Production of Authentic SARS-CoV M\(^{\text{Pro}}\) with Enhanced Activity: Application as a Novel Tag-cleavage Endopeptidase for Protein Overproduction

Xiaoyu Xue\(^1,2\), Haitao Yang\(^1,2\), Wei Shen\(^1\), Qi Zhao\(^1\), Jun Li\(^1\)
Kailin Yang\(^1\), Cheng Chen\(^1\), Yinhua Jin\(^1\), Mark Bartlam\(^1,2\)
and Zihe Rao\(^1,2,3\)⁎

\(^1\)Tsinghua-Nankai-IBP Joint Research Group for Structural Biology, Tsinghua University, Beijing 100084, China
\(^2\)National Laboratory of Biomacromolecules, Institute of Biophysics (IBP), Chinese Academy of Sciences, Beijing 100101, China
\(^3\)Nankai University, Tianjin 300071, China

The viral proteases have proven to be the most selective and useful for removing the fusion tags in fusion protein expression systems. As a key enzyme in the viral life-cycle, the main protease (M\(^{\text{Pro}}\)) is most attractive for drug design targeting the SARS coronavirus (SARS-CoV), the etiological agent responsible for the outbreak of severe acute respiratory syndrome (SARS) in 2003. In this study, SARS-CoV M\(^{\text{Pro}}\) was used to specifically remove the GST tag in a new fusion protein expression system. We report a new method to produce wild-type (WT) SARS-CoV M\(^{\text{Pro}}\) with authentic N and C termini, and compare the activity of WT protease with those of three different types of SARS-CoV M\(^{\text{Pro}}\) with additional residues at the N or C terminus. Our results show that additional residues at the N terminus, but not at the C terminus, of M\(^{\text{Pro}}\) are detrimental to enzyme activity. To explain this, the crystal structures of WT SARS-CoV M\(^{\text{Pro}}\) and its complex with a Michael acceptor inhibitor were determined to 1.6 Å and 1.95 Å resolution respectively. These crystal structures reveal that the first residue of this protease is important for sustaining the substrate-binding pocket and inhibitor binding. This study suggests that SARS-CoV M\(^{\text{Pro}}\) could serve as a new tag-cleavage endopeptidase for protein overproduction, and the WT SARS-CoV M\(^{\text{Pro}}\) is more appropriate for mechanistic characterization and inhibitor design.

© 2006 Elsevier Ltd. All rights reserved.

† X.X. and H.Y. made equal contributions to this work.

Abbreviations used: SARS, severe acute respiratory syndrome; SARS-CoV, SARS coronavirus; M\(^{\text{Pro}}\), main protease; GST, glutathione-S-transferase; WT, wild-type enzyme with authentic N and C termini; GS-WT, wild-type enzyme with two additional amino acid residues (GS) at the N terminus; GPLGS-WT, wild-type enzyme with five additional amino acid residues (GPLGS) at the N terminus; WT-GPH\(_6\), wild-type enzyme with eight additional amino acid residues (GPHHHHHH) at the C terminus.

E-mail address of the corresponding author: raozh@xtal.tsinghua.edu.cn

Keywords: SARS coronavirus; main protease; fusion protein; crystal structure; endopeptidase

Introduction

In the age of proteomics, the production of pure protein in a high-throughput manner is required...
proteases with high specificity and activity will be helpful to proteomics research.

The SARS coronavirus (SARS-CoV) is the etiological agent responsible for the global outbreak of a life-threatening disease that caused approximately 800 deaths worldwide. The coronavirus main protease (Mpro), which plays a key role in mediating viral replication and transcription, has been identified as the most attractive target for anti-SARS drug design. In 2003, our group published the first crystal structure of SARS-CoV Mpro. Thereafter, several other groups published the structures of Mpro, and its complex with an azapeptide epoxide inhibitor. The Mpro can form a homodimer both in the crystal and in solution. Each protomer consists of three domains: domains I and II resemble chymotrypsin, whereas domain III has a globular cluster of five, mostly antiparallel, α-helices. SARS-CoV Mpro has a catalytic dyad consisting of His41 and Cys145, and its substrate-binding pocket is located in the cleft between domains I and II. The N-terminal residues 1–7 of domain I (or N-finger) of Mpro are considered to have an important role in its proteolytic activity, although their importance in dimerization has been reported with inconsistent results. In particular, mutation of Arg4, which is involved in forming an ion pair, results in a fourfold decrease in activity. Deletion of the N-terminal residues 1–7 results in an almost completely inactive SARS-CoV Mpro. SARS-CoV Mpro has been characterized sufficiently and two particular aspects drew our attention. First, it is highly selective for substrate sequence. Second, it displays a high level of proteolytic activity, although inconsistent kinetic parameters have been reported for SARS-CoV Mpro with kcat/Km ranging from 20–29,000 M–1s–1. These properties indicate that SARS-CoV Mpro is a promising candidate as an endopeptidase to remove fusion tags.

