Structure of coronavirus main proteinase reveals combination of a chymotrypsin fold with an extra α-helical domain

Kanchan Anand, Gottfried J.Palm, Jeroen R.Mesters, Stuart G.Siddell1, John Ziebuhr1,2 and Rolf Hilgenfeld2

Department of Structural Biology and Crystallography, Institute of Molecular Biotechnology, D-07745 Jena and 1Institute of Virology and Immunology, University of Würzburg, D-97078 Würzburg, Germany

Corresponding authors
E-mail: hilgenfel@imb-jena.de or ziebuhr@vim.uni-wuerzburg.de

The key enzyme in coronavirus polyprotein processing is the viral main proteinase, Mpro, a protein with extremely low sequence similarity to other viral and cellular proteinases. Here, the crystal structure of the 33.1 kDa transmissible gastroenteritis (corona)virus Mpro is reported. The structure was refined to 1.96 Å resolution and revealed three dimers in the asymmetric unit. The mutual arrangement of the prote- mers in each of the dimers suggests that Mpro self-processing occurs in trans. The active site, comprised of Cys144 and His41, is part of a chymotrypsin- like fold that is connected by a 16 residue loop to an extra domain featuring a novel α-helical fold. Molecular modelling and mutagenesis data implicate the loop in substrate binding and elucidate S1 and S2 subsites suitable to accommodate the side chains of the P1 glutamate and P2 leucine residues of Mpro substrates. Interactions involving the N-terminus and the α-helical domain stabilize the loop in the orientation required for trans-cleavage activity. The study illustrates that RNA viruses have evolved unprecedented variations of the classical chymotrypsin fold.

Keywords: 3C-like/catalytic dyad/coronavirus/proteinase/ X-ray crystallography

Introduction

Transmissible gastroenteritis virus (TGEV) belongs to the Coronaviridae, a family of positive-strand RNA viruses. Coronaviruses have the largest RNA viral genomes known to date (28 500 nucleotides in the case of TGEV) and share a similar genome organization and common transcriptional and translational strategies with the Arteriviridae (den Boon et al., 1991; Cavanagh, 1997). TGEV infection is associated with severe and often fatal diarrhoea in young pigs (for reviews see Enjuanes and van der Zeijst, 1995; Saif and Wesley, 1999).

The viral proteins required for TGEV genome replication and transcription are encoded by the replicase gene (Eleouet et al., 1995; Penzes et al., 2001). This gene encodes two replicative polyproteins, pp1a (447 kDa) and pp1ab (754 kDa) that are processed by virus-encoded proteinases to produce the functional subunits of the replication complex (reviewed in Ziebuhr et al., 2000). The central and C-proximal regions of pp1a and pp1ab are processed by a 33.1 kDa viral cysteine proteinase which is called the ‘main proteinase’ (Mpro) or, alternatively, the ‘3C-like proteinase’ (3CLpro). The name ‘3C-like proteinase’ was introduced originally because of similar substrate specificities of the coronavirus Mpro and picornavirus 3C proteinases (3CLpro) and the identification of cysteine as the principal catalytic residue in the context of a predicted two-β-barrel fold (Gorbalenya et al., 1989a,b). Meanwhile, however, several studies have revealed significant differences in both the active sites and domain structures between the coronavirus and picornavirus enzymes (Liu and Brown, 1995; Lu and Denison, 1997; Ziebuhr et al., 1997, 2000; Hegyi et al., 2002). Also, the crystal structures reported for a number of picornavirus 3C proteinases (Allaire et al., 1994; Matthews et al., 1994; Bergmann et al., 1997; Mosimann et al., 1997) have not been useful in predicting the three-dimensional structures of coronavirus main proteinases. Because of the large phylogenetic distance between the two groups of enzymes, we will use the term coronavirus Mpro throughout this article.

Sequence comparisons (Figure 1) and experimental data obtained for other coronavirus homologues allow us to predict that the mature form of the TGEV Mpro is released from pp1a and pp1ab by autoproteolytic cleavage at flanking Gln-Ser sites (Eleouet et al., 1995; Hegyi and Ziebuhr, 2002). Accordingly, the TGEV Mpro has 302 amino acid residues that correspond to the pp1a/pp1ab residues 2879–3180. In vivo and in vitro analyses of avian infectious bronchitis virus (IBV), mouse hepatitis virus (MHV) and human coronavirus 229E (HCoV 229E) Mpro activities have shown consistently that the proteinase cleaves the replicase polyproteins at 11 conserved sites and, therefore, it seems reasonable to conclude that the Mpro-mediated processing pathways are conserved in all coronaviruses, including TGEV.

