Binding Site-based Classification of Coronaviral Papain-like Proteases

Traian Sulea,* Holger A. Lindner, Enrico O. Purisima, Robert Ménard*
Biotechnology Research Institute, National Research Council of Canada, Montreal, Quebec, Canada

ABSTRACT The coronavirus replicase gene encodes one or two papain-like proteases (termed PL1pro and PL2pro) implicated in the N-terminal processing of the replicase polyprotein and thus contributing to the formation of the viral replicase complex that mediates genome replication. Using consensus fold recognition with the 3D-JURY metapredictor followed by model building and refinement, we developed a structural model for the single PLpro present in the severe acute respiratory syndrome coronavirus (SCoV) genome, based on significant structural relationships to the catalytic core domain of HAUSP, a ubiquitin-specific protease (USP). By combining the SCoV PLpro model with comparative sequence analyses we show that all currently known coronaviral PLpros can be classified into two groups according to their binding site architectures. One group includes all PL2pros and some of the PL1pros, which are characterized by a restricted USP-like binding site. This group is designated the R-group. The remaining PL1pros from some of the coronaviruses form the other group, featuring a more open papain-like binding site, and is referred to as the O-group. This two-group, binding site-based classification is consistent with experimental data accumulated to date for the specificity of PLpro-mediated polyprotein processing and PLpro inhibition. It also provides an independent evaluation of the similarity-based annotation of PLpro-mediated cleavage sites, as well as a basis for comparison with previous groupings based on phylogenetic analyses. Proteins 2006;62:760–775.
© 2005 Government of Canada.

Key words: annotation; coronavirus; fold recognition; inhibition; processing site; ubiquitin-specific protease

INTRODUCTION

Coronaviruses are enveloped, single-stranded, positive-sense RNA viruses.[1] Besides economically important veterinary pathogens,[2] they include human coronaviruses (HCoVs), which are a cause of respiratory tract diseases, including the common cold, and occasional enteric infections.[3–8] The identification of a coronavirus as the infectious agent of severe acute respiratory syndrome (SARS), a life-threatening form of atypical pneumonia, has led to a renewed interest in coronaviruses.[9] Despite successful containment of the first SARS epidemic by quarantine measures, human SARS coronavirus (SCoV) infections

persist[10] without any specific therapy at hand.[9,11] Interferon treatment is currently regarded most useful,[11] whereas the broad-spectrum antiviral nucleoside analog ribavarin and the HIV protease inhibitor combination lopinavir/ritonavir proved ineffective.[12,13]

Upon cell infection, the viral replicase gene is translated directly from the viral genome.[14] Autocatalytic processing by two proteases, which are part of the replicase polyprotein, releases[14–16] nonstructural proteins (nsps).[15] These form a membrane-bound RNA replication complex.[14,16–18] One of the two coronaviral proteases, the 3CLpro, has already generated much interest as a target.[11] It resides in nsp5, and, after autocleavage, releases the downstream replicase subunits.[14] The processing of the amino-proximal nsps is carried out by one or two paralogous protease domains within nsp3, the largest of the nsps.[15,19–25] They are defined by homology to the papain-like fold[15] and constitute the peptidase family C16.[26] Mutational analyses support the presence of a Cys-His catalytic dyad.[15,22,25] Most coronaviruses harbor two such papain-like protease domains, PL1pro and PL2pro, whereas SCoV and the avian infectious bronchitis coronavirus (IBV) utilize only one, which is equivalent to PL2pro.[27] PL2pro may cleave down- and upstream of nsp3,[21,22] but only upstream cleavages were associated with PL1pro.[15,19,21,24] Additional nsp3 domains include the X domain, which is predicted to constitute a RNA processing enzyme,[27] and the hydrophobic Y domain, which likely anchors nsp3 to membranes.[21,28] The PLpro cleavage products nsp1-3 all colocalize with the replication complex.[14,16,17,28]

The synthesis of both negative- and plus-strand virus RNA require ongoing viral protein production,[29–31] and complete processing of the replicase N-terminal nsps appears to be essential for optimal virus growth.[32] The development of selective PLpro inhibitors[22] may, therefore, provide a new class of antivirals. However, little is known about the molecular basis of PLpro cleavage site sequence recognition, nor the significance for the existence

of two PLpro domains, which may or may not exhibit overlapping target site selectivities.[21,33] The cleavage site sequence specificity is limited to a preference for small residues (Gly, Ala) in the P_1 and P_2 positions for most but not all coronaviral PLpros.[20–23,25,27,32,34–37] For a structural analysis of PLpros, Herold and colleagues[39] built a homology model for the PL1pro domain of HCoV-229E based on the papain structure. The authors modeled an additional ~50-residue sequence, which connects the amino- and carboxy-terminal subdomains of the putative papain fold, as a Zn-ribbon. Indeed, the recombinant PL1pro domain binds equimolar amounts of zinc, and mutation of the predicted zinc-binding motif abolished catalytic activity.[39] Recently, we have identified a structural relationship[40] between the SCoV PLpro and the catalytic core domain of the papain-like herpesvirus-associated ubiquitin-specific protease (HAUSP), also known as USP7, of the C19 family of ubiquitin-specific proteases (USPs).[26] Instead of a classical Zn-ribbon, as proposed for the PLpros by Herold and coworkers,[39] HAUSP contains a circularly permuted Zn-ribbon-like domain inserted between the two subdomains of the papain fold.[41] We further recognized[40] that the binding site complementarity of HAUSP to the C-terminal ubiquitin sequence LRGG matches the narrow specificity profile (LXGG) of SCoV PLpro.[22,25,27]

In this study, we survey in detail the substrate interactions predicted in the binding site of SCoV PLpro, particularly in the S_1 and S_2 subsites. The structural framework provided by the modeled SCoV PLpro-binding site is then combined with comparative sequence analyses in order to understand specificity data available for other coronaviral PLpros. Despite what their names seem to imply, PL1pro and PL2pro do not represent distinctive subgroups of the coronaviral papain-like enzymes. Indeed, it has not been possible so far to cluster the PL1pro and PL2pro domains into specific groups based on clear functional comparisons. Our analysis reveals a novel classification of all currently known coronaviral PLpros, which is based on their binding-site characteristics. This classification is further used for an independent evaluation of the current annotation of coronaviral PLpro cleavage sites from public databases.

MATERIALS AND METHODS
Coronavirus Nomenclature and Sequence Accession Numbers

Coronavirus abbreviations, together with SwissProt (SW; http://www.expasy.org/sprot) or GenBank (GB; http://www.ncbi.nlm.nih.gov/entrez) accession numbers used in this article are as follows: SCoV for SARS coronavirus (strain Tor2; SW: P59641; GB: NC_004718), HCoV for human coronavirus strains 229E (SW: Q05002; GB: NC_002645), NL (GB: AY518894), OC43 (GB: NC_005147) and HKU1 (GB: AY597011), BCoV for bovine coronavirus (strain Ent; SW: Q91A29; GB: NC_003045), MHV for murine hepatitis virus (strain A59; SW: P16342; GB: NC_001846), TGEV for transmissible gastroenteritis virus (SW: Q9IW06; GB: NC_002306), PEDV for porcine epidemic diarrhea virus (SW: Q91AV2; GB: NC_003436), and

IBV for infectious bronchitis virus (strain Beaudette; SW: P27920; GB: NC_001451). Other strains for SCoV, BCoV, MHV, and IBV were omitted from the analysis in order to decrease redundancy in the datasets of sequences for the PLpros and their respective predicted cleavage sites.

Structural Bioinformatics

Fold detection was carried out at the Structure Prediction Meta Server (http://bioinfo.pl/meta).[42] Consensus sequence-to-structure scoring was achieved with the 3D-JURY method running in the best-model-scoring mode over the default set of eight threading servers, as well as over all prediction servers available including other meta-predictors.[43,44] The reported top-ranked query-to-template sequence alignments were further refined manually by considering (1) the structure-based sequence alignment of the identified templates, (2) the sequence alignment of the coronaviral PLpro family generated with the CLUSTAL W program,[45] and (3) the secondary structure alignment. Secondary structure prediction was obtained with three methods: PROFsec,[46] PSI-PRED,[47] and SAM-T99,[48] and then applying a consensus by majority voting.[49] The final sequence-to-structure alignment of SCoV and other coronaviral PLpros to the identified template structures is given in Figure 1, including predicted and experimental secondary structure elements. This alignment formed the basis for the 3D homology modeling of the SCoV PLpro structure.

Building and Refinement of the SCoV PLpro Model

We have previously reported a short outline for the construction and refinement of the SCoV PLpro homology model (residues K1632–E1847).[40] In brief, the SCoV PLpro model, comprising the Zn-ribbon domain inserted in the middle of the protease domain, was built as a chimera of the two template structures, HAUSP and foot-and-mouth disease virus leader protease (FMDV Lpro), identified by the 3D-JURY fold recognition. Detailed procedures and atomic coordinates of the final model complexed with full-length ubiquitin aldehyde (Ubal) are given in this report.

