (Fig. 1*C*). It is curious that even though the conformation of N-terminal residues 3–11 is very similar in both molecules, only the His tag of molecule A is ordered in the electron density.

The backbone conformations of the two independent molecules in both structures are very similar, with an overall root mean square deviation of only 0.24 Å. The main differences occur in the conformation of the β-hairpin formed by residues 114–124. In the C151A crystal, the β-hairpins from neighboring molecules interact to form a four-stranded antiparallel β-sheet at the dimer interface (Fig. 1*B*). The same β-hairpins assume distinctly different conformations in the two protomers that compose the S219D dimer (Fig. 1*C*). The conformational differences in loop 114–124 can be attributed to the different crystal-packing interactions made by it in the two crystal forms (see below) and suggest some conformational flexibility in this region of the molecule.

In molecule A of the S219D crystal, residues 230–236 form a clearly visible β-strand that extends a four-stranded β-sheet by interacting with the strand formed by residues 62–67 (Fig. 1*C*). However, these C-terminal residues must have originated from a symmetry-related molecule B rather than from either molecule A or B in the same asymmetric unit because the distance would be too great; and thus, this intermolecular interaction must be a crystallographic artifact. We note also that although the N terminus of the enzyme is distant from its active site, the obvious flexibility of the C terminus and its proximity to the active site suggest that the cleavage events that release the protease from the polyprotein are almost certainly intermolecular (*trans*) at the N terminus, but could very well be intramolecular (*cis*) at the C terminus.

Autoproteolysis—The catalytic domain of TEV protease has a propensity to cleave itself *in vitro* between residues 218 and 219 to yield a truncated enzyme with greatly diminished activity (15, 27). It is uncertain, however, whether autoinactivation of TEV protease plays any role in the physiology of viral infection. The C-terminal residues that are removed by autolysis have no counterpart in other trypsin-like serine or cysteine proteases. Most of this region (residues 222–229) is disordered in both crystal forms of TEV protease, suggesting that it does not form an integral part of the folded catalytic domain and may be conformationally flexible.

Several lines of evidence indicate that autoproteolysis is an intramolecular reaction (15). In accord with this proposal, we found that TEV protease would not cleave an oligopeptide form of the internal cleavage site (GGHKVFMSKPRR), even though modeling suggests that all of the side chains can be accommodated in the corresponding subsites of the enzyme active site without any obvious steric clashes (data not shown). Intramolecular autoproteolysis certainly seems feasible from a structural standpoint because the scissile bond is positioned very close to the enzyme active site. This proximity effect, in concert with the conformational flexibility of the extended C terminus,