In this study, SARS-CoV Mpro was utilized to remove the fusion domain in overproduction of calbindin D28k (a member of the calmodulin superfamily) from a novel glutathione-S-transferase (GST) fusion protein expression system. To obtain the peptidase with high catalytic efficiency, we developed a new method to produce wild-type (WT) SARS-CoV Mpro with authentic N and C termini, and investigated the activity difference among several SARS-CoV Mpro constructs. The crystal structures of WT Mpro and its complex with an inhibitor were solved to explain the mechanism of enhanced activity.

Results

Construction of a new GST fusion protein expression vector (pGSTM)

In order to take advantage of SARS-CoV Mpro as an endopeptidase, a new GST fusion protein expression vector (designated pGSTM) was constructed from another fusion vector pGEX-6p-1 (GE Healthcare). The resulting plasmid retains the whole multiple cloning sites of pGEX-6p-1. However, the linker between the GST gene and the gene of interest was replaced by nucleotides encoding a unique N-terminal autocleavage site of SARS-CoV Mpro consisting of 11 amino acid residues (see Figure 1(a)). Sequencing confirmed that the cleavage site (TSAVLQSGFRK) was inserted correctly.

Expression and purification of calbindin D28k using the pGSTM expression system

Calbindin D28k, a member of the calmodulin superfamily, has been proposed to function as an important intracellular Ca2+-buffering protein. We took it as our protein of interest for structural studies. This protein was cloned, expressed and purified using our new protein expression system. After one-step purification, calbindin D28k protein, with five additional residues (SGFRK-) at the N terminus and an authentic C terminus, was analyzed by SDS-PAGE and the purity was shown to be >90% (Figure 1(b)). Around 70–80 mg of calbindin D28k was obtained from 1 L of bacterial culture. The crystallographic analysis of calbindin D28k is underway.

Preparation of WT SARS-CoV Mpro

In an earlier study, wild-type SARS-CoV Mpro with two additional residues (GS) at the N terminus (GS-WT) was used for the determination of kinetic parameters. To avoid any potential effects from the extraneous N-terminal residues on enzyme activity, we developed a new method to produce WT SARS-CoV Mpro. First, the four amino acids AVLQ, which correspond to the P4–P1 sites of the N-terminal autocleavage sequence of SARS-CoV Mpro, were introduced between the GST tag and the first residue of the protease. Thus, the authentic N terminus would become available by autocleavage during protein expression. Second, the eight amino acids GPHHHHHH (abbreviated as GPH6, where GP correspond to the P1′ and P2′ sites of rhinovirus 3C protease; see Table 1) were added after the last glutamine residue at the C terminus (Figure 1(c)). This strategy was intended to yield an authentic C terminus following cleavage by rhinovirus 3C protease; see Table 1) were added after the last glutamine residue at the C terminus (Figure 1(c)). This strategy was intended to yield an authentic C terminus following cleavage by rhinovirus 3C protease, despite appearing unreasonable, since SARS-CoV and rhinovirus belong to different families (Coronaviridae and Picornaviridae, respectively). This strategy was used for the following reasons.

With an extra domain III, SARS-CoV Mpro differs from rhinovirus 3C protease in many aspects. However, they share several similarities in both cleavage sequence and the substrate-recognition pocket. From the native/complex structures of these two proteases published to date, the S4, S2, S1, S1′ and S2′ subsites are critical for substrate binding to the two proteases. Thus, the corresponding P4, P2, P1, P1′ and P2′ sites were
taken into consideration when designing recognition sequences. Table 1 lists the C-terminal autodigestion site for SARS-CoV M\(^{\text{pro}}\) and the substrate sequence for human rhinovirus 3C protease. It

<table>
<thead>
<tr>
<th>Protease</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>
</tr>
</thead>
<tbody>
<tr>
<td>SARS-CoV M(^{\text{pro}}) (C terminus)</td>
<td>S</td>
<td>G</td>
<td>V</td>
<td>T</td>
<td>F</td>
<td>Q</td>
<td>G</td>
<td>K</td>
</tr>
<tr>
<td>Human rhinovirus 3C</td>
<td>L</td>
<td>E</td>
<td>V</td>
<td>L</td>
<td>F</td>
<td>Q</td>
<td>G</td>
<td>P</td>
</tr>
</tbody>
</table>

**Table 1.** C-terminal autodigestion site for SARS-CoV M\(^{\text{pro}}\) and typical substrate sequence for human rhinovirus 3C protease.