Structural manipulations were performed with the SYBYL 6.6 molecular modeling software (Tripos, Inc., St. Louis, MO). First, the homology modeling program COMPOSER[50] in SYBYL was employed in order to fit various regions of the SCoV PLpro sequence onto the 2.3 Å-resolution crystal structure of the core catalytic domain of HAUSP complexed with ubiquitin aldehyde (PDB code 1NBF),[51] and onto the 1.9 Å-resolution crystal structure of FMDV Lpro (1QMY),[52] following the sequence alignment shown in Figure 1. Based on sequence similarities, deletions and/or insertions, and the disposition of secondary structure elements, the following sequence-to-template assignment was adopted: (a) the SCoV PLpro segments K1632–E1701, F1798–E1803, and T1814–P1839 were taken from FMDV Lpro segments E30–E96, F137–F142, and V150–D176, respectively, largely covering the N- and C-terminal subdomains; (b) the SCoV PLpro segments L1702–T1797, Y1804–Y1813, and V1840–E1847 were
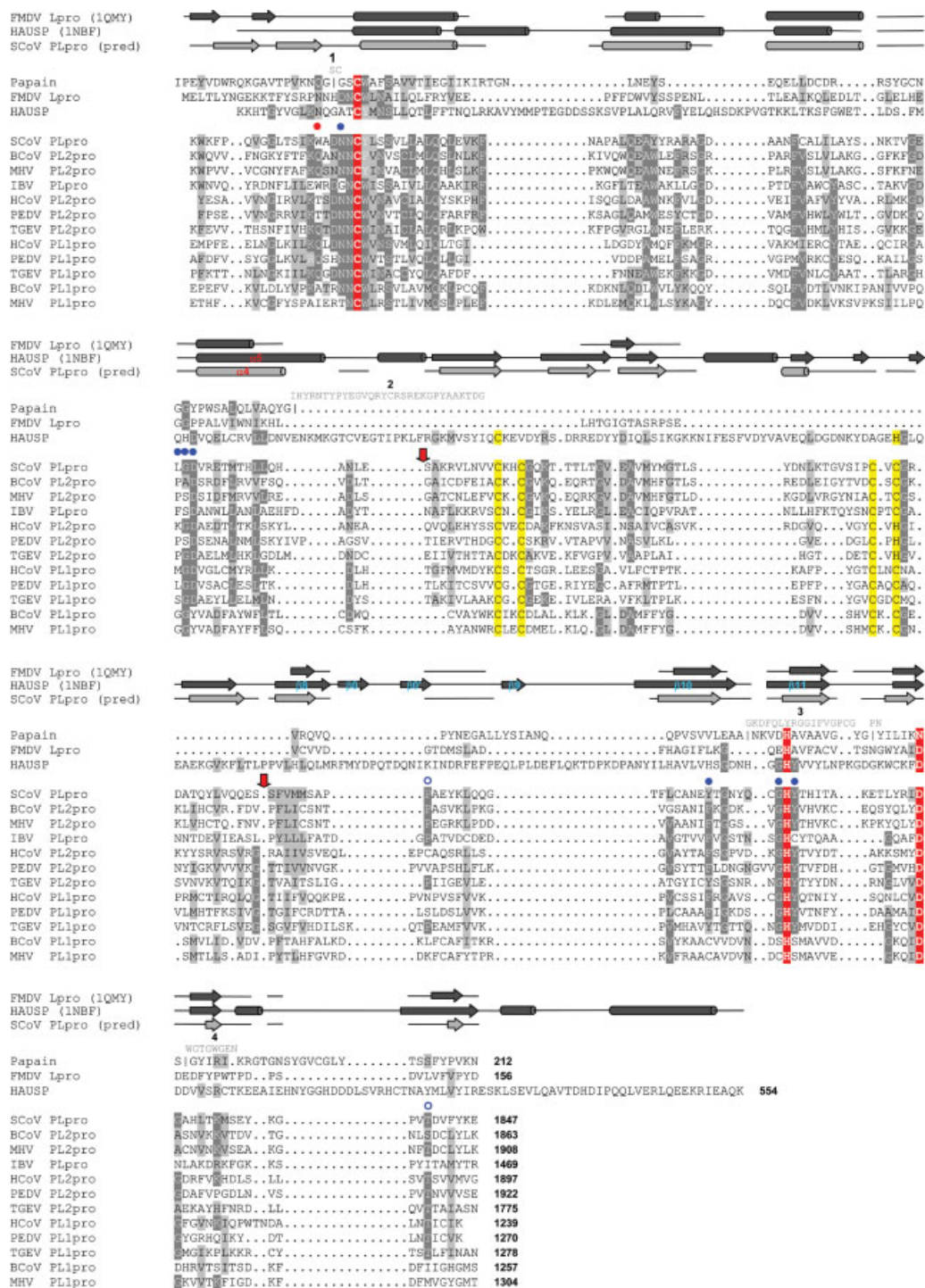
Fig. 1. Multiple sequence alignment of coronaviral PLpros to a structure-based sequence alignment of HAUSP, papain, and FMDV Lpro. Predicted secondary structure elements for SCoV PLpro are shown in gray, and the actual secondary structure elements for HAUSP and FMDV Lpro (PDB codes given in the parentheses) are shown in black above the alignment. β-strands are represented by arrows, α-helices by cylinders, and coils by lines. Selected secondary structure elements referred to in the text are labeled. Active site catalytic triad residues are shown on red background. Putative Zn-chelating Cys residues in the Zn-ribbon domain are highlighted on yellow background, as are the two reminiscent Zn-chelating residues in HAUSP. Boundaries of the Zn-ribbon domain are indicated by vertical red arrows. The position of the putative oxyanion-stabilizing residue is indicated with a red dot, and those predicted to engage in interactions at substrate positions P_1-through-P_4 (see Fig. 3) are indicated with blue dots, except for P1788 and T1841 of SCoV PLpro, which are indicated with blue circles. Papain insertions in the alignment are shown above its sequence, and those labeled 1 through 4 correspond with those indicated in Figure 2 on the papain structure. Residues identical in half or more of the coronaviral PLpro sequences are in white on dark gray background and those conserved in half or more of the cornaviral PLpro sequences are on light gray background, based on the BOXSHADE program (http://www.ch.embnet.org). The conservation highlighting is carried over to the sequences of HAUSP, FMDV Lpro and papain. HCoV refers to HCoV-229E.

taken from HAUSP segments Q293–E429, H456–Y465, and A513–R520, respectively. In all, these elements in HAUSP form the substrate-binding loop α4-α5 and part of the helix α5 in the N-terminal subdomain, the finger domain, the substrate-covering loop immediately preceding the catalytic histidine, and the two β-strands from the C-terminal subdomain adjacent to the finger domain. Loops in the SCoV PLpro, corresponding to insertions/ deletions or junctions relative to the templates, were constructed by searching protein structures from the Protein Data Bank (PDB; http://www.rcsb.org) using the PROTEIN LOOPS program in SYBYL. They include the following sequences: P1636–Q1637, E1664–A1671, and R1680–D1682 (in the N-terminal subdomain); A1716– E1719 in the region connecting the N-terminal subdomain to the finger domain of HAUSP; Y1747–L1751 (corresponding to the finger domain of HAUSP); and P1788–A1789, G1796–F1798, and K1819–E1820 (in the C-terminal subdomain). For selecting loop conformations, the search output was examined for root-mean-square (rms) deviations at the anchor positions, sequence homology, as well as suitability for the overall tertiary structure.

Using the superimposed HAUSP template structure with bound Ubal, the C-terminal portion of the Ubal (RLRG-Glycinal) was docked in the SCoV PLpro-binding site as a thiohemiacetal adduct covalently bound to the catalytic cysteine (C1651). This ligand also mimics the SCoV PLpro cleavage site sequence motif LXGG (positions P_4 to P_1).[22,25,27] The N- and C-termini of protein and ligand were blocked with acetyl and methylamino groups, respectively. Several SCoV PLpro side chains were manually repacked to improve van der Waals contacts and hydrogen bonding. Hydrogen atoms were added explicitly, and the polar hydrogens were oriented to favor hydrogen bonding. The ionization state at physiological pH was adopted. The catalytic histidine was treated as neutral due to the covalent adduct formation at the catalytic cysteine. Accordingly, a hydroxyl group was considered instead of an oxyanion in the thiohemiacetal group. Given the importance of the putative Zn-chelating cysteines for the *trans*-cleavage activity of HCoV,[39] we also carried out initial docking and coordination of a Zn ion to SCoV PLpro based on structural superimpositions with two representative C4-type Zn-ribbons from the transcription elongation factor SII (PDB code 1TFI) and RNA polymerase II subunit 9 (1QYP), and with the circularly permuted C4-type Zn-ribbon of the silent information regulator 2, Sir2 (1ICI).

Model refinement was carried out by gradual structural relaxation using a stepwise energy minimization protocol and employing the AMBER all-atom molecular mechanics force field.[53] (More details on the energy refinement procedure and the docking of ubiquitin to SCoV PLPro can be found in the Supplementary Material.) In terms of the basic stereochemical quality of the refined model, 95% of the nonglycine residues of SCoV PLpro reside in the most favored (75%) and allowed (20%) regions of the Ramachandran plot, and only one non-glycine residue (E1820) is found in the disallowed region. The refined structure

preserves the number and general disposition of predicted secondary structure elements.

RESULTS AND DISCUSSION
Structural Relationships of SCoV PLpro

We have recently mined the PDB content for further structure-to-function annotation of the coronaviral PLpros.[40] The structure of the catalytic core domain of HAUSP[51] was scored by 3D-JURY well above the significance threshold of 50, which is considered to result in a prediction accuracy of above 90%.[43] Simple application of standard homology tools (e.g., PSI-BLAST) failed to detect any statistically significant relationship between SCoV PLpro and any of the known protein structures. The structure of FMDV Lpro,[52,54] was ranked second by 3D-JURY, albeit with a borderline significant score. The structure of HAUSP and FMDV Lpro each feature a papain-like domain, with an additional circularly permuted Zn-ribbon domain inserted between the two subdomains of the papain fold only in HAUSP. 3D-JURY scored only the FMDV Lpro structure above the significance threshold when the protease domain of SCoV PLpro alone was queried (i.e., after excision of the sequence S1720– S1779).