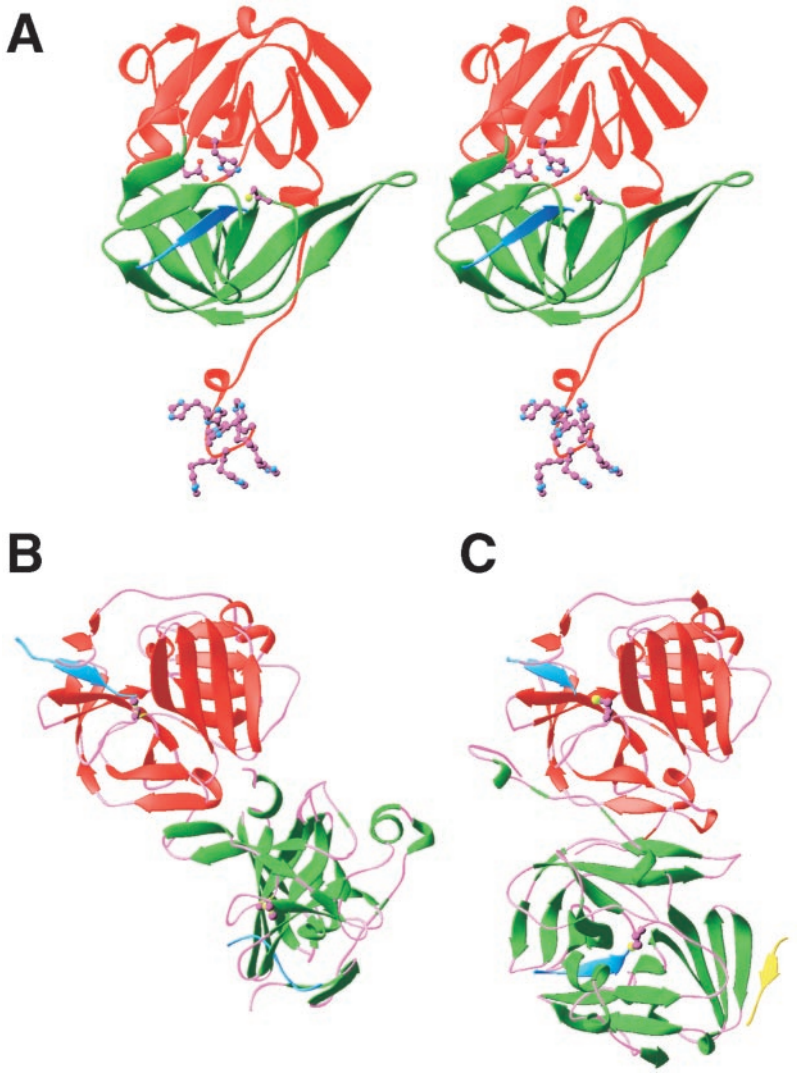

FIG. 1. **Ribbon models of the TEV(C151A) and TEV(S219D) protease structures.** *A*, stereo diagram of the S219D monomer (molecule A). The residues that compose the catalytic triad and the N-terminal His tag are depicted as ball-and-stick models (carbon, *violet*; nitrogen, *blue*; oxygen, *red*; and sulfur, *yellow*). The peptide product is also colored *blue* to distinguish it from the protein. *B*, the C151A dimer. *C*, the S219D dimer. Molecule B (*red*) is shown in the same orientation in *B* and *C* to illustrate the difference between the two dimers. The peptides are colored *blue*. Residue 151, which is Ala in the C151A protease and Cys in the S219D protease, is depicted as a ball-and-stick model. Residues 230–236, which are visible only in molecule A of the S219D protease, are colored *yellow*.

may explain why this site is readily cleaved by the protease even though the surrounding amino acid sequence (GHKVFM/S) bears little resemblance to the canonical recognition site (ENLYFQ/S). Intramolecular autoproteolysis would require a shift of the scissile bond between residues 218 and 219 by ~5 Å and its rotation into the active site, one wall of which would then be considerably modified because residues 217–221 contribute to the formation of the S3–S6 specificity pockets in TEV protease (see below). It is conceivable that the remodeled active site can better accommodate the non-canonical substrate sequence. The involvement of residues 217–221 in the structure of the active site also explains how the C-terminal truncation could increase the apparent K_m of the enzyme without affecting its k_{cat} (15).

The S219D mutation reduces the rate of autolysis by ~10-fold without affecting the catalytic activity of TEV protease, and other amino acid substitutions at this position (S219E, S219V, and S219P) give rise to enzymes with even greater resistance to autoproteolysis (15). In fact, the rank order of their stability is inversely correlated with the processing efficiency of peptide substrates containing the same amino acid substitutions in the P1′ position (28), indicating that the amino acid in the P1′ position of the internal cleavage site contributes to enzyme-substrate recognition in a manner that is consistent with the known specificity of TEV protease. It is understandable why none of these mutations impair the catalytic activity of the enzyme because the side chain of the residue in position 219 points away from the bound substrate (or product). On the other hand, a mutant protease with an amino acid substitution in the P2 position of the internal cleavage site (F217K) not only was more resistant to autoproteolysis, but also was a much less efficient catalyst due to a K_m effect (15). This, too, is understandable in light of the crystal structure because the side chain of Phe^{217} forms part of the S3 specificity pocket (see below).

Crystal Packing—As mentioned above, the unit cells of both the C151A and S219D mutants of TEV protease contain protein dimers. In the crystal structure of C151A, we found a rather unusual arrangement of intermolecular disulfide bonds formed between molecule A and its crystallographic symmetry mate A′ around the 4-fold screw axis, as well as between molecule B and its symmetry mate B′ around the 2-fold screw axis (Fig. 2). Cys^{130} is the only residue that is involved in the formation of these intermolecular disulfide bonds. TEV protease is a monomer in solution and was maintained in the presence of a reducing agent (dithiothreitol) throughout all steps of purification. Consequently, these disulfide bonds must have formed during the crystallization process. Although rare, similar phenomena have been reported previously, *e.g.* for RNA 3′-phosphate cyclase (29).

The dimers of the S219D mutant are much more closely packed than those of the C151A mutant, as reflected by the large differences between their V_m values of 2.25 and 4.35 (20). The surface area buried at the interface between molecules A