Previous theoretical studies and experimental data have led to the following conclusions (Bazan and Fletterick, 1988; Gorbalenya et al., 1989a,b; Liu and Brown, 1995; Lu et al., 1995; Ziebuhr et al., 1995, 1997, 2000; Lu and Denison, 1997; Seybert et al., 1997; Ziebuhr and Siddell, 1999; Ng and Liu, 2000; Hegyi et al., 2002): (i) Coronavirus main proteinases employ conserved cysteine and histidine residues in the catalytic site. In TGEV Mpro, these are Cys144 and His41. There has been some debate on the existence of a third residue in the catalytic centre. In common with picornavirus 3C proteinases, the catalytic centre of the coronavirus Mpro is predicted to be embedded in a chymotrypsin-like, two-β-barrel structure in which cysteine (rather than serine) serves as the principal nucleophile. (ii) Coronavirus main proteinases have well-defined substrate specificities. All known cleavage sites contain bulky hydrophobic residues (mainly leucine) at the P2 position, glutamine at the P1 position, and small...
aliphatic residues at the P1 position. (iii) Coronavirus main proteinases possess a large C-terminal domain of ~110 amino acid residues that is not found in other RNA virus 3C-like proteinases. The characterization of recombinant proteins, in which 33, 28 and 34 C-terminal amino acid residues were deleted from the IBV, MHV and HCoV main proteinases, respectively, resulted consistently in dramatic losses of proteolytic activity, suggesting that the C-terminal domain of Mpro contributes to proteolytic activity through undefined mechanisms.

The 1.96 Å TGEV Mpro crystal structure reported herein reveals the structural details of a unique catalytic system and facilitates the interpretation of previously published mutagenesis studies that have, at least in part, remained speculative due to the complete lack of structural information on ‘3C-like’ enzymes.

Results and discussion

Structure determination by MAD phasing

The presence of 10 methionine residues in the TGEV Mpro molecule suggested that selenomethionine-based multi-wavelength anomalous dispersion (MAD; Hendrickson et al, 1990) could be used to solve the phase problem. The unit cell dimensions of the crystals (a = 72.8 Å, b = 160.1 Å, c = 88.9 Å, β = 94.3°, space group P2₁) and self-rotation calculations indicated the presence of as many as six TGEV Mpro molecules per asymmetric unit. In the MAD phasing process, we finally succeeded in locating 48 (out of 60) crystallographically independent selenium sites by the ‘Shake & Bake’ approach to direct methods (Weeks and Miller, 1999), without recourse to heavy atom derivatives or other methods of phasing (see Materials and methods). The phases obtained resulted in a readily interpretable electron density map.

Quality of the model

All six copies (designated A–F) of the TGEV Mpro in the asymmetric unit of the crystal could be built into well-defined electron density (Figure 2), which covered almost all of the 302 amino acid residues of each monomer. The only exceptions were the two C-terminal residues which were not visible in five of the six chains. Monomers A, E and F also lacked electron density for residue 300.

Fig. 1. Sequence comparison of coronavirus main proteinases. The alignment was produced using CLUSTAL X, version 1.81 (Thompson et al., 1997), and corrected manually on the basis of the three-dimensional structure of TGEV Mpro. The corresponding sequences of FIPV (strain 79–1146), HCoV (strain 229E), bovine coronavirus (BCoV, isolate LUN), MHV (strain JHM) and IBV (strain Beaudette) were derived from the replicative polyproteins of the respective viruses whose sequences are deposited at the DDBJ/EMBL/GenBank database (accession Nos: FIPV, AF326575; HCoV, X69721; BCoV, AF391542; MHV, M55148; IBV, M95169; TGEV, AJ271965). The β-strands and α-helices as revealed in the TGEV Mpro crystal structure (this study) are shown above the sequence alignment (see also Figures 4 and 5). Black background colour indicates the catalytic cysteine and histidine residues. Grey background colour indicates the key residue of the S1 subsite (TGEV Mpro His162) and its equivalents in other coronavirus main proteinases. Also shown in grey are the phenylalanine and tyrosine residues (TGEV Mpro Phe139 and Tyr160) that are proposed to stabilize the neutral state of His162 (see text for details).
The final model comprises 1798 amino acid residues and 1006 water molecules, as well as 27 sulfate ions, nine dioxane molecules and six 2-methyl-2,4-pentanediol (MPD) molecules from the crystallization medium. The refinement converged to a final $R$-factor of 0.210 and an $R_{free}$ (Bruènger, 1992) of 0.256, with good stereochemistry. Altogether, 88.4% of the amino acid residues were found in the most favoured regions of the Ramachandran plot, and 10.8% were in additionally allowed regions. Residues Asn70, Asn71 and Ser279 were in regions only generously allowed, but had clear electron density.

**Domain structure**

The six TGEV M$^{\text{pro}}$ monomers present in the asymmetric unit are arranged in three dimers (Figure 3). Each monomer is folded into three domains, the first two of which are antiparallel β-barrels reminiscent of those found in serine proteinases of the chymotrypsin family (Figure 4). Residues 8–100 form domain I, and residues 101–183 make up domain II. The connection to the C-terminal domain III is formed by a long loop comprising residues 184–199. Domain III (residues 200–302) contains a novel arrangement of five α-helices. A deep cleft between domains I and II, lined by hydrophobic residues, constitutes the substrate-binding site. The catalytic site is situated at the centre of the cleft.

The interior of the β-barrel of domain I consists entirely of hydrophobic residues. A short α-helix (helix A; Tyr53–Ser58) closes the barrel like a lid. Domain II is smaller than domain I and also smaller than the homologous domain II of chymotrypsin and hepatitis A virus (HAV) 3C$^{\text{pro}}$ (Tsukada and Blow, 1985; Allaire et al., 1994; Bergmann et al., 1997). Several secondary structure elements of HAV 3C$^{\text{pro}}$ (strands b2II and cII and the intervening loop) are missing in the TGEV M$^{\text{pro}}$. Also, the domain II barrel of the TGEV M$^{\text{pro}}$ is far from perfect (Figure 4). The segment from Gly135 to Ser146 forms a part of the barrel, even though it consists mostly of consecutive loops and turns. In fact, in contrast to domain I, a structural alignment of domain II has proven difficult. The superposition of domains I and II of the TGEV M$^{\text{pro}}$ onto those of the HAV 3C$^{\text{pro}}$ yields an r.m.s.d. of 1.85 ± 0.05 Å for 114 equivalent (out of 184 compared) C$_\alpha$ pairs, while domain II alone displays an r.m.s.d. of 3.25 ± 0.28 Å for 57 (out of 85) C$_\alpha$ pairs.