As already mentioned, this additional inserted domain was previously proposed to adopt a Zn-ribbon fold.[39] Our sequence alignment (Fig. 1) only detects a cysteine residue in the first of the four putative Zn-chelating positions in the HAUSP sequence. However, when we extended this comparison to related USPs, it became clear that all four positions are occupied by cysteine residues in ~68% of the 251 members of the C19 family as aligned in the MEROPS database.[26] We recognized that in the context of the HAUSP finger domain structure, these residues would form the Zn-binding motif of a circularly permuted Zn-ribbon.[40] Independently, a structural relationship between the finger domain of HAUSP and the circularly permuted C4-type Zn-ribbon has recently been recognized by Krishna and Grishin.[41]

Although no statistically significant scores were reported by 3D-JURY for the putative Zn-ribbon domain of SCoV PLpro alone (sequence S1720–S1779), all the top-ranked structures represented rubredoxins (e.g., PDB codes 1S24, 1SMM, 1BQ8), which, as HAUSP, feature a circularly permuted Zn-ribbon domain. Together with Zn-β-ribbons, they belong to the rubredoxin-like fold family according to the SCOP database.[55] Members of this family contain two $C(X)_{1,2}C$ motifs that typically coordinate Fe^{2+}/Fe^{3+} in rubredoxins and Zn^{2+} in Zn-ribbons. A more in-depth discussion of the circularly permuted Zn-ribbon is given in Appendix A.

Overall modeled structure of SCoV PLpro

A 3D model of SCoV PLpro (K1632–E1847) was constructed as a chimera between the HAUSP and FMDV Lpro template structures. Figure 2 compares the refined model of SCoV PLpro with the crystal structures of HAUSP, FMDV Lpro and papain. Relative to the SCoV PLpro protease domain (i.e., excluding the Zn-ribbon do-
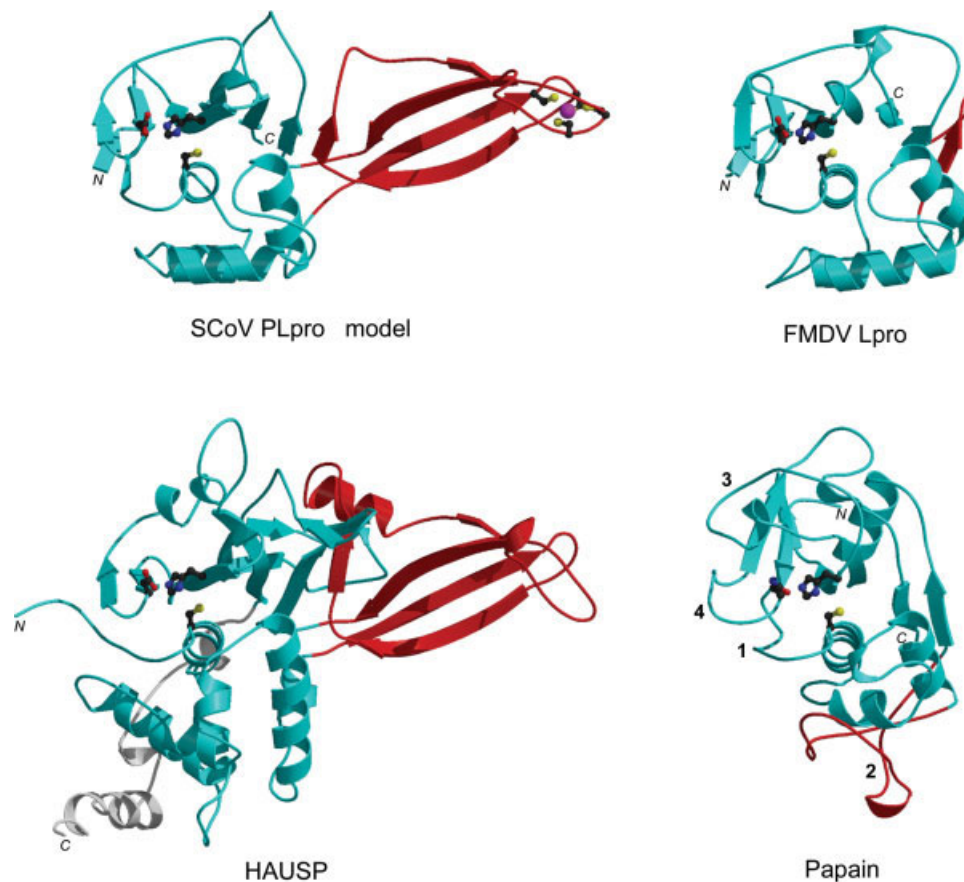
Fig. 2. Comparison of the modeled structure of the catalytic core domain of SCoV PLpro with the crystal structures of the catalytic core domain of HAUSP (PDB code 1NBF), FMDV Lpro (1QYM), and papain (1PPN). The protease domains are colored in cyan, the insertions in the middle of the protease domain (residues S1720–S1779 in SCoV PLpro, R325–P399 in HAUSP, T113–E123 in FMDV Lpro, and G79–G109 in papain) are in red, and the C-terminal extension of HAUSP (S552–K554) is rendered in white. Catalytic triads are shown in ball-and-stick representation. The four cysteine residues that coordinate the Zn ion (magenta sphere) in the SCoV PLpro model are also shown. Structural differences in papain relative to the other three enzymes (see also Fig. 1) are numbered 1 to 4: (1) the sequence preceding the catalytic cysteine and harboring the oxyanion hole residue; (2) the insertion between the N- and C-terminal subdomains of the protease domain that folds back onto the N-terminal subdomain rather than extending the β-sheet of the C-terminal subdomain as in the other structures; (3) a long loop folded onto the C-terminal subdomain and replacing the shorter, substrate-covering loop in the other structures; (4) a Trp-containing eight-residue loop inserted after the asparagine of its catalytic triad and shielding it from solvent (while the corresponding aspartate in the other structures is solvent exposed).

main), the larger protease domain of HAUSP has two additional α-helices in the N-terminal subdomain and three additional β-strands in the C-terminal subdomain, together with longer intervening loops. In fact, the smaller FMDV Lpro is a more suitable template for most of the SCoV PLpro protease domain, because of its similar size and an exact match of secondary structure elements. However, the residues predicted to shape the substrate-binding subsites S_1 through S_4 in SCoV PLpro (described in more detail in the following section) clearly resemble the HAUSP-binding site that accommodates the ubiquitin C-terminal sequence LRGG.[51] Among the several sizable differences, which led to the prediction of a less-elaborated structure of SCoV PLpro compared to HAUSP, we noted a shorter loop after the first β-strand of the C-terminal subdomain in the former protease. The corresponding loop

in HAUSP (β8-β9) becomes ordered as a β-hairpin (β0-β0′) upon ubiquitin binding, presumably, because of its contacts with the ubiquitin residues in the P_4 through P_6 positions.[51] The β10-β11 hairpin loop of HAUSP, however, which also covers the ubiquitin C-terminal residues, appears conserved in SCoV PLpro, but is three residues shorter in FMDV Lpro (see also Fig. 1). Figures 1 and 2 further highlight significant differences between the SCoV PLpro model and papain structure (see Fig. 2 for details).

The presence of a Zn-ribbon domain in SCoV PLpro is compatible with the existence of a circularly permuted Zn-ribbon domain in HAUSP,[56] in terms of its size, sequence location, and predicted secondary structure. As in the HAUSP template, the Zn-ribbon domain of SCoV PLpro extends the β-sheet in the C-terminal subdomain of the protease domain by a parallel β-strand, which serves

to anchor the orientation of the Zn-ribbon domain relative to the protease domain. Further interdomain contacts established in HAUSP between an additional α-helix (α7) in the Zn-ribbon domain and a longer loop β9-β10 in the protease domain, are absent in our model of SCoV PLpro. In FMDV Lpro, the inserted Zn-ribbon domain is reduced to just one β-strand that preserves the parallel interaction with the β-sheet of the protease domain. Further discussion on the predicted crossover loop conformation of the SCoV PLpro circularly permuted Zn-ribbon domain, and its implications for interdomain orientation, is given in Appendix A.

Predicted Substrate-binding Site of SCoV PLpro

The structure of the catalytic core domain of HAUSP, in complex with Ubal,[51] is a suitable template for reliable modeling of the substrate-binding cleft of SCoV PLpro. In order to allow a detailed view of specific enzyme-substrate interactions in the nonprimed side of the binding groove, structural refinement of SCoV PLpro was carried out in the presence of RLRG-Glycinal bound covalently to the catalytic cysteine as a thiohemiacetal adduct and interacting with subsites S_5 through S_1. As we have pointed out previously,[40] this peptidyl aldehyde not only corresponds to the Ubal C-terminal sequence, but also matches the general P_4-P_1 specificity motif of SCoV PLpro, LXGG, derived from the predicted PLpro-processing sites of the polyprotein.[22,25,27] The details of the substrate interactions in subsites S_4 to S_1 are shown in Figure 3.

In the P_1 substrate position, the Glycinal moiety is covalently bound to the catalytic residue C1651, which together with H1812 and D1826 forms the putative catalytic triad in a canonical spatial arrangement. The tetrahedral hemiacetal oxygen atom is stabilized by three hydrogen bonds, namely, with the indole NH group of the oxyanion hole residue W1646, the main chain NH group of C1651, and the side chain amide group of N1649. Six of the seven main-chain heteroatoms of the substrate P_1 to P_4 positions are engaged in direct intermolecular hydrogen bonds with enzyme residues G1811 (one H-bond to P_1 backbone), G1703 (two H-bonds to P_2 backbone), Y1804 (one H-bond to P_3 backbone), and D1704 and Y1813 (two H-bonds to P_4 backbone). Such an extensive hydrogen-bonding network indicates not only high levels of complementarity in the recognition of the substrate main chain, but also that the substrate can achieve substantial binding affinity without additional interactions through its side chains.