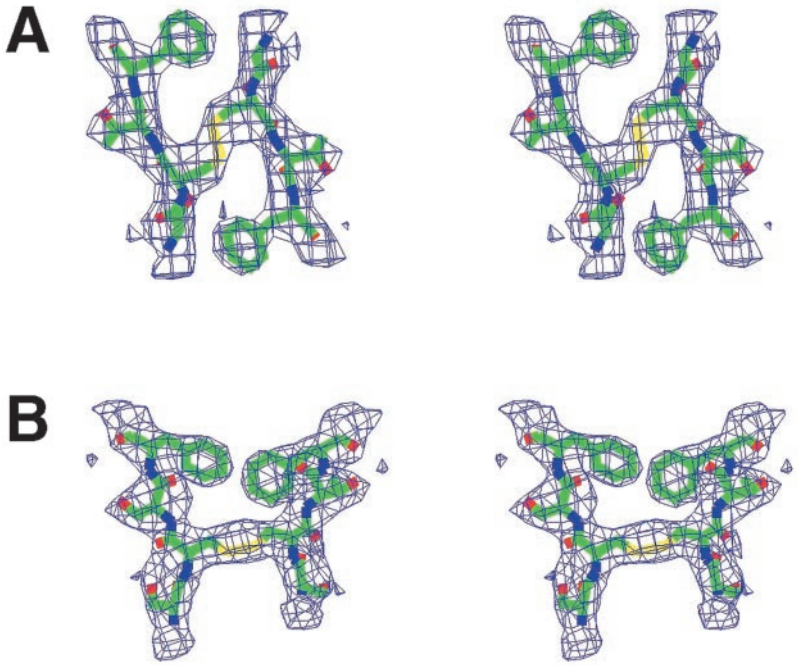

FIG. 2. **Stereo diagram of the intermolecular disulfide bonds in the C151A crystal lattice (prepared using CHAIN (37)).** Electron density is from the experimental multi-wavelength anomalous diffraction map contoured at 1.5 σ. *A*, the disulfide bond between molecules A and A′; *B*, the disulfide bond between molecules B and B′.

and B of the C151A dimer is 1534 Å^2, compared with 1209 and 491 Å^2 for the AA′ and BB′ disulfide-linked crystallographic dimers, respectively. The corresponding inaccessible surface area in S219D is twice as large (3138 Å^2).

Comparison with Related Proteins—A search of coordinates available from the Protein Data Bank performed with the program DALI (30) identified a number of proteins with structural similarity to TEV protease. Seven of these structures exhibit extremely high similarity, with *Z*-scores in excess of 12. However, all of these proteins have only limited sequence identity to TEV protease, ranging from 11 to 19% for the aligned segments that did not exceed two-thirds of the total length; thus, their full sequence identity is even lower (Fig. 3*A*). The two structures with the greatest similarity to TEV protease are those of the 3C proteases from hepatitis A virus (Protein Data Bank code 1hav) (12) and rhinovirus (code 1cqq) (31). Both are cysteine proteases with a trypsin-like fold that play a role similar to that of TEV protease in their respective viruses. Serine proteases showing significant similarity to TEV protease are β-trypsin (code 5ptp) (32), the protease domain of factor B (code 1dle) (33), *Staphylococcus aureus* epidermolytic toxin A (code 1agj) (34), and lysine-specific protease I from *Achromobacter lyticus* (code 1arb) (35). Finally, the human heparin-binding protein, also known as azurocidin (code 1a7S) (36), although missing the catalytic triad, also shows considerable structural conservation. These are representative structures for their respective families, and a number of other Protein Data Bank entries could also be used for comparisons.

Although the similarity between the overall fold of TEV protease and those of the related proteins listed above is very high (Fig. 3*B*), the actual atomic coordinates are not very close. The root mean square deviation for C-α atoms in the areas that could be aligned range between 2.4 and 3.5 Å, with the remaining loops showing little (if any) similarity. Even if the alignment is limited only to the central parts of the β-strands that define the conserved topology of the enzymes, the coordinates are >1.8 Å apart. This explains our inability to solve the structure of TEV protease by molecular replacement, although the models used for this purpose included most of the enzymes mentioned above.

The similarity of the catalytic sites (if present) of these enzymes is, of course, quite high. All three cysteine proteases are characterized by comparatively long distances between the S-γ atom of the catalytic cysteine and the N-ϵ2 atom of the histidine member of the triad, in the range of 4.0 Å. The distance between the carboxylate oxygen of the third member of the triad (Asp^{81} in TEV protease) and N-δ1 of His^{46} is 2.95–3.08 Å for the two molecules in the active enzyme, whereas the corresponding distance to the carboxylate of Glu^{71} in rhinovirus 3C protease is 2.8 Å. (Hepatitis A virus 3C protease is found in an inactive conformation, with the corresponding residue Asp^{84} turned away from the histidine.)

Structural Determinants of Substrate Specificity—As discussed above, the catalytically inactive (C151A) and catalytically active (S219D) TEV proteases were crystallized with similar but non-identical peptides bound in their respective active sites. The active site of the S219D mutant contained a product consisting of residues P6–P1 (ENLYFQ), whereas a longer peptide substrate (TTENLYFQSGT) was present in the active site of the C151A mutant. All of the residues in the substrate except for the N-terminal Thr had well defined electron density, thus delineating subsites P7–P3′ (Fig. 4). Biochemical studies have established that the specificity determinants for TEV protease reside between the P6 and P1′ positions of the substrate, with P6, P3, P1, and, to a lesser degree, P1′ being of greatest importance (7). Although the P7, P2′, and P3′ residues of the substrate are clearly visible in the electron density map, their side chains do not engage in any noteworthy interactions with the protease. Consequently, we will limit our discussion to the interactions involving residues P6–P1, which adopt very similar conformations in both crystal structures, and the P1′ Ser in the C151A structure. A comprehensive list of the interactions between the peptide substrate and the C151A protease is provided in Table II.

The peptide backbone of the substrate (or product) of TEV protease makes extended β-sheet interactions with the enzyme. These types of conformations have also been seen for the inhibitors complexed to the related enzymes. For example, human rhinovirus 3C protease was crystallized in the presence of a modified tetrapeptide inhibitor (AG7088), in which almost all of the side chains have been changed from their standard forms, that occupies the S4–S1 subsites in the enzyme (31).