Figure 1. A new GST fusion protein expression system. (a) The map of the pGSTM vector design. The cleavage site for SARS-CoV M\(^{\text{pro}}\) is labeled. (b) Expression and purification of calbindin D28k using the pGSTM expression system. Lane 1, total cell extract for calbindin D28k before induction; lane 2, total cell extract for calbindin D28k after induction overnight; lane 3, supernatant of the cell lysate; lanes 4 and 5, purified calbindin D28k; lane 6, protein molecular mass marker. (c) Schematic plot of WT SARS-CoV M\(^{\text{pro}}\) construct designed. (d) Expression and purification of WT M\(^{\text{pro}}\). Lane 1, protein molecular mass marker; lane 2, total cell extract for WT M\(^{\text{pro}}\) before induction; lane 3, total cell extract for WT M\(^{\text{pro}}\) after induction overnight; lane 4, WT-GPH, after affinity chromatography; lane 5, WT after cleavage by rhinovirus 3C protease; lane 6, GPLGS-WT; lane 7, GS-WT. This shows that the five sites are identical for the two proteases, with the exception of P2’. The substantial difference embodied in the two substrate sequences is limited to P2’. Hence, we reasoned that rhinovirus 3C protease could recognize the octapeptide substrate SGVTGQGP, whose P1–P6 sites were derived from the C-terminal autoprocessing site of SARS-CoV M\(^{\text{pro}}\). However, replacing lysine with proline at the P2’ site would arrest autocleavage. Our group has recently solved the crystal structure of a mutant of SARS-CoV M\(^{\text{pro}}\) in complex with an 11 amino acid residue peptidyl substrate (unpublished results). The complex structure revealed that substitution of
P2′ with Pro would result in steric hindrance between P2′ (Pro) and the main chain of SARS-CoV, hindering substrate binding. On the basis of the above analysis, we hypothesized that rhinovirus Mpro would not recognize the octapeptide substrate, SGVTFAQ1GP. SDS-PAGE analysis (see Figure 1(d)) showed that after Ni-affinity chromatography, the SARS-CoV Mpro is in the WT-GPH6 form, implying that the N-terminal GST tag was removed efficiently by autolysis during protein expression in Escherichia coli. The difference between lane 4 and lane 5 indicates that the GPH6 tag was indeed removed by rhinovirus 3C protease cleavage, confirming that the modified substrate of SARS-CoV Mpro could be recognized by rhinovirus 3C protease.

Matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS) analysis indicated that WT-GPH6 and WT Mpros were in accord with their predicted molecular mass (see Table 2).

**Enzyme activity assay of four types of SARS-CoV Mpros**

To ascertain whether the additional residues would interfere with the catalytic activity of SARS-CoV Mpro, we determined the kinetic parameters for GPLGS-WT (with five additional residues (GPLGS) at the N terminus of WT), GS-WT, WT-GPH6 and WT SARS-CoV Mpros (see Table 3). The catalytic efficiency of an enzyme is best defined by $k_{cat}/K_m$. Table 3 shows that the WT protease has the highest cleavage efficiency ($k_{cat}/K_m = 26,500 \text{ M}^{-1} \text{s}^{-1}$) while GPLGS-WT has the lowest ($k_{cat}/K_m = 167 \text{ M}^{-1} \text{s}^{-1}$). The activity of the WT protease with authentic N and C termini was more than 150-fold greater than that of GPLGS-WT and 20-fold greater than that of GS-WT. Furthermore, our results suggest that increasing the number of additional residues at the N terminus would result in a greater decrease in activity. However, the activity of WT-GPH6 was about one-third of that of the WT enzyme, which suggests that additional residues at the C terminus had less effect on activity.

**Crystal structure of WT SARS-CoV Mpro**

In order to shed light on the difference in activity, we determined the crystal structure of the WT SARS-CoV Mpro to 1.6 Å resolution. In our published crystal structure of SARS-CoV Mpro (in the GPLGS-WT form), the crystal belongs to the space group $P_2_1$ and the asymmetric unit contains a dimer (with the two protomers designated A and B). Within the dimer, protomer A was in the active form, while protomer B showed an inactive form, resulting from a partially collapsed S1 subsite. In contrast, the crystal of WT SARS-CoV Mpro belongs to space group $C_2$, and each asymmetric unit contains only one protomer of a typical dimer (with the two protomers designated A′ and B′). The two protomers in the dimer are related by a crystallographic 2-fold symmetry axis, and each has a catalytically competent conformation. All residues of the protomer (residues 1–306) were identified from electron density maps. Apart from differences in the S1 subsite, the SARS-CoV Mpro protomer as seen in the new structure is very similar to each member of the dimer in the original structure (see Figure 2). Compared to the latter structure, the protomer in the new crystal forms display overall rmsd for $C_\alpha$ atoms of 0.6 Å for protomer A and 0.7 Å for protomer B. In the following discussion, we consider the differences between the substrate-binding sites of the two structures in further detail.