Furthermore, the side chains of residues N1649 and L1702 restrict the S_1 pocket to hinder the accommodation of large P_1 side chains. In the S_2 subsite, the side chains of residues Y1813 and Y1804 completely occlude the S2 pocket and clearly prevent binding of P_2 side chains larger than Ala. As mentioned earlier, these two Tyr side chains are also involved in the anchoring of the substrate main chain at the P_4 and P_3 positions, respectively. In addition, Y1813 and Y1804 side chains are conformationally restricted, particularly, the more buried Y1813 adjacent to the catalytic H1812 residue. The available space around

the P_2 main chain is also reduced by the β-hairpin loop between Y1804 and Y1813. Closure of the loop on the substrate main chain also brings it in contact with the L1702 side chain, effectively creating a narrow tunnel into which the P_1-P_2 di-glycine can fit snugly [Fig. 3(c)]. From a structural viewpoint, the overall importance in determining the strict P_2 specificity appears to be Y1813 > Y1804 > Y1804–Y1813 loop. The model clearly explains the observed S_1 and S_2 specificities of SCoV PLpro for glycine residues.[22,25]

The Arg side chain modeled at the P_3 substrate position is largely solvent-accessible, which is in agreement with the consensus processing site sequence for SCoV PLpro containing a variable P_3 residue.[22,25] The only specific interaction of the P_3 Arg side chain is a long hydrogen bond (not shown) between its guanidinium group and the substrate-covering loop Y1804–Y1813 of the enzyme. Leu is conserved at the P_4 position of the three polyprotein-processing sites by SCoV PLpro. The modeled P_4 Leu side chain binds in a relatively defined pocket of the enzyme, where it contacts the side chains of residues Y1804, as well as P1788 and T1841. Low levels of target-template sequence conservation (see Fig. 1) decrease the prediction reliability for the contacts with the latter two side chains. The P_5 Arg side chain was readily modeled in a salt-bridge interaction with the E1707 carboxylate group (not shown). Because of its surface exposure, it is not expected that this electrostatic interaction would play a major role in substrate affinity and specificity. Accordingly, different P_5 residues are found in the putative SCoV PLpro cleavage site sequences.

The HAUSP-like topology of the SCoV PLpro-binding site differs significantly from that of papain. In papain, SCoV PLpro residues D1704, Y1804, and Y1813 are replaced with residues Y67, V133, and A160, respectively. This precludes hydrogen-bond formation between papain and the substrate main chain in the P_3 and P_4 positions, as outlined above for SCoV PLpro. Importantly, substitutions of the S_2-occluding residues Y1804 and Y1813 of SCoV PLpro result in a well-shaped substrate-accessible S_2 pocket in papain, suitable for the accommodation of bulky hydrophobic P_2 side chains, such as Leu or Phe.[57] Instead of SCoV PLpro residues N1649 and L1702, which sterically block its S_1 pocket, glycine residues are found at the corresponding positions in papain (Gly23, Gly65) and related cathepsins, which tolerate a variety of P_1 side chains in the open S_1 subsite. Mutation of any of these two glycine residues in cathepsin B to the corresponding non-glycine residues at these positions in papaya proteinase IV, which only accepts Gly at the substrate P_1 position, has been shown to restrict the P_1 specificity of cathepsin B to glycine.[58] The β-hairpin loop Y1804–Y1813 of SCoV PLpro is replaced in papain by a long insertion (labeled 3 in Figs. 1 and 2) that folds against the C-terminal subdomain of the protease. Also different from SCoV PLpro, papain does not have a defined S_4 subsite, in agreement with its broad specificity at the substrate P_4 position.
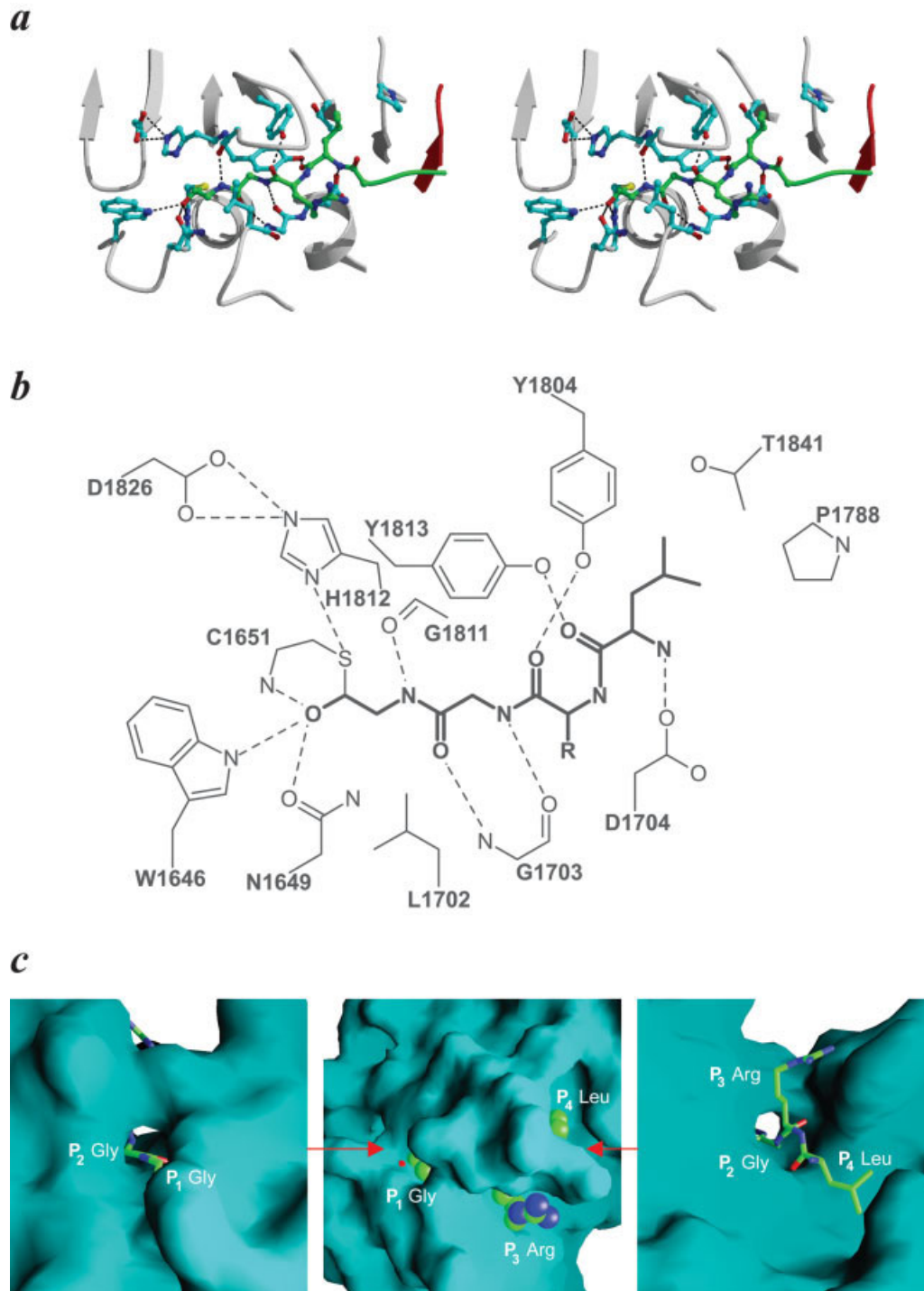
Fig. 3. Substrate recognition in the subsites S_1 through S_4 of SCoV PLpro. (**a**) Stereo view of the modeled interactions between the LRG-Glycinal peptidyl aldehyde and the substrate-binding site of SCoV PLpro. Carbon atoms are colored in cyan for the protease and in green for the ligand. Hydrogen bonds are indicated with dashed lines. The color scheme applied for rendering the protein chains is as in Figure 2, except for the protease domain shown here in white. (**b**) Schematic representation of the interactions show in (**a**). Protein residues are shown with thin lines, the ligand is shown with thick lines, and hydrogen bonds are shown with dashed lines. (**c**) Steric fit of LRG-Glycinal in the substrate-binding site of SCoV PLpro. The protease is represented by its molecular surface, and the ligand is shown as a CPK model in the middle panel and with sticks in the two side panels. The view in the central panel is similar to the orientation shown in (**a**). The left and right panels are opposite side views as indicated by the red arrows, through the narrow tunnel in the S_1-S_2 region.
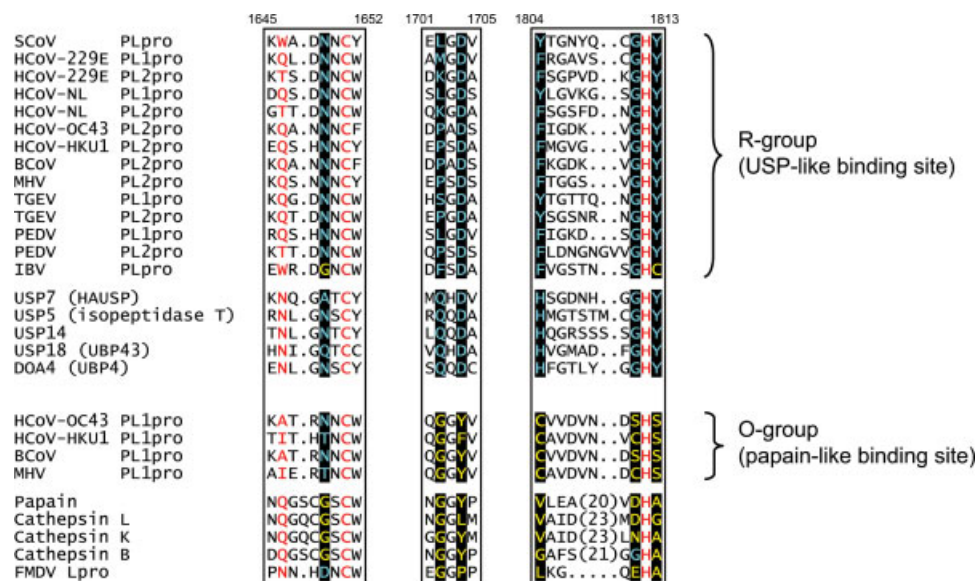
Fig. 4. Binding site-based classification of coronaviral PLpros. Sequence alignment of binding site signature motifs of coronaviral PLpros predicted from the SCoV PLpro model (cf. Fig. 3). Key substrate-binding residues are highlighted on black background. The USP-like binding site signature residues are in cyan, and the papain-like binding site signature residues are in yellow. The conserved catalytic Cys and His residues, as well as the variable residue at the putative oxyanion hole position, are in red. In the grouping of coronaviral PLpros, R indicates a restricted, USP-like binding site, and O is used to indicate an open, papain-like binding site according to the nature of, primarily, the S_2 subsite. As reference, the corresponding blocks of representative USPs are aligned bellow the R-group coronaviral PLpros and of representative papain-like proteases and FMDV Lpro below the O-group. Insertions in the papain-like enzymes are indicated by the length of their sequences. Entries shown are from the SwissProt database: USP7 (Human; SW: Q93009); USP5 (Human; SW: P45974); USP14 (Human; SW: P54578); USP18 (Mouse; SW: Q9WTV6); DOA4 (*Saccharomyces cerevisiae*; SW: P32571); Papain (*Carica papaya*; SW: P00784); Cathepsin L (Human; SW: P07711); Cathepsin K (Human; SW: P43235); Cathepsin B (Human; SW: P07858); and FMDV Lpro (Foot-and-mouth disease virus strain O1; SW: P03305). See Materials and Methods for coronavirus nomenclature and sequence accession numbers.