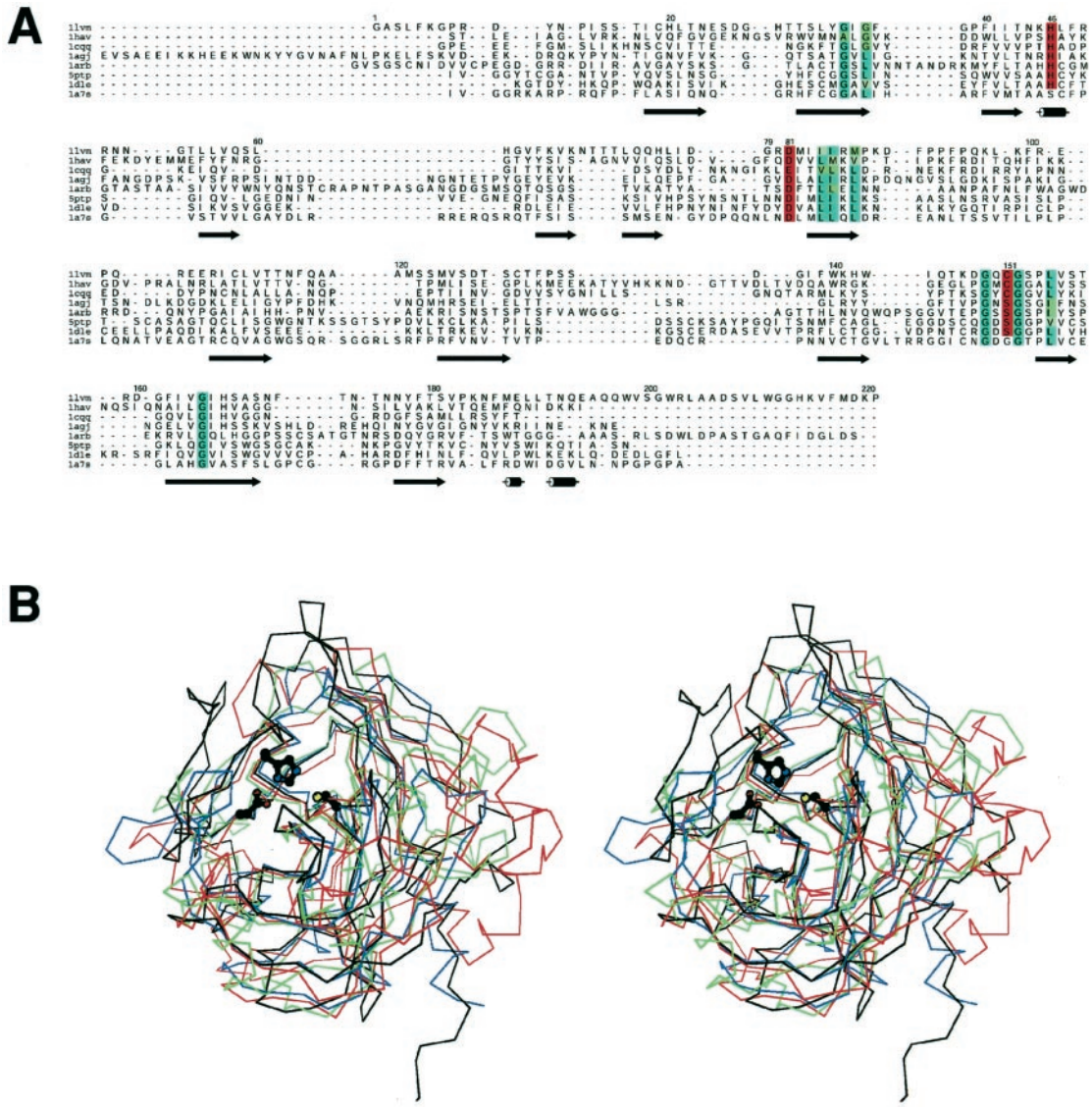

FIG. 3. **Comparison of TEV protease to other proteins with similar folds.** *A*, sequence alignment of TEV protease with seven other proteins that have similar folds (see "Results and Discussion"). The structures are identified by their Protein Data Bank codes. Those residues that are strictly conserved in more than half of the structures are colored *blue*, whereas those that are similar in the remaining ones are colored *green*. Secondary structure elements are marked for TEV protease. *B*, superposition of the backbone of TEV protease (*black*) on rhinovirus 3C protease (*blue*; code 1cqq) (31), β-trypsin (*green*; code 5ptp) (32), and lysine-specific protease I from *A. lyticus* (*red*; code 1arb) (35).