Domain III is composed of five, mostly antiparallel, α-helices and the loops connecting them. The crossover angles are ~90° between helices B and E, ~30° between B and D, ~20° between C and E, and ~80° between E and F, whereas C–B and B–F are parallel to each other (see Figure 5). Interhelical contacts are mediated by hydrophobic side chains. The loops between the helices are quite long and fill up most of the interstitial space of domain III. Database searches (Holm and Sander, 1993; Gilbert et al., 1999) did not reveal other proteins or protein domains with the same topology as domain III. The N-terminal segment (residues 1–5) of the polypeptide chain folds onto domain III, placing the N-terminus of the protein within 17.0 ± 2.7 Å of the C-terminus (Figure 4). The six copies of the TGEV M$^{\text{pro}}$ in the asymmetric unit of the crystal are highly similar. The core regions of domains I and II display an r.m.s.d. of 0.29 ± 0.09 Å for 130 equivalent C$_\alpha$ atoms (monomer A as a reference; herein, geometrical values given are the r.m.s. over the six monomers, with the corresponding standard deviation). If all 299 well-determined C$_\alpha$ positions are included, the average r.m.s.d. for all monomers is 0.57 ± 0.18 Å. The largest deviations of the main chain trace are in: (i) the N-terminal segment from residues 1 to 4 (average r.m.s.d. 1.69 ± 0.91 Å); (ii) the flexible surface loop from residues 216 to 225 (average r.m.s.d. 0.99 ± 0.51 Å); (iii) the C-terminus of helix E and the loop region between residues 267 and 276 (average r.m.s.d. 0.99 ± 0.42 Å); and (iv) the segment 294–300 following the C-terminal F helix (average r.m.s.d. 1.55 ± 0.44 Å). In addition to being flexible and at the surface of the molecules, segments...
(ii) and (iii) are involved in interdimer crystal contacts in some but not all of the six protomers. Surprisingly, the regions with the highest r.m.s.d. are not the regions with the highest temperature factors, except for the C-terminal domain of monomer F which does have high temperature factors (~70 Å²; whole model 47 Å², including all 1006 water molecules).

**Active site**

The active site of the coronavirus M\(^{\text{pro}}\) is similar to those of the picornavirus 3C proteinases, as had been predicted earlier (Gorbalenya et al., 1989b). The mutual arrangement of the nucleophilic Cys144 and the general acid–base catalyst His41 of TGEV M\(^{\text{pro}}\) is identical to that of the HAV 3C\(^{\text{pro}}\) Cys172 and His44 residues and the Ser195 and His57 residues of chymotrypsin. The distance between the sulfur atom of Cys144 and the \(\text{Ne}^2\) of His41 is 4.05 (±0.04) Å, i.e. longer than the corresponding cysteine–histidine distances in HAV 3C\(^{\text{pro}}\) (3.92 Å; Bergmann et al., 1997), poliovirus (PV) 3C\(^{\text{pro}}\) (3.4 Å; Mosimann et al., 1997) and papain (3.65 Å; Kamphuis et al., 1984) (Figure 6B and C). In contrast to papain, but in agreement with the picornavirus 3C proteinases, the sulfur atom is in the plane of the histidine imidazole. There are clear indications from the difference Fourier synthesis (Figure 6A) that Cys144 is oxidized, at least to the stage of the sulfinic acid, -SO\(_2\)H, and probably to the sulfonic acid, -SO\(_3\)H, in all six copies of TGEV M\(^{\text{pro}}\) in the crystal. Such oxidation could occur during the time required for crystallization or during X-ray data collection, and would lead to inactivation of the enzyme. Refinement of the corresponding derivatives was, however, not successful.

It is generally assumed that the native state of the active site of papain-like cysteine proteinases is a thiolate–imidazolium ion pair formed by cysteine and histidine residues (Polgár, 1974). In proteinases of the papain family, an asparagine is the third member of the catalytic triad. Chymotrypsin and other members of this serine proteinase family have a catalytic triad consisting of Ser195...His57...Asp102. In HAV 3C\(^{\text{pro}}\), Asp84 is present at the required position, although its side chain points away from His44, making its role disputable (Malcolm, 1995; Bergmann et al., 1997). PV 3C\(^{\text{pro}}\), human rhinovirus (HRV) 3C\(^{\text{pro}}\) and HRV 2A\(^{\text{pro}}\) have a glutamate or aspartate in the proper orientation to accept a hydrogen bond from the active site histidine (Matthews et al., 1994; Mosimann et al., 1997; Petersen et al., 1999). In contrast, TGEV M\(^{\text{pro}}\) has Val84 in the corresponding position, with its side chain pointing away from the catalytic site (Figure 6B and C). A buried water molecule is found in the place that normally would be occupied by the side chain of the third member of the catalytic triad. This water molecule makes hydrogen bonds to His41 N\(^{\delta1}\), His163 N\(^{\delta1}\) and Asp186 O\(^{\delta1}\) (Figure 6B). His163 is not conserved among coronavirus main proteinases and its substitution by leucine (M\(^{\text{pro}}\)-H163L) had no significant effect on the proteolytic activity in the standard peptide assay (see Materials and methods), as compared with the activity of the wild-type M\(^{\text{pro}}\) (Table I). Asp186 makes a salt bridge to Arg40 that appears to be required to maintain the active site geometry, since both Asp186 and Arg40 are absolutely conserved among coronaviruses. Through this (and other) interaction(s), the polypeptide segment 184–199, which connects domains II and III and is probably involved in substrate binding (see below), is held in the proper
position. Taken together, the data contradict a direct involvement of His163 or Asp186 in catalysis, making the TGEV Mpro a clear case of a viral cysteine proteinase employing only a catalytic dyad.