Comparative Analysis and Classification of Coronaviral PLpros

The modeled architecture and interactions in the non-primed side of the SCoV PLpro substrate-binding cleft, combined with the multiple sequence alignment presented in Figure 1, provide a structural framework for comparative analysis and classification of the other currently known coronaviral PLpros. The resulting binding site signature motifs, which characterize the entire coronaviral PLpro family, are delineated in Figure 4. SCoV PLpro residue numbering will be used in the following comparisons.

One group of coronaviral PLpros is characterized by a HAUSP-like binding site and includes, besides SCoV PLpro, the PL2pros from HCoV-229E, HCoV-NL, HCoV-OC43, HCoV-HKU1, BCoV, MHV, TGEV, and PEDV and the PL1pros from HCoV-229E, HCoV-NL, TGEV, and PEDV. In the S_1 subsite of these enzymes (cf. Fig. 3), N1649 is absolutely conserved, and L1702 is a non-Gly residue; in the S_2 subsite, Y1813 is absolutely conserved, and Y1804 is conservatively substituted by Phe in some of the homologs. The occluded S_1 and S_2 subsites of all these enzymes are suitable for recognition of P_1-P_2 di-glycine and appear to hinder accommodation of P_1 and P_2 side chains larger than Ala. We expect the binding mode of the substrate P_1-P_4 main chain to these coronoviral PLpros to be also similar, because of conservation of the hydrogen-bonding residues G1811, D1704, and Y1813, and conservative substitutions of residues G1703 and Y1804. Owing to the restricted nature of the S_1 and S_2 subsites, we term this group of coronaviral PLpros the R-group. Overall, the binding site signature for the R-group of coronaviral PLpro appears to be remarkably similar to that characteristic for USPs.[59,60]

The coronaviral PL1pros from HCoV-OC43, HCoV-HKU1, BCoV, and MHV share a papain-like binding site that is clearly distinct from that predicted for SCoV PLpro and form a second group. One major difference from the R-group of coronaviral PLpros is seen in the putative S_2 subsites of these enzymes. Here, Y1813 and Y1804 are replaced by smaller residues, namely, Ser and Cys, respectively. As in papain and related cathepsins, this opens the S_2 pocket for the recognition of bulkier P_2 side chains (Fig. 5). Together with the replacement of D1704 by Tyr (another papain-like substitution), this also eliminates three hydrogen bonds to the substrate P_3-P_4 main chain as modeled for SCoV PLpro. Replacement of G1811 and G1806, which are both conserved in the R-group coronaviral PLpros, with larger residues may affect the conformation and flexibility of the substrate-covering loop (loop
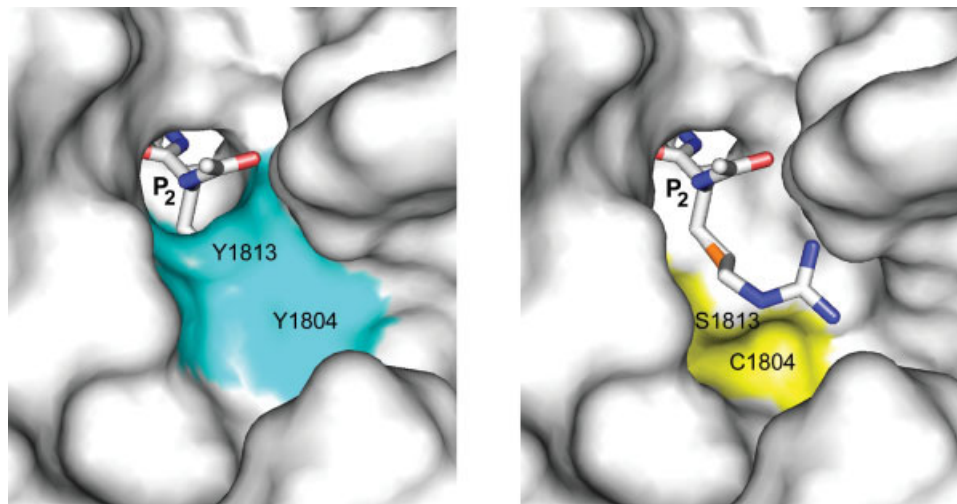
Fig. 5.   Opening of the S_2 subsite due to the replacement of the bulky SCoV PLpro residues Y1804 and Y1813 (conserved in the R-group coronaviral PLpros) to the smaller residues Cys and Ser found at these positions, respectively, in the O-group coronaviral PLpros. The panel on the left shows the molecular surface of SCoV PLpro with the surface patches associated with the side-chain atoms of residues Y1804 and Y1813 colored in cyan. The panel on the right shows the molecular surface of the SCoV PLpro double mutant Y1804C,Y1813S with the surface patches belonging to the mutated side-chains colored in yellow. Two substrate P_2 side chains, Cys and Arg, characteristic of the nsp1-nsp2 and nsp2-nsp3 processing sites sequences, respectively, by the O-group enzymes, are shown penetrating the S_2 molecular surface of the R-group SCoV PLpro (left panel), but being accommodated well in the spacious S_2 pocket of the O-group-like double mutant (right panel).

Y1804–Y1813, SCoV PLpro numbering). Interestingly, changes in the size of the S_2 pocket also impact the relative location of other subsites: the S_4 subsite of R-group coronaviral PLpros effectively forms the base of the S_2 subsite in the O-group. For example, residues encompassing positions T1841 and P1788, which putatively contribute to the P_4 recognition in SCoV PLpro, might actually impact P_2 recognition in MHV PL1pro. The extent of the steric hindrance at the S_1 subsite in the SCoV PLpro model yet differs from papain. Although the papain-characteristic Gly replaces the bulkier L1702, a non-Gly residue is still present at the N1649 position, which may nevertheless suffice in blocking accommodation of large P_1 side chains, as shown by mutation of the corresponding Gly27 in cathepsin B.[58] Owing to the open nature primarily at the S_2 subsite but also at the S_1 subsite, we termed the second group of coronaviral PLpros the O-group. The presence of hydrophobic residues at the putative oxyanion hole position is another interesting feature of O-group coronaviral PLpros, contrasting with the hydrogen-bond-capable oxyanion-stabilizing residues found in the R-group (Gln, Trp, or Thr), as well as in HAUSP and other USPs, FMDV Lpro, papain, and related cathepsins (Asn or Gln).

Although the IBV PLpro-binding site does not fit perfectly into the above bipartite classification, it appears more related to the R-group of coronaviral PLpros. At the S_1 subsite, the removal of the N1649 side chain through replacement by Gly does not generate a more accessible S_1 pocket because a bulkier Phe, in turn, replaces L1702. Similarly, although the S_2 pocket may become more spacious because of the replacement of Y1813 with Cys, the

conservative substitution of Y1804 for Phe is expected to still prevent the recognition of large P_2 side chains. Additionally, conservation of the SCoV PLpro residues D1704, G1811, and G1806 suggests similarities in the binding mode of the substrate main chain between IBV PLpro and the R-group of coronaviral PLpros.