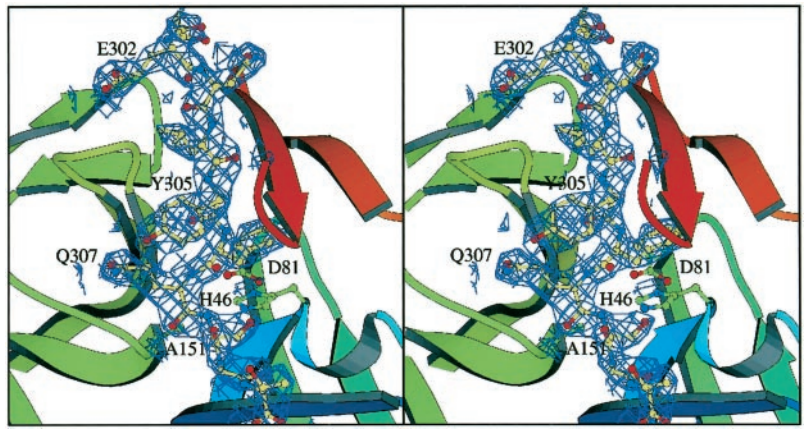

FIG. 4. **Stereo diagram of the peptide substrate bound to TEV(C151A) protease.** A ribbon model of the enzyme active-site cleft and a ball-and-stick model of the peptide substrate are overlaid on difference electron density contoured at 1.5 σ from an omit map. The carbon atoms in the substrate and the catalytic triad residues in the protease are colored *yellow* and *green*, respectively.

Because rhinovirus 3C and TEV proteases share some common specificity determinants (*e.g.* Gln in the P1 position), a comparison of the two co-crystal structures is instructive. The peptide backbone is retained in AG7088 and is also involved in forming β-sheet interactions with the enzyme. In both structures, many interactions are formed by the side chains that occupy the S1–S4 substrate-binding pockets.

The P1 site is occupied by Gln in the TEV protease substrate,

TABLE II
Interactions between the TEV(C151A) protease and the peptide substrate

Amino acids		Hydrogen bonds	Hydrophobic interactions
Enzyme	Substrate		
		Å	Å
His^{46} C-δ2	Ser^{308} C-β		3.79
Ser^{170} C-β	Gln^{307} C-δ		3.88
Asp^{148} O-δ1	Gln^{307} N-ε2	2.98	
Thr^{146} O	Gln^{307} N-ε2	3.04	
His^{167} N-ε2	Gln^{307} O-ε1	2.65	
Thr^{146} O-γ1	Gln^{307} O-ε1	2.55	
Val^{209} C-γ1	Phe^{306} C-ε1		3.89
Trp^{211} C-ζ3	Phe^{306} C-ζ		3.62
Val^{216} C-γ1	Phe^{306} C-ε1		3.48
Met^{218} C-ε	Phe^{306} C-ε1		3.63
His^{46} C-ε1	Phe^{306} C-δ2		3.44
Asn^{174} N-δ2	Tyr^{305} O-η	2.95	
Asp^{148} O-δ1	Tyr^{305} O-η	2.56	
Lys^{220} C-ε	Tyr^{305} C-ζ		3.46
Val^{216} C-γ1	Tyr^{305} C-α		3.81
His^{214} C-γ	Leu^{304} C-δ1		3.63
Tyr^{178} C-ε1	Leu^{304} C-δ2		3.43
Asn^{171} C-α	Leu^{304} C-δ2		3.76
Ala^{169} C-β	Leu^{304} C-β		4.15
Wat51 O	Glu^{302} O-ε2	2.66	
Tyr^{178} O-η	Glu^{302} O-ε1	2.59	
Asn^{176} N-δ2	Glu^{302} O-ε1	2.82	
Asn^{171} N-δ2	Glu^{302} O-ε2	2.95	
Asn^{171} C-β	Glu^{302} C-δ		3.40
Gly^{213} O	Thr^{301} O-γ1	2.55	
His^{214} C-ε1	Thr^{301} C-β		3.44

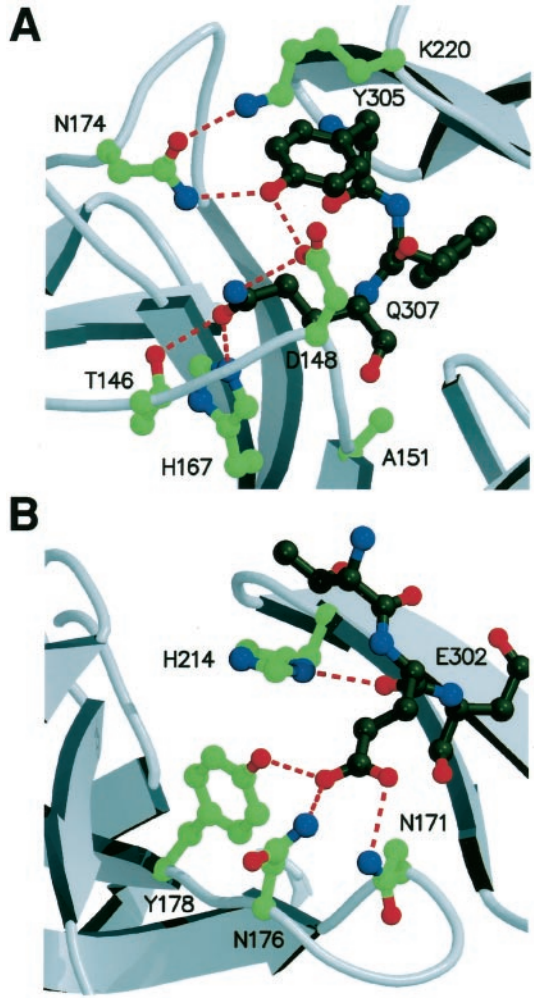

FIG. 5. **Networks of hydrogen bonds in the S1 (*A*) and S6 (*B*) specificity pockets of TEV protease.** The carbon and nitrogen atoms are colored *dark green* and *blue* in the peptide and *green* and *blue* in the protease, respectively. The ribbons are colored *gray*, and the *dashed red lines* represent hydrogen bonds (prepared using Molscript (38) and Raster3d (39)).