The S1 site of SARS-CoV Mpro, which has absolute specificity for Gln in the P1 site, consists of the side-chains of His163 and Phe140, and the main-chain atoms of GluA166, Asn142, Gly143 and HisA172. In the S1 site of protomer A in our original structure (GPLGS-WT), the NH group of SerB1 was unable to donate hydrogen bonds simultaneously to the carboxylate group of GluA166 and the main-chain carbonyl group of PheA140, due to the presence of additional residues at the N terminus, although the

---

**Table 2.** Molecular mass of WT-GPH6 and WT measured by MALDI-TOF MS

<table>
<thead>
<tr>
<th>SARS-CoV Mpro</th>
<th>Predicted (Da)</th>
<th>Measured (Da)</th>
</tr>
</thead>
<tbody>
<tr>
<td>WT-GPH6</td>
<td>34,822</td>
<td>34,819</td>
</tr>
<tr>
<td>WT</td>
<td>33,846</td>
<td>33,817</td>
</tr>
</tbody>
</table>

**Table 3.** Comparison of enzyme activities of four types of SARS-CoV Mpro

<table>
<thead>
<tr>
<th>SARS-CoV Mpro</th>
<th>$K_m$ (μM)</th>
<th>$k_{cat}$ (s$^{-1}$)</th>
<th>$k_{cat}/K_m$ (M$^{-1}$ s$^{-1}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPLGS-WT</td>
<td>126±8</td>
<td>0.021±0.001</td>
<td>167</td>
</tr>
<tr>
<td>GS-WT$^{15}$</td>
<td>129±7</td>
<td>0.14±0.01</td>
<td>1100</td>
</tr>
<tr>
<td>WT-GPH6</td>
<td>61.0±2.9</td>
<td>0.41±0.02</td>
<td>6800</td>
</tr>
<tr>
<td>WT</td>
<td>40.0±0.8</td>
<td>1.06±0.04</td>
<td>26,500</td>
</tr>
</tbody>
</table>

---

**Figure 2.** Superposition of WT and GPLGS-WT Mpros. The substrate-binding pocket of one protomer is in surface representation. GPLGS-WT is in blue, WT is in magenta.
distance is suitable for hydrogen bond formation. However, in the newly solved WT structure, the amino group (NH\textsubscript{2}) of SerB*1 in one protomer donates a 3.0 Å hydrogen bond to the carboxylate group of GluA*166 (Figure 3(a)) and a 2.7 Å hydrogen bond to the main-chain carbonyl group of PheA*140, thus stabilizing the S1 pocket. This induces a series of conformational changes. For instance, the NH of GlyA*143, which participates directly in the formation of the oxyanion hole, moves by 0.8 Å towards the active site; the main chain of residues 142-143 moves towards the S1 subsite; and the side-chain of AsnA*142 flips over with a 6 Å shift. Stabilized by SerB*1, protomer A* displays a more catalytically competent conformation in the S1 subsite than protomer A in our original structure. In our original structure, the S1 pocket of protomer B is partly collapsed compared with the WT protomer: no electron density was visible for residues A1 and A2; GluB166 reorients to interact with the possibly protonated HisB163; PheB140 undergoes a dramatic conformational change, with the phenyl ring moving by as much as 10 Å; and GlyB143 moves about 3 Å towards the active site.

Figure 3. Superposition of the S1 pockets of GPLGS-WT and WT SARS-CoV M\textsuperscript{pro} (in stereo). (a) Superposition of the S1 pockets in protomer A of GPLGS-WT and that of protomer A* of WT SARS-CoV M\textsuperscript{pro}. Protomer A* of WT is in blue; protomer A of GPLGS-WT is in yellow; protomer B* of WT is in magenta; protomer B of GPLGS-WT is in red. In the WT structure, the amino group (NH\textsubscript{2}) of Ser1 in protomer B* donates a 3.0 Å hydrogen bond to the carboxylate group of Glu166 and a 2.7 Å hydrogen bond to the main-chain carbonyl group of Phe140 in protomer A*, stabilizing the S1 pocket. The NH of Gly143 moves 0.8 Å toward the activity site; the main chain of residues 142-143 moves toward the S1 subsite; the side-chain of Asn-A*142 flips over with a 6 Å shift compared with protomer A of GPLGS-WT. (b) Superposition of the S1 pockets in protomer B of GPLGS-WT and that of ProA* of WT SARS-CoV M\textsuperscript{pro}. Protomer A* of WT is in blue; protomer B of GPLGS-WT is in yellow; protomer B* of WT is in magenta; protomer A of GPLGS-WT is in red. The S1 pocket of protomer B collapses partly with reorientation of Glu166 and residues 140–143. No electron density was visible for residues A1 and A2.
leaving no space to accommodate a tetrahedral reaction intermediate (Figure 3(b)). These structural variations account for the higher activity of the WT protease compared with the GPLGS-WT and GS-WT proteases. We observed also that the GPLGS-WT protease has lower activity than the GS-WT protease. This might result from the additional flexible residues at the N terminus, which are located close to the active site and would hinder substrate binding.