Consistency of Coronaviral PLpro Classification with Experimental Data

After the demonstration that SCoV PLpro cleaves at the nsp2-nsp3 boundary by Thiel and colleagues,[22] Baker and coworkers[25] have recently demonstrated that SCoV PLpro mediates cleavages at all three putative SCoV PLpro processing sites. These occurred most likely at the highly conserved P_4 to P_1 motif LXGG, consistent with earlier predictions.[27] Baker and coworkers have also demonstrated different P_2 specificities for MHV PL1pro and PL2pro using extensive cleavage site-directed mutagenesis of the polyprotein. For MHV PL1pro, these studies revealed a stringent requirement for Gly in P_1 and a preference for Arg at the P_2 position, where several substitutions, including Gly, precluded PL1pro cleavage.[34–36] In contrast, the presence of Gly at both P_1 and P_2 is critical for recognition and processing of the nsp3-nsp4 cleavage site by MHV PL2pro.[23] Liu and colleagues[20,38] investigated the specificity of IBV PLpro by site-directed mutagenesis at the p41 and p87 cleavage sites, which are equivalent to the nsp3-nsp4 and nsp2-nsp3 sites, respectively, of the other coronavirus replicase polyproteins.[21] These two highly conserved cleavage sites feature Lys, Ala, and Gly at P_3, P_2, and P_1, respectively. A Gly is also

```
                nspX ↓ nspY      P 654321 ↓ 12 P'

SCoV            nsp1-nsp2        RELNGG180 -AV      PLpro
                nsp2-nsp3        FRLKGG818 -AP      PLpro
                nsp3-nsp4        ISLKGG2740-KI      PLpro
HCoV-229E       nsp1-nsp2        GKRGGG111 -NV      PL1pro / PL2pro
                nsp2-nsp3        FTKAAG897 -GK      PL1pro / PL2pro
                nsp3-nsp4        VAKQGA2484-GD      PL1pro / PL2pro
HCoV-NL         nsp1-nsp2        FGHGAG110 -SV      PL1pro / PL2pro
                nsp2-nsp3        FTKLAG898 -GK      PL1pro / PL2pro
                nsp3-nsp4        VAKQGA2457-GF      PL1pro / PL2pro
HCoV-OC43       nsp3-nsp4        FSLKGG2750-AV      PL2pro
HCoV-HKU1       nsp3-nsp4        FSLKGG2838-VV      PL2pro
BCoV            nsp3-nsp4        FSLKGG2750-AV      PL2pro
MHV             nsp3-nsp4        FSLKGG2837-AV      PL2pro
TGEV            nsp1-nsp2ᵃ       RTGRGA111 -IY      PL1pro / PL2pro
                nsp2-nsp3        YNKMGG879 -GD      PL1pro / PL2pro
                nsp3-nsp4ᵇ       SPKSGS2389-GF      PL1pro / PL2pro
PEDV            nsp1-nsp2        FGRRGG110 -NI      PL1pro / PL2pro
                nsp2-nsp3        FKKKGG895 -GD      PL1pro / PL2pro
                nsp3-nsp4ᶜ       ANKKGA2517-GL      PL1pro / PL2pro
IBV             nsp1-nsp2        VVCKAG673 -GK      PLpro
                nsp2-nsp3        VEKKAG2265-GI      PLpro                       } R-group


HCoV-OC43       nsp1-nsp2        IRGYRG246 -VK      PL1pro
                nsp2-nsp3        WRVPCA851 -GR      PL1pro
HCoV-HKU1       nsp1-nsp2        LRQYRG222 -IK      PL1pro
                nsp2-nsp3        WRFPCA809 -GR      PL1pro
BCoV            nsp1-nsp2        IRGYRG246 -VK      PL1pro                       } O-group
                nsp2-nsp3        WRVPCA851 -GR      PL1pro
MHV             nsp1-nsp2        LKGYRG247 -VK      PL1pro
                nsp2-nsp3        WRFPCA832 -GK      PL1pro
```

Fig. 6. Assignment of confirmed/predicted cleavage site sequences in coronavirus replicase polyproteins and processing PLpros based on predicted requirements at the S_1 and S_2 subsites. nspX-nspY indicates cleavage between nonstructural proteins X and Y of the polyprotein. The P_1 and P_2 positions of the cleavage site sequence are highlighted on black background and are colored in cyan for small residues (Gly, Ala) and yellow otherwise. The right column lists the PLpros responsible for the processing event at each site. Enzyme names are given on black background if the respective cleavage event is supported by experimental data. The annotated predicted PLpro-mediated cleavage sites were retrieved form the SwissProt (SW) and/or Genbank (GB) databases, except those for HCoV-NL, HCoV-OC43, and HCoV-HKU1, which were derived by similarity in this work, and for the TGEV nsp1-nsp2 and nsp3-nsp4 cleavage sites, reannotated in this work based on the predicted binding site architectures. [a] The SW and GB annotations for the TGEV nsp1-nsp2 cleavage site are as ARTGRG$_{110}$-AI and KIARTG$_{108}$-RG, respectively. [b] The SW and GB annotation for the TGEV nsp3-nsp4 cleavage site is VSPKSG$_{2388}$-SG. [c] The shown PEDV nsp3-nsp4 cleavage site corresponds to the GB annotation; the SW annotation for this site is IANKKG$_{2516}$-AG. See Materials and Methods for nomenclature and sequence accession numbers.

found in P_1'. Mutational data suggest that the presence of P_1 Gly and P_2 Ala, but not P_1' Gly are essential for cleavage. The substrate specificities of HCoV-229E PL1pro and PL2pro were established by determination of the polyprotein processing sites by sequence analysis in the laboratories of Siddell and Ziebuhr.[19,21] Importantly, both enzymes exhibited overlapping substrate specificities at the nsp2-nsp3 cleavage site,[21] and the two experimentally confirmed PLpro-processing sites of HCoV-229E feature P_1 Gly and P_2 Gly/Ala.[21,37]

In summary, the confirmed sites processed by R-group coronaviral PLpros show a stringent requirement for Gly/Ala in P_1 and P_2, which agrees with the restricted nature of the S_1 and S_2 subsites predicted for this group. The O-group MHV-PL1pro processes the polyprotein at sites with Gly/Ala at P_1 and Arg/Cys at P_2, which corresponds to the more open S_2 subsite in this group. Thus, our classification of coronaviral PLpros, which is based on the predicted topology of the nonprimed side of the substrate-binding site, correlates with specificity and activity data available for some of these enzymes

(see also Fig. 6). It is interesting to note that MHV PL1pro (O-group) and MHV PL2pro (R-group), in addition to their different substrate specificities, also display distinct behaviors toward E-64d, a membrane-permeable derivative of the cysteine protease-specific irreversible epoxysuccinyl inhibitor E-64. In virus-infected cells, E-64d was shown to block the MHV PL1pro-mediated processing of nsp1 and nsp2.[28,31] MHV PL2pro-mediated nsp2-nsp3 cleavage, however, appeared to be E64d-insensitive.[61] The molecular basis for E-64d specificity can be attributed to a Leu residue that normally binds into S_2 subsite of most cellular PLpros.[62,63] The steric occlusion of the S_2 pocket in MHV PL2pro most likely precludes the accommodation of large P_2 substrate side chains or the bulky Leu side chain of the E-64d inhibitor. In contrast, MHV PL1pro has an open papain-like S_2 pocket, which can accommodate bulky moieties, such as the side chains of Leu (from the E-64d inhibitor), Arg (from the nsp1-nsp2 processing site), or Cys (from the nsp2-nsp3 processing site), but would not establish a productive contact with a small Gly residue (Fig. 5).

Evaluation of Annotated PLpro-mediated Processing Sites

In the absence of experimental specificity data for other coronaviral PLpros, Figure 6 summarizes the PLpro-mediated processing sites sequences as annotated by the SwissProt and GenBank databases, based on similarity to confirmed processing sites. We recognize a remarkable complementarity between these sites and our binding site-based classifications of coronaviral PLpros. Specifically, the majority of annotated processing sites for R-group PLpros feature small residues (Gly, Ala) in P_1 and P_2, which is in agreement with the restricted nature of the binding sites of the processing PLpros. For PEDV, however, there are different predictions in the SwissProt and GenBank databases for the nsp1-nsp2 cleavage sites. Our classification is consistent with the GenBank annotation. In the case of TGEV, the annotated nsp1-nsp2 and nsp3-nsp4 PLpro-mediated cleavage sites contain larger P_2 residues, i.e., Arg (according to SwissProt) or Thr (according to GenBank) in the nsp1-nsp2 cleavage site, and Ser for nsp3-nsp4. Given the restricted S_1 and S_2 binding subsites predicted for both PL1pro and PL2pro of TGEV, we revised the nsp1-nsp2 cleavage site annotation to A111-I112, one residue downstream of the SwissProt annotation, thus placing Ala and Gly in the P_1 and P_2 positions, respectively. The nsp3-nsp4 cleavage site of TGEV may also be subject to revision. Processing may rather occur between S2389 and G2390, which is also one residue downstream to the current annotation. Although this reannotation positions Ser instead of Gly in P_1, it substitutes Ser for Gly in the more restricted S_2 subsite and displaces Pro from P_4 to P_5. In our model of SCoV PLpro, D1704, which is fully conserved in the R-group PLpros (Fig. 4), forms a hydrogen bond to the P_4 main-chain NH group (Fig. 3). However, this would be incompatible with the presence of a P_4 Pro as predicted by the current database annotation. Another alternative cleavage site, in our opinion less favorable, would be between G2390 and F2391, two residues downstream to the current annotation, which although positions Pro in P_6 and Gly in P_1, introduces Ser in the more restricted S_2 subsite and places the bulky hydrophobic Phe in P_1' (unique among coronaviral PLpro cleavage sites).

In the O-group coronaviral PLpros, the nsp1-nsp2 and nsp2-nsp3 processing sites confirmed for MHV PL1pro and the predicted corresponding sites for the PL1pros from HCoV-OC43, HCoV-HKU1, and BCoV are highly conserved (Fig. 6). As presented earlier, bulky P_2 residues (Arg, Cys) are predicted to fit into the spacious S_2 pocket of these enzymes (Fig. 5). Obviously, in the coronaviral genomes that do not contain an O-type PLpro, the P_1 and P_2 side chains of both the nsp1-nsp2 and the nsp2-nsp3 cleavage sites are reduced in size to fit the R-type PLpro binding site.

On the Evolution of Coronaviral PLpros

Our results base the classification of coronaviral PLpros on structural binding site relationships, superseding previous classification attempts. Different from a previously reconstructed phylogenetic tree of coronaviral PLpros,[21] for example, our classification does not group the PL2pros of MHV and BCoV together with their PL1pros, but rather with the PL1pros and PL2pros from HCoV-229E and TGEV. The putative representation of PLpros, whether R- or O-type, in the primordial nsp3, and their evolution in the contemporary lineages of coronaviruses,[64] remains speculative. It cannot be ruled out that the involvement of PLpros in processes other than polyprotein processing has played a part in the diversification of their structural relationships, and influenced the co-evolution of their cleavage sites. Interestingly, the O-type signature along with its corresponding cleavage site sequences appears to be less diverse (Figs. 4 and 6), maybe owing to a more recent evolutionary origin than for the R-type signature. Our results, however, suggest that a conversion of PLpro specificity in either direction would have been associated with major structural active site rearrangements, requiring considerable evolutionary pressure.