whereas the equivalent position in AG7088 contains a five-member lactam. Both N-ϵ2 and O-ϵ1 of P1 Gln are involved in forming hydrogen bonds with the side chains of His^{167} and Thr^{146} in TEV protease (Fig. 5*A*), whereas very similar interactions in 3C protease are made by the nitrogen and oxygen atoms of the lactam with the side chains of the equivalent residues His^{161} and Thr^{142}. The character of the hydrophobic interactions is also conserved. Thus, the absolute requirement for glutamine in the P1 position by TEV and human rhinovirus 3C proteases can be explained.

The S2 pocket in TEV protease is lined with exclusively hydrophobic residues (Val^{209}, Trp^{211}, Val^{216}, and Met^{218}), as well as with a face of His^{46}. The pocket is closed and protected from solvent. By comparison, because there is no strand corresponding to the C-terminal part of TEV protease in 3C protease, its S2 pocket is completely open on the side and therefore not very hydrophobic.

The S3 pocket in TEV protease is occupied by a tyrosine that assumes two alternative conformations in different monomers. In the first conformation, the side chain is oriented such that its hydroxyl makes a short hydrogen bond to the N-δ2 atom of Asn^{174} and also with O-δ1 of Asp^{148} (Fig. 5*A*). Hydrophobic interactions are mediated primarily by the side chains of Phe^{172} and Lys^{220}, as well as by the main chain atoms of residues 171–172. In the second conformation, the Tyr side chain rotates 118° around the C-α–C-β bond into a position previously occupied by Phe^{172}, which, in turn, moves away toward Pro^{221}. In its new position, the side chain of Tyr^{305} is located in the middle of the hydrophobic pocket made by Phe^{172}, Phe^{217}, the hydrophobic part of Lys^{220}, and Pro^{221}. By contrast, P3 Val of AG7088 points into an open area and is not involved in any interactions with rhinovirus 3C protease.

The S4 pocket of TEV protease is occupied by a leucine side chain that engages in hydrophobic interactions with the side chains of Phe^{139}, Ala^{169}, His^{214}, Tyr^{178}, and Val^{216}. In rhinovirus 3C protease, the isoxazole S4 group of AG7088 is involved in several hydrogen bonds provided by Asn^{165} and the main chain carbonyl of the preceding residue. Thus, the character of the S4 subsites in these two viral proteases is considerably different. Even more extensive differences may be expected for hepatitis A virus 3C protease (12) because its residues 143–156 occupy some of the substrate-binding areas of TEV and rhinovirus 3C proteases, as well as the C-terminal strand of TEV protease, suggesting either that the substrate binds in a significantly different way or that substantial reorganization of the structure of the enzyme must accompany substrate binding.

There is no S5 pocket in TEV protease. Rather, the side chain in this position points away from the protein. This agrees with the observation that practically any residue can occupy the P5 position with little or no impact on the efficiency of processing (7).

TEV protease exhibits a stringent requirement for Glu in the P6 position of its substrates (7). This residue is involved in an intricate network of hydrogen-bonding interactions in both crystal structures (Fig. 5*B*). Its O-ϵ2 atom is within hydrogen-bonding distance of N-δ2 of Asn^{171}, and O-ϵ1 of the latter residue accepts a hydrogen bond from the main chain amide of Asn^{176}. The O-ϵ1 atom of P6 Glu accepts a hydrogen bond from the N-δ2 atom of Asn^{176}. O-δ1 of the latter residue accepts a

hydrogen bond from N-ϵ1 of Trp^{143}. In addition, there is a hydrogen bond between Glu O-ϵ1 and the hydroxyl of Tyr^{178}. All of these hydrogen bonds can be formed only if the P6 side chain is Glu because any other residue would interrupt this cooperative network. For example, a Gln in the P6 position would place two nitrogens in close proximity to one another. The remaining hydrogen bonds would prevent the P6 side chain from rotating 180° to alleviate this unfavorable interaction. Thus, electrostatic repulsion between two side chain amide nitrogens in the case of P6 Gln and the advantage of being able to make a good hydrogen bond in the case of P6 Glu probably explain the strong preference exhibited by TEV protease for Glu in the P6 position.