In contrast to the N terminus, the C terminus of the WT protomer (residues 301–306) is located far from the substrate-binding pocket of its partner protomer (~10 Å). Therefore, the additional residues (GPH6) at the C terminus are expected to have less effect on enzyme activity (see Table 3).

**Inhibition assay of SARS-CoV Mpro**

In our previous study, we designed an irreversible anti-coronavirus inhibitor (designated as N3, Figure 4(a)) consisting of an α,β-unsaturated ester (one type of Michael acceptor) incorporated with a peptidyl portion. The evaluation of this series of time-dependent inhibitors requires a pseudo second-order rate constant (k3/Ki). Ki and k3 represent the equilibrium binding constant and inactivation rate constant for covalent bond formation, respectively. We assayed the inhibition of N3 against different types of Mpros to determine whether or not the additional residues would affect inhibitor binding. In our preliminary inhibition assays, we observed that N3 could completely inactivate WT and WT-GPH6 proteases after its preincubation with the proteases (fivefold molar excess of the enzyme) for 5 min, but not GPLGS-WT and GS-WT proteases (see Supplementary Data Figure S2). The strict kinetic parameters listed in Table 4 show that the second-order rate constant of N3 against WT protease (k3/Ki = 18,800 M⁻¹ s⁻¹) is approximately equal to that against WT-GPH6 (k3/Ki = 15,400 M⁻¹ s⁻¹). However, the second-order rate constant

---

**Figure 4.** Differences between the complex structures of WT and GPLGS-WT. (a) Inhibitor N3. (b) Superposition of the substrate-binding pockets in protomer A of GPLGS-WT and that in protomer A* of WT. In the WT-N3 complex structure, the NH₂ group of Ser1 in protomer B* was still hydrogen-bonded to the carboxylate group of Glu166 and the carbonyl group of Phe140 in protomer A*, stabilizing the S1 pocket. In the GPLGS-WT-N3 complex structure, however, the two hydrogen bonds described above were not found. Instead, an ordered water molecule was observed in the S1 pocket. Protomer A* of WT is in blue; protomer A of GPLGS-WT is in yellow; inhibitor N3 (complexed with WT) is in magenta; inhibitor N3 (complexed with GPLGS-WT) is in red; protomer B* of WT is in green; protomer B of GPLGS-WT is in cyan.
Table 4. Enzyme inhibition data of inhibitor N3 against four types of SARS-CoV Mpro.

<table>
<thead>
<tr>
<th>Protease</th>
<th>$K_i$ (μM)</th>
<th>$k_{cat}$ (s$^{-1}$)</th>
<th>$k_{cat}/K_i$ (M$^{-1}$ s$^{-1}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPLGS-WT</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>GS-WT</td>
<td>9.0±0.8</td>
<td>0.034±0.0005</td>
<td>340±27</td>
</tr>
<tr>
<td>WT-GPH$_6$</td>
<td>2.3±0.1</td>
<td>15,400±1,200</td>
<td>18,800±1,800</td>
</tr>
<tr>
<td>WT</td>
<td>1.9±0.1</td>
<td>0.035±0.002</td>
<td>18,800±1,800</td>
</tr>
</tbody>
</table>

The crystal structure of WT SARS-CoV Mpro in complex with a Michael acceptor

We determined the crystal structure of WT SARS-CoV Mpro in complex with inhibitor N3 to 1.9 Å resolution. The substantial difference between the structures of the WT complex and the GPLGS-WT complex reported previously is still lies in the S1 subsite. In the GPLGS-WT complex structure, N3 binds to protomers A and B of SARS-CoV Mpro in an identical and normal manner, thus we discuss only protomer A and A$. In protomer A of GPLGS-WT Mpro complexed with N3, the NH group of SerB1 is 3.4 Å from the carboxylate group of GluA166 and 4.7 Å from the main-chain carbonyl group of PheA140, both of which are beyond the distance for formation of a hydrogen bond. However, an ordered water molecule is situated at the bottom of the S1 subsite, connecting N3 and the protease. This water donates two hydrogen bonds, to the carboxylate group of GluA166 and the main-chain carbonyl group of PheA140, and accepts two hydrogen bonds from the lactam of N3. As a consequence, the carboxylate group of GluA166 moves 1.7 Å upwards to form a 2.8 Å hydrogen bond with the NH of the lactam. The competent conformation of the S1 subsite is still maintained via the interaction of NH$_2$ SerB$^\dagger$ with the carboxylate group of GluA$^\dagger$ (Figure 4(b)) and the main-chain carbonyl group of PheA$^\dagger$.140. These structural data demonstrate that the first Ser1 residue is important also for inhibitor binding, and accounts for the more potent inhibition of N3 against the WT than the GPLGS-WT and GS-WT proteases.