On the Predicted Deubiquitinating Activity of Coronaviral PLpros

Our prediction of deubiquitinating activity of SCoV PLpro[40] can safely be extended to the R-group enzymes that cleave the polyprotein at sites that contain the motif LXGG in P_4 to P_1. These enzymes are the PL2pros from HCoV-OC43, HCoV-HKU1, BCoV, and MHV (Fig. 6). In order to comment on the ability of other R-group PLpros to deubiquitinate proteins, further experimental and theoretical studies are needed to elucidate whether those coronaviral PLpros can accommodate in their binding sites a P_4 Leu and a P_3 Arg, the residues found in ubiquitin. Owing to the requirement for a bulky P_2 residue[34] and the predicted spacious S_2 subsite (Fig. 5), it is unlikely that the O-group PLpros will possess deubiquitinating activity. An interesting observation is that, for those coronaviruses where a P_4 Leu residue is found at the replicase cleavage site by an R-group PLpro, the coronavirus also has an O-group PLpro (i.e., HCoV-OC43, HCoV-HKU1, BCoV, and MHV). The R-group enzyme performs a single cleavage of the polyprotein at nsp3-nsp4, whereas the O-group enzyme cleaves at nsp1-nsp2 and nsp2-nsp3. The SARS coronavirus is an exception, because it has an R-group PLpro that processes at sites containing Leu in the P_4 position, but it lacks an O-group PLpro. In the coronaviruses where there is no Leu in the P_4 of the processing site, there are two R-group enzymes, or in the case of IBV, only one R-group PLpro.

CONCLUSION

Owing to the wealth of protein 3D structural data coupled with the constant improvement of fold recognition algorithms, a significant structural relationship could be detected between the catalytic core domains of SCoV PLpro and HAUSP cysteine proteases, both featuring a circularly permuted Zn-ribbon domain inserted in the middle of a papain-like fold. One can thus reconsider the current classification of coronaviral PLpros and USPs into families C16 and C19, respectively, in the MEROPS

peptidase database (http://merops.sanger.ac.uk). [26] Comparative sequence analysis data superimposed onto a binding site structural framework show that coronaviral PLpros can be classified into two groups according to their binding site architectures. One group, termed R and present in all currently known coronaviruses, is predicted to feature sterically restricted S_1 and S_2 substrate-binding subsites and a P_1-P_4-substrate-binding mode characteristic of USPs. The other group, termed O, particularly features an open S_2 subsite and a substrate-binding mode that resembles more papain and related cathepsins. This classification, which differs in part from those extracted from the phylogenetic trees of coronaviral replicases and PLpro domains, is a first step toward the understanding of the molecular basis for the processing specificity and inhibition selectivity data that has become available for several members of the family. For the remaining coronaviral PLpros, the R/O classification can be used to critically evaluate and, in a few instances, to revise the publicly available annotations of polyprotein cleavage sites. The ubiquitous presence of the R-group binding site in all coronaviruses can be advantageously exploited to design PLpro inhibitors with a wide-spectrum efficacy against all coronaviruses. Certainly, experimental structure determinations, at least for one family member, will be valuable for a more reliable identification of those structural details that may be required to overcome the predicted inhibitory cross-reactivity with host enzymes, particularly with the USPs.

REFERENCES

1. Cavanagh D. Nidovirales: a new order comprising Coronaviridae and Arteriviridae. Arch Virol 1997;142:629–633.
2. Ziebuhr J, Siddell SG. Nidoviruses. In: Atlas R, Bynum WF, Cox M, editors. Encyclopedia of life sciences. London: Stockton Press; 2002. p 190–198.
3. Lai MM, Holmes KV. Coronaviridae: the viruses and their replication. In: Knipe DM, Howley PM, Griffin DE, Lamb RA, Martin MA, Roizman B, Straus SE, editors. Fields virology. Philadelphia: Lippincott Williams & Wilkins; 2001. p 1163–1185.
4. Holmes KV. Coronaviruses. In: Knipe DM, Howley PM, Griffin DE, Lamb RA, Martin MA, Roizman B, Straus SE, editors. Fields virology. Philadelphia: Lippincott Williams & Wilkins; 2001. p 1187–1203.
5. Fouchier RAM, Hartwig NG, Bestebroer TM, Niemeyer B, de Jong JC, Simon JH, Osterhaus ADME. A previously undescribed coronavirus associated with respiratory disease in humans. Proc Natl Acad Sci USA 2004;101:6212–6216.
6. van der Hoek L, Pyrc K, Jebbink MF, Vermeulen-Oost W, Berkhout RJ, Wolthers KC, Wertheim-van Dillen PM, Kaandorp J, Spaargaren J, Berkhout B. Identification of a new human coronavirus. Nat Med 2004;10:368–373.
7. Esper F, Weibel C, Ferguson D, Landry ML, Kahn JS. Evidence of a novel human coronavirus that is associated with respiratory tract disease in infants and young children. J Infect Dis 2005;191: 492–498.
8. Woo PCY, Lau SKP, Chu Cm, Chan Kh, Tsoi Hw, Huang Y, Wong BHL, Poon RWS, Cai JJ, Luk Wk, Poon LLM, Wong SSY, Guan Y, Peiris JSM, Yuen Y. Characterization and complete genome sequence of a novel coronavirus, coronavirus HKU1, from patients with pneumonia. J Virol 2005;79:884–895.
9. Skowronski DM, Astell C, Brunham RC, Low DE, Petric M, Roper RL, Talbot PJ, Tam T, Babiuk L. Severe acute respiratory syndrome (SARS): a year in review. Annu Rev Med 2005;56:357–381.
10. Song HD, Tu CC, Zhang GW, Wang SY, Zheng K, Lei LC, Chen QX, Gao YW, Zhou HQ, Xiang H, Zheng HJ, Chern SWW, Cheng F, Pan CM, Xuan H, Chen SJ, Luo HM, Zhou DH, Liu YF, He JF, Qin PZ, Li LH, Ren YQ, Liang WJ, Yu YD, Anderson L, Wang M, Xu RH, Wu XW, Zheng HY, Chen JD, Liang G, Gao Y, Liao M, Fang L, Jiang LY, Li H, Chen F, Di B, He LJ, Lin JY, Tong S, Kong X, Du L, Hao P, Tang H, Bernini A, Yu XJ, Spiga O, Guo ZM, Pan HY, He WZ, Manuguerra JC, Fontanet A, Danchin A, Niccolai N, Li YX, Wu CI, Zhao GP. Cross-host evolution of severe acute respiratory syndrome coronavirus in palm civet and human. Proc Natl Acad Sci USA 2005;102:2430–2435.
11. Oxford JS, Balasingam S, Chan C, Catchpole A, Lambkin R. New antiviral drugs, vaccines and classic public health interventions against SARS coronavirus. Antivir Chem Chemother 2005;16:13–21.
12. Sung JJ, Wu A, Joynt GM, Yuen KY, Lee N, Chan PK, Cockram CS, Ahuja AT, Yu LM, Wong VW, Hui DS. Severe acute respiratory syndrome: report of treatment and outcome after a major outbreak. Thorax 2004;59:414–420.
13. Chu CM, Cheng VC, Hung IF, Wong MM, Chan KH, Chan KS, Kao RY, Poon LL, Wong CL, Guan Y, Peiris JS, Yuen KY. Role of lopinavir/ritonavir in the treatment of SARS: initial virological and clinical findings. Thorax 2004;59:252–256.
14. Ziebuhr J. The coronavirus replicase. Curr Top Microbiol Immunol 2005;287:57–94.
15. Ziebuhr J, Snijder EJ, Gorbalenya AE. Virus-encoded proteinases and proteolytic processing in the Nidovirales. J Gen Virol 2000;81: 853–879.
16. Brockway SM, Lu XT, Peters TR, Dermody TS, Denison MR. Intracellular localization and protein interactions of the gene 1 protein p28 during mouse hepatitis virus replication. J Virol 2004;78:11551–11562.
17. Prentice E, McAuliffe J, Lu X, Subbarao K, Denison MR. Identification and characterization of severe acute respiratory syndrome coronavirus replicase proteins. J Virol 2004;78:9977–9986.
18. Salonen A, Ahola T, Kaariainen L. Viral RNA replication in association with cellular membranes. Curr Top Microbiol Immunol 2005;285:139–173.
19. Herold J, Thiel V, Siddell SG. Characterization of a papain-like cysteine-proteinase encoded by gene 1 of the human coronavirus HCV 229E. Adv Exp Med Biol 1998;440:141–147.
20. Lim KP, Ng LFP, Liu DX. Identification of a novel cleavage activity of the first papain-like proteinase domain encoded by open reading frame 1a of the coronavirus avian infectious bronchitis virus and characterization of the cleavage products. J Virol 2000;74:1674–1685.
21. Ziebuhr J, Thiel V, Gorbalenya AE. The autocatalytic release of a putative RNA virus transcription factor from its polyprotein precursor involves two paralogous papain-like proteases that cleave the same peptide bond. J Biol Chem 2001;276:33220–33232.
22. Thiel V, Ivanov KA, Putics A, Hertzig T, Schelle B, Bayer S, Weissbrich B, Snijder EJ, Rabenau H, Doerr HW, Gorbalenya AE, Ziebuhr J. Mechanisms and enzymes involved in SARS coronavirus genome expression. J Gen Virol 2003;84:2305–2315.
23. Kanjanahaluethai A, Jukneliene D, Baker SC. Identification of the murine coronavirus MP1 cleavage site recognized by papain-like proteinase 2. J Virol 2003;77:7376–7382.
24. Teng H, Weiss SR. Further in vitro characterization of mouse hepatitis virus papain-like proteinase 1: cleavage sequence requirements within pp1a. J Neurovirol 2002;8:143–149.
25. Harcourt BH, Jukneliene D, Kanjanahaluethai A, Bechill J, Severson KM, Smith CM, Rota PA, Baker SC. Identification of severe acute respiratory syndrome coronavirus replicase products and characterization of papain-like protease activity. J Virol 2004;78:13600–13612.
26. Rawlings ND, Tolle DP, Barrett AJ. MEROPS: the peptidase database. Nucleic Acids Res 2004;32:D160–D164.
27. Snijder EJ, Bredenbeek PJ, Dobbe JC, Thiel V, Ziebuhr J, Poon LLM, Guan Y, Rozanov M, Spaan WJM, Gorbalenya AE. Unique and conserved features of genome and proteome of SARS-coronavirus, an early split-off from the coronavirus group 2 lineage. J Mol Biol 2003;331:991–1004.
28. Gosert R, Kanjanahaluethai A, Egger D, Bienz K, Baker SC. RNA Replication of mouse hepatitis virus takes place at double-membrane vesicles. J Virol 2002;76:3697–3708.
29. Perlman S, Ries D, Bolger E, Chang LJ, Stoltzfus CM. MHV nucleocapsid synthesis in the presence of cycloheximide and accumulation of negative strand MHV RNA. Virus Res 1986;6:261–272.