The S1′ pocket of TEV protease is a shallow, narrow groove on its surface. Consequently, the side chain of the P1′ residue is partially exposed to solvent rather than completely buried within the complex. Experiments using genetically engineered fusion proteins and peptides with different residues in the P1′ position of an otherwise canonical TEV protease recognition site demonstrated that the enzyme can tolerate a wide variety of side chains in the S1′ subsite (28). The most efficient substrates were those with short, aliphatic side chains (Gly, Ser, Ala, Met, and Cys). These residues would readily fit into the S1′ groove. Longer side chains could also be accommodated in the groove, but this would bring them into close physical proximity with one edge of the His^{46} imidazole ring, which is part of the catalytic triad. The tolerance of TEV protease for longer side chains in the P1′ position could also be explained if they can rotate away from the protein and project into the solvent. The least efficient substrates were those with Pro or the β-branched hydrophobic residues Leu, Ile, and Val in the P1′ position. The bulky substituents of the β-branched side chains would sterically clash with the narrow and shallow dimensions of the S1′ subsite, and the conformational constraints on the polypeptide backbone imposed by Pro would create a similar problem.

Prospects for Creating Mutant Proteases with Altered Specificity—Having elucidated the structural basis for the substrate specificity of TEV protease, the prospects for designing mutant proteases with altered specificity can now be considered. The two co-crystal structures reveal that the P6, P4, P3, P2, P1, and P1′ substituents of the substrate make direct contacts with the enzyme active site (Table II). Although they are among the most important specificity determinants, the S3 and S1 subsites would present a substantial challenge from an engineering standpoint because they are intimately interconnected. In addition, as discussed above, the presence of the catalytic triad imidazole ring in the bottom of the S1′ pocket would complicate any effort to alter specificity at this position. The remaining pockets (S6, S4, and S2) therefore appear to be the most promising targets for specificity engineering. Replacing Asn^{171} with Asp might create a more favorable environment for Gln than Glu in the S6 subsite of TEV protease. The S4 pocket has the potential to be enlarged by replacing Tyr^{178} with a smaller residue (*e.g.* Val), which might enable it to accommodate bulkier residues such as Phe and Tyr. It may also be possible to alter the specificity of the S2 pocket by replacing Val^{209} with Ser because the Ser hydroxyl would then be in a perfect position to form a hydrogen bond with the hydroxyl group of a Tyr in the P2 position. Of course, additional possibilities for the rational design of mutant proteases with altered specificity also exist. Another attractive approach would be to employ localized random mutagenesis in conjunction with a genetic selection or screen for TEV protease activity to isolate mutants with the desired phenotypes.

Conclusion—The co-crystal structures of TEV protease in complex with a peptide substrate on the one hand and a peptide product on the other have illuminated, for the first time, the structural basis of its stringent sequence specificity and helped to explain why the protease readily cleaves itself at a specific site near the C terminus even though the surrounding sequence does not closely resemble the canonical recognition site. Moreover, the two structures lay the groundwork for the rational design of TEV protease mutants with altered substrate specificity. Mutants with alternative specificities would be useful reagents for cleaving fusion proteins in cases where the target protein happens to contain a sequence that closely resembles a canonical TEV protease recognition site. It is also possible that the catalytic activity and stability of the protease can be improved by structure-based protein engineering.

Acknowledgments—We thank Holly Baden, Fan Yang, and D. Eric Anderson for early contributions to this work; Suzanne Specht and Terry Copeland for peptide synthesis; Scott Cherry for assistance with the preparation of selenomethionine-labeled TEV protease; and Zbigniew Dauter for the opportunity to collect diffraction data on beamline X9B at the National Synchrotron Light Source. Additional diffraction data were collected on beamline 17ID in the facilities of the Industrial Macromolecular Crystallography Association Collaborative Access Team at the Advanced Photon Source.