Discussion

Although SARS-CoV is notorious for causing a lethal disease in humans, some positive elements may result from this life-threatening virus. In this study, SARS-CoV Mpro was engineered to serve as a novel endopeptidase to remove fusion tags in recombinant protein overproduction. Table 5 shows the advantages of SARS-CoV Mpro compared with other routinely used proteases in methods for production, substrate specificity and cleavage efficiency. Human thrombin and bovine factor Xa are both extracted from plasma, which requires more complicated procedures, although they possess a high level of cleavage efficiency. In addition, they are not as selective for substrate as viral proteases. WT SARS-CoV Mpro, which can be suitably overexpressed in E. coli, is highly specific for substrate. Furthermore, it has superior activity ($k_{cat}/K_m = 26500 \text{ M}^{-1} \text{s}^{-1}$) to rhinovirus and TEV proteases. Therefore, SARS-CoV Mpro is a suitable candidate for site-specific cleavage in fusion protein expression systems. The cleavage efficiency of WT-GPH$_6$ is still high, despite additional residues at the C terminus. Highly purified Mpro in this form could be obtained by a simple one-step purification, as shown by the overloaded SDS-PAGE gel (Figure 1(d)) For the sake of an additional six histidine residues at the C terminus, it can be readily separated from the protein of interest by affinity chromatography after removal of the fusion tags. These advantages suggest this form of SARS-CoV Mpro could have important industrial applications.

Table 5. Comparison of the different proteases

<table>
<thead>
<tr>
<th>Proteases</th>
<th>Methods for production</th>
<th>Substrate specificity</th>
<th>Cleavage efficiency $k_{cat}/K_m$ (M$^{-1}$ s$^{-1}$)</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human thrombin</td>
<td>Extraction from plasma</td>
<td>-XLVPR</td>
<td>GSX-</td>
<td>94,600</td>
</tr>
<tr>
<td>Bovine Factor Xa</td>
<td>Extraction from plasma</td>
<td>-XIEGR</td>
<td>X-</td>
<td>39,000</td>
</tr>
<tr>
<td>Human rhinovirus 3C protease</td>
<td>Recombinant expression</td>
<td>-XLEVLQ</td>
<td>GPX-</td>
<td>920</td>
</tr>
<tr>
<td>TEV protease</td>
<td>Recombinant expression</td>
<td>-XENLYFQ</td>
<td>GSX-</td>
<td>2620</td>
</tr>
<tr>
<td>SARS-CoV Mpro</td>
<td>Recombinant expression</td>
<td>-XTSAVLQ</td>
<td>SGFRKX-</td>
<td>26,500</td>
</tr>
</tbody>
</table>
the autocleavage mechanism of this enzyme for three reasons: (1) autoprocessing will not produce additional residues at the N terminus; (2) SARS-CoV M\textsuperscript{pro} has been reported to have highly efficient expression in GST fusion systems; and (3) this method could be used to characterize the autocleavage efficiency of SARS-CoV in vitro. As for point (3), the affinity GST tag was a mimic of the transmembrane domain upstream of SARS-CoV M\textsuperscript{pro} in polyprotein 1a and 1ab, representing an autocleavage model of the M\textsuperscript{pro} in vitro. SDS-PAGE analysis shows that the GST tag was removed entirely from the N terminus through autoprocessing during expression of SARS-CoV M\textsuperscript{pro} in E. coli, exhibiting its high efficiency for autocleavage in vitro. In previous studies, it was hypothesized that the autocleavage of M\textsuperscript{pro} may occur in proximity to the membrane. Our data suggest that completion of autocleavage might be achieved in the cytoplasm, although the precise whereabouts of the assembly of replicase complex components remains to be identified. According to separate reports by Hsu and Lin,\textsuperscript{35,36} using autocleavage to remove tags such as thioredoxin can further support our hypothesis.

Rhinoviruses and coronaviruses belong to the Picornaviridae and the Coronaviridae, respectively. No previous report has shown that any CoV Mpro enzyme could efficiently process the substrate of picornavirus 3C proteases, or vice versa. In our study, it is interesting to observe that only substitution at the P2' site resulted in the conversion of the SARS-CoV main protease substrate into that of rhinovirus 3C protease.