Substrate hydrolysis by cysteine and serine proteinases occurs through a covalent tetrahedral intermediate resulting from attack of the active site nucleophile on the carbonyl carbon of the scissile bond. The developing oxyanion is stabilized by strong hydrogen bonds donated by amide groups of the enzyme. This so-called ‘oxyanion hole’ is also found in TGEV Mpro (Figure 6B).

Substrate-binding site

The specificity of Mpro for a very limited range of amino acids at the P1, P2 and P4 positions resembles the substrate specificity of picornavirus 3C proteinases (Palmenberg, 1990; Ziebuhr et al., 2000). This leads us to believe that, similarly to 3Cpro (Matthews et al., 1994; Bergmann et al., 1997; Mosimann et al., 1997), specific substrate binding by Mpro is ensured by well-defined S4, S2 and S1 specificity pockets. In order to visualize potential interactions with the substrate, we have modelled a pentapeptide representing the P5±P1 residues of a TGEV Mpro cleavage site (Asn±Ser±Thr±Leu±Gln, pp1a amino acids 2874±2878; Hegyi and Ziebuhr, 2002) into the substrate-binding cleft of Mpro (Figure 7). The model is based on the assumption that Mpro binds substrates in a manner analogous to that found in complexes of chymotrypsin-like proteinases with peptide inhibitors. X-ray structures have shown that the P4±P1 residues of peptide inhibitors assume a common main chain conformation when bound to these proteinases, with the P4 and P3 residues adopting a β conformation and the P2 and P1 residues assuming a specific main-chain conformation suitable to place their side chains in the pre-formed S1 and S2 specificity pockets.
These studies lead us to suggest that the residues P5 to P3 of Mpro substrates may form an antiparallel β-sheet with segment 164–167 of the long strand eII on one side, and with the segment 186–191 (which links domains II and III) on the other. Hydrogen bonding interactions are likely between the main chain amide and carbonyl oxygen atoms of substrate residues Thr(P3), Ser(P4) and Asn(P5) and the main chain atoms of TGEV Mpro residues Glu165, Ser189 and Gly167 (see Figure 7).

Table I. Enzymatic activities of TGEV Mpro mutants

<table>
<thead>
<tr>
<th>Plasmid</th>
<th>Oligonucleotides used for cloning or mutagenesis (5′→3′)</th>
<th>Protein</th>
<th>Mpro amino acids</th>
<th>Activity (%)a</th>
</tr>
</thead>
<tbody>
<tr>
<td>pMal-Mpro</td>
<td>TCAGGTGTTGCAGGAAATGACAC, AAAAGGTACCTTATGGAAGATTTACACCATACATTGT</td>
<td>Mpro</td>
<td>Ser1–Gln302</td>
<td>100</td>
</tr>
<tr>
<td>pMal-MproΔ184–302</td>
<td>TCAGGTGTTGCAGGAAATGACAC, AAAAGGTACCTTATGGAAGATTTACACCATACATTGT</td>
<td>MproΔ184–302</td>
<td>Ser1–Gly183</td>
<td>&lt;0.02</td>
</tr>
<tr>
<td>pMal-MproΔ200–302</td>
<td>AAAAGGTACCTTATGGAAGATTTACACCATACATTGT</td>
<td>MproΔ200–302</td>
<td>Ser1–Ser199</td>
<td>0.4</td>
</tr>
<tr>
<td>pMal-MproΔ1–5Δ200–302</td>
<td>ATGGCAACAGCTTTAGTGGCTTGTGA, AAAAGGTACCTTATGGAAGATTTACACCATACATTGT</td>
<td>MproΔ1–5Δ200–302</td>
<td>Met6–Ser199</td>
<td>0.6</td>
</tr>
<tr>
<td>pMal-MproΔ1–5</td>
<td>AAAGGTACCTTATGGAAGATTTACACCATACATTGT</td>
<td>MproΔ1–5</td>
<td>Met6–Gln302</td>
<td>0.3</td>
</tr>
<tr>
<td>pMal-Mpro-H163L</td>
<td>ATGGCAACAGCTTTAGTGGCTTGTGA, AAAAGGTACCTTATGGAAGATTTACACCATACATTGT</td>
<td>Mpro-H163L</td>
<td>Ser1–Gln302 (His163→Leu)</td>
<td>98</td>
</tr>
<tr>
<td>pMal-Mpro-C144A</td>
<td>ATGGCAACAGCTTTAGTGGCTTGTGA, AAAAGGTACCTTATGGAAGATTTACACCATACATTGT</td>
<td>Mpro-C144A</td>
<td>Ser1–Gln302 (Cys144→Ala)</td>
<td>&lt;0.02</td>
</tr>
</tbody>
</table>

The sequence of the 15mer substrate peptide, H2N-VSVNSTLQSLRKM-AOOH, was derived from the N-terminal Mpro autoprocessing site (residues shown in bold indicate the scissile bond). The activity of wild-type Mpro (encompassing 302 residues) was taken as 100% and the mean value of three experiments, which did not vary by more than 15%, is shown.

aProteolytic activities were determined using a peptide-based cleavage assay (Ziebuhr et al., 1997; see Materials and methods).