REFERENCES

1. Ryan, M. D., and Flint, M. (1997) *J. Gen. Virol.* **78,** 699–723
2. Stanway, G. (1990) *J. Gen. Virol.* **71,** 2483–2501
3. Seipelt, J., Guarne, A., Bergmann, E., James, M., Sommergruber, W., Fita, I., and Skern, T. (1999) *Virus Res.* **62,** 159–168
4. Wang, Q. M. (2001) *Prog. Drug Res. Spec. No.*, 229–253
5. Stevens, R. C. (2000) *Structure* **8,** R177–R185
6. Malcolm, B. A. (1995) *Protein Sci.* **4,** 1439–1445
7. Dougherty, W. G., Cary, S. M., and Parks, T. D. (1989) *Virology* **171,** 356–364
8. Cordingley, M. G., Callahan, P. L., Sardana, V. V., Garsky, V. M., and Colonno, R. J. (1990) *J. Biol. Chem.* **265,** 9062–9065
9. Allaire, M., Chernaia, M. M., Malcolm, B. A., and James, M. N. (1994) *Nature* **369,** 72–76
10. Matthews, D. A., Smith, W. W., Ferre, R. A., Condon, B., Budahazi, G., Sisson, W., Villafranca, J. E., Janson, C. A., McElroy, H. E., and Gribskov, C. L. (1994) *Cell* **77,** 761–771
11. Mosimann, S. C., Cherney, M. M., Sia, S., Plotch, S., and James, M. N. (1997) *J. Mol. Biol.* **273,** 1032–1047
12. Bergmann, E. M., Mosimann, S. C., Chernaia, M. M., Malcolm, B. A., and James, M. N. (1997) *J. Virol.* **71,** 2436–2448
13. Bergmann, E. M., Cherney, M. M., Mckendrick, J., Frormann, S., Luo, C., Malcolm, B. A., Vederas, J. C., and James, M. N. (1999) *Virology* **265,** 153–163
14. Dragovich, P. S., Webber, S. E., Babine, R. E., Fuhrman, S. A., Patick, A. K., Matthews, D. A., Reich, S. H., Marakovits, J. T., Prins, T. J., Zhou, R., Tikhe, J., Littlefield, E. S., Bleckman, T. M., Wallace, M. B., Little, T. L., Ford, C. E., Meador, J. W., III, Ferre, R. A., Brown, E. L., Binford, S. L., DeLisle, D. M., and Worland, S. T. (1998) *J. Med. Chem.* **41,** 2819–2834
15. Kapust, R. B., Tözsér, J., Fox, J. D., Anderson, D. E., Cherry, S., Copeland, T. D., and Waugh, D. S. (2001) *Protein Eng.* **14,** 993–1000
16. Ho, S. N., Hunt, H. D., Horton, R. M., Pullen, J. K., and Pease, L. R. (1989) *Gene* (*Amst.*) **77,** 51–59
17. Kapust, R. B., and Waugh, D. S. (1999) *Protein Sci.* **8,** 1668–1674
18. Doublié, S. (1997) *Methods Enzymol.* **276,** 523–530
19. Wlodawer, A., and Hodgson, K. O. (1975) *Proc. Natl. Acad. Sci. U. S. A.* **72,** 398–399
20. Matthews, B. W. (1968) *J. Mol. Biol.* **33,** 491–497
21. Otwinowski, Z., and Minor, W. (1997) *Methods Enzymol.* **276,** 307–326
22. Terwilliger, T. C., and Berendzen, J. (1999) *Acta Crystallogr. Sect. D Biol. Crystallogr.* **55,** 849–861
23. Terwilliger, T. C., and Berendzen, J. (1997) *Acta Cryst. Acta Crystallogr. Sect. D Biol. Crystallogr.* **53,** 571–579
24. Kleywegt, G. J., and Jones, T. A. (1997) *Methods Enzymol.* **277,** 208–230
25. Brünger, 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., Read, R. J., Rice, L. M., Simonson, T., and Warren, G. L. (1998) *Acta Crystallogr. Sect. D Biol. Crystallogr.* **54,** 905–921
26. Navaza, J. (1994) *Acta Crystallogr. Sect. A* **50,** 157–163
27. Parks, T. D., Howard, E. D., Wolpert, T. J., Arp, D. J., and Dougherty, W. G. (1995) *Virology* **210,** 194–201
28. Kapust, R. B., Tözsér, J., Copeland, T. D., and Waugh, D. (2002) *Biochem. Biophys. Res. Commun.* **294,** 949–955
29. Palm, G. J., Billy, E., Filipowicz, W., and Wlodawer, A. (2000) *Structure* **8,** 13–23
30. Holm, L., and Sander, C. (1993) *J. Mol. Biol.* **233,** 123–138
31. Matthews, D. A., Dragovich, P. S., Webber, S. E., Fuhrman, S. A., Patick, A. K., Zalman, L. S., Hendrickson, T. F., Love, R. A., Prins, T. J., Marakovits, J. T., Zhou, R., Tikhe, J., Ford, C. E., Meador, J. W., Ferre, R. A., Brown, E. L., Binford, S. L., Brothers, M. A., DeLisle, D. M., and Worland, S. T. (1999) *Proc. Natl. Acad. Sci. U. S. A.* **96,** 11000–11007
32. Finer-Moore, J. S., Kossiakoff, A. A., Hurley, J. H., Earnest, T., and Stroud,

R. M. (1992) *Proteins* **12,** 203–222
33. Jing, H., Xu, Y., Carson, M., Moore, D., Macon, K. J., Volanakis, J. E., and Narayana, S. V. (2000) *EMBO J.* **19,** 164–173
34. Cavarelli, J., Prevost, G., Bourguet, W., Moulinier, L., Chevrier, B., Delagoutte, B., Bilwes, A., Mourey, L., Rifai, S., Piemont, Y., and Moras, D. (1997) *Structure* **5,** 813–824
35. Tsunasawa, S., Masaki, T., Hirose, M., Soejima, M., and Sakiyama, F. (1989) *J. Biol. Chem.* **264,** 3832–3839
36. Karlsen, S., Iversen, L. F., Larsen, I. K., Flodgaard, H. J., and Kastrup, J. S. (1998) *Acta Crystallogr. Sect. D Biol. Crystallogr.* **54,** 598–609
37. Sack, J. S. (1988) *J. Mol. Graph.* **6,** 224–225
38. Kraulis, P. J. (1991) *J. Appl. Crystallogr.* **24,** 946–950
39. Meritt, E. A., and Murphy, M. E. P. (1994) *Acta Crystallogr. Sect. D Biol. Crystallogr.* **50,** 869–873