In our study, we demonstrated that the first Ser1 residue at the N terminus of SARS-CoV M\textsuperscript{pro} is important for its activity and inhibitor binding. The critical interactions involved in stabilizing the substrate-binding pocket are two hydrogen bonds formed by the free amino group of the first residue of one protomer with two residues, Phe140 and Glu166, comprising the S1 subsite of its partner protomer within the dimer structure. The stabilizing effect of the free amino group is obvious. In our published GPLGS-WT structure, the S1 subsite of protomer B is partially collapsed without stabilization by the first residue of protomer A, suggesting that the S1 pocket would be less stable due to micro-environmental changes. During the crystallization step of the structure determination procedure, a large quantity of WT SARS-CoV M\textsuperscript{pro} crystals could be produced over one night, which could easily diffract to very high resolution under the conditions described previously. In contrast, it usually took five to ten days to produce a very limited amount of GPLGS-WT SARS-CoV M\textsuperscript{pro} or GS-WT SARS-CoV M\textsuperscript{pro} crystals suitable for diffraction, albeit with comparatively lower resolution. As a consequence, the WT SARS-CoV M\textsuperscript{pro} reported here will be more suitable for activity assays, inhibitor screening and crystallization. It should accelerate development of anti-coronavirus inhibitors through a structure-assisted approach to drug design.

### Materials and Methods

#### Construction of the pGSTM expression vector

The 654–965 region of the pGEX-6p-1 vector was cloned into the pMD18-T vector with two primers:

- forward, 5′-TTCCGAGATCCTTATATTGATAAA-3′
- reverse, 5′-GGATCCCTTCTCAAAACACTCCTGCAGACTGCACATGATCCGGATGGAGTGG-3′

The 33 nucleotides encoding the specific 11 amino acids TSVLQSQGFRK recognized by SARS-CoV M\textsuperscript{pro} were introduced by the reverse primer. The recombinant pMD18-T plasmid was double-digested with BstBI and BamHI. The segment of interest was recombined into the pGEX-6p-1 vector, which had also been double-digested by BstBI and BamHI, to construct the new protein expression vector pGSTM.

#### Cloning, expression and purification of calbindin D28k by the pGSTM protein expression system

### Expression and purification

The resulting recombinant plasmid was transformed into the E. coli strain BL21 (DE3). The cells were cultured in LB medium containing 0.1 mg ml\textsuperscript{-1} ampicillin. When the absorbance at 600 nm (A\textsubscript{600}) reached 0.6, IPTG was added to 0.5 mM and the cell culture was incubated at 16 °C for 10 h. After harvesting by centrifugation at 4600 g (Beckman JLA-10-5), the pellet was resuspended in PBS (140 mM NaCl, 2.7 mM KCl, 10 mM Na\textsubscript{2}HPO\textsubscript{4}, 1.8 mM KH\textsubscript{2}PO\textsubscript{4}, pH 7.3) and sonicated on ice. The lysate was centrifuged at 27,000 g (Beckman JA-25-50) for 30 minutes and the precipitate was discarded. The supernatant was loaded onto 2 ml GST-glutathione affinity columns (Pharmacia) equilibrated with PBS and washed with 30 column volumes of PBS. After that, 0.1 mg of SARS-CoV M\textsuperscript{pro} (GS-WT) was added to the column at 4 °C for 12 h to remove the GST tag. The protein of interest was collected and analyzed by SDS-PAGE.

#### Expression and purification of SARS-CoV M\textsuperscript{pro}

Expression and purification of SARS-CoV M\textsuperscript{pro} with different numbers of additional residues (5, 2, 0) at the N terminus is described below.

**SARS-CoV M\textsuperscript{pro} (with five additional residues, GPLGS, at the N terminus)**

Expression and purification of GPLGS-WT SARS-CoV M\textsuperscript{pro} has been reported.\textsuperscript{14} Briefly, the coding sequence of
the SARS-CoV M<sup>pro</sup> was inserted into the BamHI and Xhol sites of pGEX-6p-1 plasmid DNA (GE Healthcare). The resulting plasmid was used to transform <i>E. coli</i> BL21 (DE3) cells. The GST fusion protein, GST-SARS-CoV M<sup>pro</sup>, was purified by GST-glutathione affinity chromatography, cleaved with GST rhinovirus 3C protease, and the recombinant SARS-CoV M<sup>pro</sup> was further purified by anion-exchange chromatography.

**SARS-CoV M<sup>pro</sup> (with two additional residues, GS, at the N terminus)**

Expression and purification of GS-WT SARS-CoV M<sup>pro</sup> has been reported. Briefly, the coding sequence was inserted into the BamHI and Xhol sites of the pGEX-4T-1 vector (GE Healthcare). The following procedure is similar to that used for the expression and purification of SARS-CoV M<sup>pro</sup> in pGEX-6p-1 plasmid, except that the GST fusion protein was cleaved by thrombin.