S1 subsite

It has been shown for the HAV, HRV and PV 3Cpro enzymes that the imidazole side chain of a conserved histidine, which is located in the centre of a hydrophobic pocket, interacts with the P1 carboxamide side chain of the substrate. This interaction is generally accepted to determine the picornavirus 3Cpro specificity for glutamine at P1 (Matthews et al., 1994, 1999; Bergmann et al., 1997; Mosimann et al., 1997). Mutational analyses revealed that any replacement of His162 completely abolished the proteolytic activities of the HCoV and feline infectious...
peritonitis virus (FIPV) M^pro enzymes (Ziebuhr et al., 1997; Hegyi et al., 2002). The structure shows that the imidazole side chain of His162 is positioned suitably to interact with a P1 glutamine side chain. His162 is located at the very bottom of a hydrophobic pocket which is formed by residues Phe139 and the main-chain atoms of Ile140, Leu164, Glu165 and His171. The side chain of Glu165 forms an ion pair (2.96 ± 0.14 Å) with His171. This salt bridge is itself on the periphery of the molecule, forming part of the ‘outer wall’ of the S1 subsite. Accordingly, mutants of the HCoV 229E M^pro, in which the residue equivalent to His171 had been replaced by alanine, serine or threonine, retained significant proteolytic activities (Ziebuhr et al., 1997). In order to interact with the P1 glutamine side chain of the substrate, His162 has to maintain a neutral state over a wide pH range. Most probably, this is achieved by two important interactions: (i) stacking onto the phenyl ring of Phe139, at a distance of 3.53 ± 0.18 Å; and (ii) accepting a hydrogen bond from the buried Tyr160 hydroxyl group which has no other hydrogen-bonding partner. The role proposed for the hydroxyl group of Tyr160 is strongly supported by FIPV M^pro mutagenesis studies in which the proteolytic activities of Y160F, Y160G, Y160A and Y160T mutants were shown to be dramatically reduced (Hegyi et al., 2002). Tyr160 is part of the absolutely conserved coronavirus M^pro sequence signature, ^160^Tyr-X-His^162 (Figures 1 and 2), whereas Gly(Ala)-X-His is found at the equivalent sequence position in most 3C and 3C-like proteinases (Gorbalenya et al., 1989a). Accordingly, in the 3C and 3C-like proteinases, stabilization of histidine in the neutral tautomeric state has to be ensured by other residues. Notably, in the case of PV 3C^pro, this involves a tyrosine residue (Tyr138) which, however, is provided by a different part of the structure (β-strand cII; Mosimann et al., 1997). For HAV 3C^pro, other mechanisms are proposed (Bergmann et al., 1997).

Halfway down the S1 subsite of TGEV M^pro, there is dumbbell-shaped electron density which we have assigned to two water molecules, although theoretically they are too close to one another (2.10 ± 0.16 Å). One of them makes a hydrogen bond with N^2 of His162, while the second one, unusually for water, makes no additional contacts. In our model of the substrate complex, these two water molecules mark the position of the carboxamide group of the P1 glutamine side chain.

S2 subsite
Coronavirus main proteinases have a strong preference for leucine at the P2 position (Ziebuhr et al., 2000). The putative S2 subsite identified in the structure is a hydrophobic pocket that is suitably positioned and large enough to accommodate a leucine side chain easily. The S2 pocket is lined by the side chains of Leu164 (the main chain of which forms part of the S1 subsite, see above), Pro188, Ile51, His41 and Thr47 (Figure 7). In our electron density maps, part of the S2 subsite (of all six copies of the monomer) harbours extra electron density that we interpreted as an MPD molecule from the crystallization medium. In the HAV 3C^pro, the corresponding subsite is formed by different parts of the polypeptide chain. It is also smaller and can accommodate the side chains of serine and threonine (Bergmann et al., 1997).

Quaternary structure
The quaternary arrangement of the proteinase is a homodimer, with three copies in the asymmetric unit (monomers A and B, C and D, and E and F). All dimers have approximate C2 symmetry (Figure 3) and ~1580 (±199) Å^2 of each monomer, i.e. 11–12% of its solvent-accessible surface, are buried upon dimerization. The dimer formation is driven mainly by intermolecular interactions between domains II and III of one monomer and the N-terminal residues of the other (see below for
further details). In contrast, the domain III–domain III interface appears to be the consequence rather than the cause of other intermolecular interactions. It involves a relatively small area of \(337 \pm 45 \text{ Å}^2\) and comprises only two hydrogen bonds, between the amide group of Gly281 (molecule A) and the main-chain oxygen of Ser279 (molecule B), as well as its symmetry mate, Gly281B...Ser279A (3.22 \pm 0.37 \text{ Å}, averaged over all six monomers).

Interestingly, the N-terminal residues of each monomer are relatively close to the substrate-binding site of the other monomer in the dimer. The following observations for monomer A hold true for all other monomers. The NH\(_3^+\) group of Ser1A, which is the P1′ residue of the autocleavage reaction of TGEV M\(^{pro}\), is 11.9 \pm 1.6 \text{ Å} from the active site Cys144B S\(^{\gamma}\) of the second molecule in the dimer but as much as 34.2 \pm 0.9 \text{ Å} away from its own active site cysteine. Ser1A is in contact with residues participating in the substrate-binding site of monomer B. Its NH\(_3^+\) group makes a salt bridge (4.99 \pm 1.04 \text{ Å}) to the carboxylate of Glu165B (Figure 8). This glutamate, which is absolutely conserved among coronaviruses, is part of the S1 subsite (see above), where it also interacts with His171. Although these two side chains form the ‘wall’ of the specificity site, they have their polar groups oriented towards the surface of the protease molecule and away from the substrate’s P1 glutamine. An intermolecular ionic interaction between Arg4A and Glu286B (6.0 \pm 0.7 \text{ Å}) appears to play a role in positioning the N-terminal residues. Because of the 2-fold non-crystallographic symmetry (NCS), the same interaction occurs between Arg4B and Glu286A. Residues 6A–8A form a short \(\beta\)-strand interacting with strand cII of monomer B (at Val124B). Most of the interactions between the N-terminus of molecule A and the region next to the S1 subsite of molecule B constitute a perfect fit. Given the fact that the P′ residues in serine and cysteine proteinases constitute the leaving group of the cleavage reaction and, in coronavirus main proteases, are not subject to stringent specificity requirements, it is quite conceivable that, after autoproteolysis, the N-terminus of one monomer slides over the active site of the partner monomer and adopts the position seen in our crystal structure, i.e. with Ser1A interacting with Glu165B at the ‘outer wall’ of the

Fig. 8. Intra- and intermolecular contacts of the TGEV M\(^{pro}\) N-terminus. (A) MOLSCRIPT stereo representation of a TGEV M\(^{pro}\) dimer. Molecule A is coloured from blue at the N-terminus, via green (domain II), to red (C-terminus), while molecule B is shown in grey. The catalytic Cys144 and His41 residues are labelled in both monomers. (B) Detailed view of the interactions made by the N-terminal segment (blue) and domains II/III of monomer A as well as domains II/III of monomer B. Residues critically involved in these interactions are designated by the single-letter code and shown in ball-and-stick representation (see text for details). The N- and C-termini of molecule A are indicated.
S1 subsite. This, in turn, would suggest that the dimer we are seeing corresponds to the product of the autolysis reaction and that this occurs in trans. Molecular modelling revealed that binding of the Mpro N-terminus in the active site cleft of the same molecule would require remodelling of the entire N-terminal segment and beyond (residues 1–13; data not shown), making cleavage in cis less likely. There is additional experimental evidence supporting these conclusions. First, dilution experiments with MHV Mpro translated in vitro contradict cis-cleavage activity (Lu et al., 1996). Secondly, the fact that, early in infection, Mpro remains part of a relatively stable 150 kDa precursor protein in which it is flanked by hydrophobic domains (Schiller et al., 1998) argues against rapid autoprocessing in cis. The proposed model of intermolecular self-processing would imply that components of the replication complex could first be anchored to membranes (i.e. the site of RNA replication) in an uncleaved form, and only later, when the precursor proteins accumulate to high local concentrations, will Mpro release itself by intermolecular cleavage, thereby triggering the complete spectrum of trans-processing reactions.