**WT SARS-CoV M<sup>pro</sup> (without additional residues at the termini)**

The coding sequence for SARS-CoV M<sup>pro</sup> was amplified by polymerase chain reaction (PCR) using the PCR primers:

forward, 5′-CCGGATTCGCGTACTGCAGATGTTTCACTCAAAATGGCA3′

reverse 3′-CCGGCTCAAGTCTAGTGGTGTTGGTGTTGGTGAG-GGTGTAGCTGGAAAGTAACCT3′

The 12 nucleotides coding for the four amino acids AVLQ (corresponding to the P<sub>1</sub>–P<sub>4</sub> autocleavage sites at the N terminus) was inserted into the BamHI and Xhol sites of the pGEX-4T-1 vector (GE Healthcare). The following procedure is similar to that used for the expression and purification of SARS-CoV M<sup>pro</sup> with an authentic C terminus. The resulting plasmid was used to transform <i>E. coli</i> BL21 (DE3) cells. The GST fusion protein, GST-SARS-CoV M<sup>pro</sup>, was cleaved by thrombin, and concentrated WT SARS-CoV M<sup>pro</sup> (10 mg ml<sup>−1</sup>) was stored in 50 mM Tris–HCl (pH 7.3), 1 mM EDTA at −80 °C for enzyme activity assays and crystallization.

**Enzymatic activity and inhibition assays**

Enzymatic activity assays of GS-WT SARS-CoV M<sup>pro</sup> have been described. Activity assays of GPLGS-WT, WT-GPH<sub>6</sub> and WT SARS-CoV M<sup>pro</sup> followed a similar protocol. Briefly, the substrate of the N and C-terminal authentic SARS-CoV M<sup>pro</sup> was the fluorogenic compound MCA-AVLQSGFR-Lys(Dnp)-Lys-NH<sub>2</sub> (greater than 95% purity, GL Biochem Shanghai Ltd, Shanghai, China). The excitation and emission wavelengths of the fluorogenic substrate were 320 nm and 405 nm, respectively. A buffer consisting of 50 mM Tris–HCl (pH 7.3), 1 mM EDTA was used for enzyme activity assays at a temperature of 30 °C. The reaction was initiated by adding protease (final concentration of 0.2 μM for WT and WT-GPH<sub>6</sub>, 2 μM for GPLGS-WT) to a solution containing different final concentrations of the substrate (3.2–40 μM for WT and WT-GPH<sub>6</sub> and 6.4–80 μM for GPLGS-WT). The kinetic constants <i>K<sub>m</sub></i> and <i>K<sub>cat</sub></i> were obtained from a double-reciprocal plot (Supplementary Data Figure S1). Strict
kinetic parameters were determined for the inhibition assay.12

Crystallization, data collection and structure determination

Crystallization of WT SARS-CoV M\textsuperscript{pro} was carried out as described.14 The preparation of the co-crystals of SARS-CoV M\textsuperscript{pro} in complex with the inhibitor N3 has been reported.12 A set of WT SARS-CoV M\textsuperscript{pro} data was collected from a single crystal on beamline BL19-ID of the Advanced Photon Source (APS), Argonne National Lab at a wavelength of 1.00 Å. Data for the SARS-CoV M\textsuperscript{pro} complex were collected at 100 K in-house on a Lab at 40 kV and 20 mA (1.5418 Å) with a Rigaku R-AXIS IV++ image-plate detector. Data were processed, integrated, scaled and merged using HKL2000.15 The methods used for structure determination were as described.12 Briefly, the structures were determined by molecular replacement from our native structure of SARS-CoV Mpro (pH 7.6) (PDB ID 1UK3). Data collection and structure refinement statistics are summarized in Table 6.

Data Bank with accession numbers

Coordinates and structure factors for WT SARS-CoV M\textsuperscript{pro} have been deposited in the Protein Data Bank with accession number 2H2Z. Coordinates and structure factors for WT SARS-CoV M\textsuperscript{pro} in complex with the inhibitor N3 have been deposited in the Protein Data Bank with accession number 2HOB.

Acknowledgements

We thank Xuemei Li and Sheng Ye for technical assistance; Huanming Yang, Jian Wang, and Jun Yu for providing cDNA of SARS-CoV M\textsuperscript{pro}; Rongguang Zhang and Andrzej Joachimiak for data collection of WT SARS-CoV M\textsuperscript{pro}. This work was supported by Project 973 of the Ministry of Science and Technology of China (grant number 2004BA519A30), the NSFC (grant number 30221003), the Sino-German Center (grant number GZ236(202/9)), the Sino-European Project on SARS Diagnostics and Antivirals (SEPSDA) of the European Commission (grant number 003831) and the Tsinghua University PhD student innovation fund.

Supplementary Data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jmb.2006.11.073

References


*Edited by R. Huber*