**Intramolecular interactions of the N-terminus**

A specific conformation of the N-terminal segment allows it to ‘squeeze’ residues 1–8 in between domains II and III of the same monomer and domains II and III of monomer B (see above and Figure 8). In this context, the N-terminus also interacts with domains II and III of its own protomer. For example, the side-chain amino group of Lys5A makes strong intramolecular hydrogen bonds with Ser110A O' of domain II (2.83 ± 0.15 Å), and with the Glu286A main chain oxygen (2.80 ± 0.07 Å), as well as with Glu291A O' (2.74 ± 0.13 Å) of domain III. Furthermore, the side chain of Leu3A completes a hydrophobic patch on domain III which includes Phe206A, Ala209A, Phe287A, Val292A, the Cα atom of Gln295A and Met296A; these residues belong to helices B and F. All sequenced members of the coronavirus proteinase family have a hydrophobic residue in position 3, while glycine is absolutely conserved in position 2 (see Figure 1). The latter residue adopts the αL conformation which is easily accessible only to glycine. To investigate the functional significance of these interactions, a recombinant protein, MproΔ1–5, in which the N-terminal residues Ser1–Lys5 were removed from the Mpro sequence, was expressed and tested for proteolytic activity in a trans-cleavage assay using a 15mer peptide representing the N-terminal TGEV Mpro autoprocessing site. As shown in Table I, the activity of MproΔ1–5 was decreased to only 0.3% of the Mpro activity. We conclude from these data that, indeed, residues 1–5 may be critically involved in stabilizing the mutual orientation of domains II and III and thus, indirectly, in maintaining the proper orientation of the intervening loop region (residues 184–199). If this hypothesis is correct, then the deletion of domain III should have similarly detrimental effects on the proteolytic activity and, in fact, the published data (see Introduction) seem to support this conclusion. To corroborate this hypothesis further, an additional set of Mpro mutants was characterized in which we used the structural information to remove domain III completely. In this approach, the probability of domain III misfolding, which might have been the cause of Mpro inactivation in previous studies using randomly ‘truncated’ coronavirus main proteinases (Lu and Denison, 1997; Ziebuhr et al., 1997; Ng and Liu, 2000), should be significantly reduced. The TGEV Mpro deletion mutants tested for activity comprised (i) domains I and II (MproΔ184–302); (ii) domains I and II together with the entire loop region (MproΔ200–302); or (iii) domains I and II combined with the loop region but lacking the five N-terminal residues (MproΔ1–5/Δ200–302). As Table I shows, MproΔ200–302 had clearly detectable (albeit significantly reduced) activity (0.4% of Mpro). Similarly, the mutant MproΔ1–5/Δ200–302 had significantly reduced activity (0.6% of Mpro). In sharp contrast, no activities were detectable for MproΔ184–302 and the active site mutant, Mpro-C144A (the latter being used as a negative control). The fact that residues 184–199 proved to be indispensable for proteolytic activity supports our model of substrate binding (Figure 7) in which residues of the loop are predicted to be critically involved in the formation of a β-sheet-type structure with the substrate (see above). The data also show that an intact N-terminus and the C-terminal domain are required for full activity. The structure suggests that the additional α-helical domain III as well as the N-terminal residues help fix domains II and the loop 184–199 in a catalytically competent orientation. It will be interesting to investigate whether similar mechanisms are also operating in other 3C-like proteinases with (smaller) C-terminal domains (e.g. arteriviruses and potyviruses; Ziebuhr et al., 2000; Hegyi et al., 2002).

Beyond its presumed role in proteolytic activity, domain III may have other functions, which remain to be determined. In contrast to picornavirus 3C proteinases for which RNA-binding activities are well established (Andino et al., 1993; Leong et al., 1993; Xiang et al., 1995), the Mpro structure does not support such an activity for the coronavirus main proteinase. Thus, calculation of the electrostatic potential (Nicholls et al., 1991) does not reveal an overall basic character of domain III, nor are there distinct patches of basic or aromatic residues (data not shown). The same applies to domains I and II. Also, the conserved picornavirus sequence motif, KFRDI, located between domains I and II, as well as the small helices and reverse turns that together form the RNA-binding site of HAV 3Cpro (Bergmann et al., 1997) are missing in the TGEV Mpro structure.

**Conclusion**

The crystal structure of TGEV Mpro shows that coronaviruses have evolved proteinases in which a thiolate-imidazolium catalytic dyad has been combined with a two-β-barrel fold. This framework is extended further by a novel α-helical domain that, together with the N-terminal residues 1–5, appears to be involved in proteolytic activity by maintaining the proper positioning of the presumed substrate-binding loop, 184–199. We are confident that the first crystal structure of a non-picornaviral chymotrypsin-like cysteine proteinase will facilitate further molecular modelling of other members of the huge family of RNA viral ‘3C-like’ enzymes for which structural information is still lacking.
Materials and methods

Protein purification and crystallization

Recombinant TGEV M\textsuperscript{apo} was expressed and purified as previously described for the HCoV and FIPV main proteinases (Ziebuhr et al., 1997; Hegyi et al., 2002). Briefly, the coding sequence of the TGEV M\textsuperscript{apo} was inserted into the XmnI and BamHI sites of pMal-c2 plasmid DNA (New England Biolabs). The resulting plasmid, pMal-M\textsuperscript{apo}, was used to transform Escherichia coli TB1 cells. The maltose-binding protein (MBP)–TGEV M\textsuperscript{apo} fusion protein was purified by amylose–agarose chromatography, cleaved with factor Xa, and the recombinant M\textsuperscript{apo} (residues Ser1–Gln302) was purified by hydrophobic interaction, anion exchange and size exclusion chromatography (Hegyi et al., 2002). The purified and concentrated TGEV M\textsuperscript{apo} (12.5 mg/ml) was stored in 12 mM EDTA. This protein solution was used to crystallize Mpro by the hanging drop vapour diffusion method at 4°C.

Incorporation of selenomethionine

The M\textsuperscript{apo} structure could not be solved using conventional molecular replacement techniques. Therefore, selenomethionine (SeMet)-substituted MBP±TGEV Mpro was produced. The coding sequence of the MBP±TGEV Mpro was inserted into pMal-c2 plasmid DNA (New England Biolabs). The resulting plasmid, pMal-M\textsuperscript{apo}, was used to transform Escherichia coli TB1 cells. The maltose-binding protein (MBP)–TGEV M\textsuperscript{apo} fusion protein was purified by amylose–agarose chromatography, cleaved with factor Xa, and the recombinant M\textsuperscript{apo} (residues Ser1–Gln302) was purified by hydrophobic interaction, anion exchange and size exclusion chromatography (Hegyi et al., 2002). The purified and concentrated TGEV M\textsuperscript{apo} (12.5 mg/ml) was stored in 12 mM EDTA. This protein solution was used to crystallize Mpro by the hanging drop vapour diffusion method at 4°C. The best crystals, which were of triangular shape and had dimensions of ~0.3 × 0.3 × 0.3 mm, were obtained by using 100 mM HEPES pH 8.8, 1.8 M ammonium sulfate, 6% MPD, 5 mM DTT and 4% dioxane as the reservoir and grew in ~10 days.

Diffraction data collection

Crystals used for data collection were rinsed with mustard oil and cryo-cooled in liquid nitrogen. Diffraction data up to 1.95 Å resolution were collected from native crystals at 100 K on the X-ray diffraction beamline at ELETTRA, Trieste, equipped with a Mar CCD detector.

Structural determination

The unit cell dimensions, as well as the self-rotation function (ALMN; CCP4, 1994), implied that several monomers were present in the asymmetric unit. A Matthews coefficient (Matthews, 1968) of 2.3 Å\textsuperscript{3}/Da and a solvent content of 51% were obtained assuming six molecules in the asymmetric unit. The bottleneck of the structure determination was the identification of the 60 selenium atoms (six monomers with 10 Se each). Solving the problem by SnB v2.0 (Weeks and Miller, 1999) required data of increased precision, which were obtained by averaging of several data sets and monitoring the process by \( R_{	ext{merge}} \) (Weiss and Hilgenfeld, 1997). Only after we had combined three merged peak-wavelength data sets with two merged edge-wavelength data sets (redundancy = 18) were we able to obtain 105 solutions (from 5000 trials) with significantly reduced minimal function values (\( R_{	ext{merge}} = 0.49, \text{CC} = 0.51) \) (Hauptman, 1991) (details to be published elsewhere).

The positions of the best 60 atom solutions from SnB were examined for NCS. In total, 37 positions were found to obey a 2-fold NCS. This symmetry predicted a further 11 positions. All 48 positions were used in MLPHARE (CCP4, 1994) for phasing, followed by solvent flattening and NCS averaging in DM (Cowtan and Main, 1996). The resulting electron density maps were of sufficient quality for chain tracing. The first monomer was built manually into the experimental electron density map, using the program ‘O’ (Jones et al., 1991). All other monomers were generated by NCS. NCS restraints were applied during the initial stages of refinement at low resolution and later gradually released as the resolution limit was extended to 1.96 Å.

Cycles of adjustments to the model with O and subsequent refinement using the program CNS (Brünger et al., 1998) converged to an \( R_{	ext{free}} \) of 0.256 and a crystallographic R-factor of 0.210. Data quality and refinement statistics are given in Table III. The quality of the structural model and its agreement with the structure factors were checked with programs PROCHECK (Laskowski et al., 1993), WHATCHECK (Friend, 1990) and SFCHECK (Vaguine et al., 1999). Solvent accessibility was calculated using the algorithm of Lee and Richards (1971; program NACCESS), using a solvent probe of radius 1.4 Å. The molecular diagrams were drawn using MOLSCRIPT (Kraulis, 1991) and rendered with RASTER 3D (Bacon and Anderson, 1988). Atomic coordinates and structure factors have been submitted to the RCSB Protein Data Bank under accession code ILVO.
Phasing

- FOM after solvent flattening: 0.48
- FOM after solvent flattening (no averaging): 0.72
- FOM after solvent flattening (with averaging): 0.27

Re-refinement

Resolution (Å): 50–1.96

No. of non-hydrogen atoms (average): 
- Protein (main chain): 7198 (46.1)
- Water: 1006 (50.3)
- MPD: 48 (67.6)
- Sulfate: 135 (57.1)
- Dioxane: 54 (71.7)

R.m.s. deviation from ideal geometry

- Bonds (Å): 0.017
- Angles (%): 1.9
- Improper dihedral angles (%): 1.16

\[ \text{FOM} = \text{figure of merit.} \]

\[ \text{R-factor} = \frac{\sum |F_o| - k|F_i|}{\sum |F_o|}, \text{ where } k \text{ is the scale factor.} \]

**Proteolytic activities of TGEV Mpro mutations**

For the expression of Mpro proteins with N- and C-terminal deletions (Mpro(18–302), Mpro(200–302), Mpro(1–5) and Mpro(1–5/43–48)), the corresponding Mpro coding sequences were amplified by PCR and inserted into XmnI–BamHI-digested pMal-c2 plasmid DNA. To substitute the Mpro residues Cys144 (by Ala) and His163 (by Leu), the corresponding codons were replaced in pMal-Mpro by site-directed mutagenesis using a recombination-PCR method (Yao et al., 1992). The details of the primers used for cloning and mutagenesis and the amino acid sequences of the recombinant proteins expressed and tested for proteolytic activity are given in Table I. The plasmid DNAs were transformed into E.coli TB1 cells and the recombinant proteins were synthesized, affinity purified and cleaved with factor Xa as described previously (Hegyi et al., 2002). The purity and structural integrity of the mutant proteins were analysed by SDS–PAGE. The control protein for using a peptide cleavage assay (Ziebuhr et al., 1999-00017, respectively). We thank M.S.Weiss and D.Pal for their support in the 5’–factor = \(|F_o| - |k|F_i|)/|F_o|, \text{ where } \text{FOM}_a \text{ after solvent flattening (with averaging) } 0.79.

**Structure of TGEV main proteinase**


protease 3C (3C\textsuperscript{pro}) binds specifically to the 5' non-coding region of the viral RNA. J. Virol. Chem., 268, 25735–25739.