30. Sawicki SG, Sawicki DL. Coronavirus minus-strand RNA synthesis and effect of cycloheximide on coronavirus RNA synthesis. J Virol 1986;57:328–334.

31. Kim JC, Spence RA, Currier PF, Lu X, Denison MR. Coronavirus protein processing and RNA synthesis is inhibited by the cysteine proteinase inhibitor E64d. Virology 1995;208:1–8.

32. Denison MR, Yount B, Brockway SM, Graham RL, Sims AC, Lu X, Baric RS. Cleavage between replicase proteins p28 and p65 of mouse hepatitis virus is not required for virus replication. J Virol 2004;78:5957–5965.

33. Teng H, Pinon JD, Weiss SR. Expression of murine coronavirus recombinant papain-like proteinase: efficient cleavage is dependent on the lengths of both the substrate and the proteinase polypeptides. J Virol 1999;73:2658–2666.

34. Dong S, Baker SC. Determinants of the p28 cleavage site recognized by the first papain-like cysteine proteinase of murine coronavirus. Virology 1994;204:541–549.

35. Hughes SA, Bonilla PJ, Weiss SR. Identification of the murine coronavirus p28 cleavage site. J Virol 1995;69:809–813.

36. Bonilla PJ, Hughes SA, Weiss SR. Characterization of a second cleavage site and demonstration of activity in trans by the papain-like proteinase of the murine coronavirus mouse hepatitis virus strain A59. J Virol 1997;71:900–909.

37. Herold J, Gorbalenya AE, Thiel V, Schelle B, Siddell SG. Proteolytic processing at the amino terminus of human coronavirus 229E gene 1-encoded polyproteins: identification of a papain-like proteinase and its substrate. J Virol 1998;72:910–918.

38. Lim KP, Liu DX. Characterization of the two overlapping papain-like proteinase domains encoded in gene 1 of the coronavirus infectious bronchitis virus and determination of the C-terminal cleavage site of an 87-kDa protein. Virology 1998;245:303–312.

39. Herold J, Siddell SG, Gorbalenya AE. A human RNA viral cysteine proteinase that depends upon a unique Zn2+-binding finger connecting the two domains of a papain-like fold. J Biol Chem 1999;274:14918–14925.

40. Sulea T, Lindner HA, Purisima EO, Menard R. Deubiquitination, a new function of the severe acute respiratory syndrome coronavirus papain-like protease? J Virol 2005;79:4550–4551.

41. Krishna SS, Grishin NV. The finger domain of the human deubiquitinating enzyme HAUSP is a zinc ribbon. Cell Cycle 2004;3:1046–1049.

42. Bujnicki JM, Elofsson A, Fischer D, Rychlewski L. Structure prediction meta server. Bioinformatics 2001;17:750–751.

43. Ginalski K, Elofsson A, Fischer D, Rychlewski L. 3D-Jury: a simple approach to improve protein structure predictions. Bioinformatics 2003;19:1015–1018.

44. Ginalski K, Rychlewski L. Detection of reliable and unexpected protein fold predictions using 3D-Jury. Nucleic Acids Res 2003;31:3291–3292.

45. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 1994;22:4673–4680.

46. Rost B. Protein secondary structure prediction continues to rise. J Struct Biol 2001;134:204–218.

47. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.

48. Karplus K, Barrett C, Hughey R. Hidden Markov models for detecting remote protein homologies. Bioinformatics 1998;14:846–856.

49. Albrecht M, Tosatto SCE, Lengauer T, Valle G. Simple consensus procedures are effective and sufficient in secondary structure prediction. Protein Eng 2003;16:459–462.

50. Blundell T, Carney D, Gardner S, Hayes F, Howlin B, Hubbard T, Overington J, Singh DA, Sibanda BL, Sutcliffe M. Knowledge-based protein modelling and design. Eur J Biochem 1988;172:513–520.

51. Hu M, Li P, Li M, Li W, Yao T, Wu JW, Gu W, Cohen RE, Shi Y. Crystal structure of a UBP-family deubiquitinating enzyme in isolation and in complex with ubiquitin aldehyde. Cell 2002;111:1041–1054.

52. Guarne A, Hampoelz B, Glaser W, Carpena X, Tormo J, Fita I, Skern T. Structural and biochemical features distinguish the foot-and-mouth disease virus leader proteinase from other papain-like enzymes. J Mol Biol 2000;302:1227–1240.

53. Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, Ferguson DM, Spellmeyer DC, Fox T, Caldwell JW, Kollman PA. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J Am Chem Soc;117:5179–5197.

54. Guarne A, Tormo J, Kirchweger R, Pfistermueller D, Fita I, Skern T. Structure of the foot-and-mouth disease virus leader protease: a papain-like fold adapted for self-processing and eIF4G recognition. EMBO J 1998;17:7469–7479.

55. Andreeva A, Howorth D, Brenner SE, Hubbard TJP, Chothia C, Murzin AG. SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res 2004;32:D226–D229.

56. Anand K, Ziebuhr J, Wadhwani P, Mesters JR, Hilgenfeld R. Coronavirus main proteinase (3CLpro) structure: basis for design of anti-SARS drugs. Science 2003;300:1763–1767.

57. Bromme D, Bonneau PR, Lachance P, Storer AC. Engineering the S2 subsite specificity of human cathepsin S to a cathepsin L- and cathepsin B-like specificity. J Biol Chem 1994;269:30238–30242.

58. Fox T, Mason P, Storer AC, Mort JS. Modification of S1 subsite specificity in the cysteine protease cathepsin B. Protein Eng 1995;8:53–57.

59. Quesada V, Diaz-Perales A, Gutierrez-Fernandez A, Garabaya C, Cal S, Lopez-Otin C. Cloning and enzymatic analysis of 22 novel human ubiquitin-specific proteases. Biochem Biophys Res Commun 2004;314:54–62.

60. Amerik AY, Hochstrasser M. Mechanism and function of deubiquitinating enzymes. Biochim Biophys Acta 2004;1695:189–207.

61. Kanjanahaluethai A, Baker SC. Identification of mouse hepatitis virus papain-like proteinase 2 activity. J Virol 2000;74:7911–7921.

62. Varughese KI, Su Y, Cromwell D, Hasnain S, Xuong NH. Crystal structure of an actinidin-E-64 complex. Biochemistry 1992;31:5172–5176.

63. Zhao B, Janson CA, Amegadzie BY, D'Alessio K, Griffin C, Hanning CR, Jones C, Kurdyla J, McQueney M, Qiu X, Smith WW, Abdel-Meguid SS. Crystal structure of human osteoclast cathepsin K complex with E-64. Nat Struct Biol 1997;4:109–111.

64. Gonzalez JM, Gomez-Puertas P, Cavanagh D, Gorbalenya AE, Enjuanes L. A comparative sequence analysis to revise the current taxonomy of the family Coronaviridae. Arch Virol 2003;148:2207–2235.

65. Min J, Landry J, Sternglanz R, Xu RM. Crystal structure of a SIR2 homolog-NAD complex. Cell 2001;105:269–279.

APPENDIX A
Circularly Permuted Zn-ribbon Domain of Coronaviral PLpros

A recent report demonstrates that the previously uncharacterized finger domain, inserted between the two subdomains of the papain fold of the HAUSP catalytic core domain, represents a circularly permuted Zn-ribbon.[41] Although in HAUSP this domain has lost its zinc-binding ability because of mutation of two of the Zn-chelating residues, intact Zn-chelating capability appears to be present in a number of USPs (family C19 in the MEROPS database, http://merops.sanger.ac.uk) that are close homologs of HAUSP (Fig. A1). Our model predicts that the Zn-ribbon domain of SCoV PLpro resembles that of HAUSP, but essentially differs from the previous model of HCoV-229E PL1pro, in which the topology of the corresponding sequence was based on classical Zn-ribbons.[39] Although this prediction can be fully validated only by an experimental structure, there are several lines of evidence supporting the existence of a circularly permuted instead of a classical Zn-ribbon topology for the intermediate domain of coronaviral PLpros.

The first line of evidence is represented by the fold recognition result itself, with the detection of HAUSP as the only statistically significant structural template for SCoV PLpro. Sequence comparisons (Fig. A1) suggest that the putative Zn-chelating residues of coronaviral PLpros
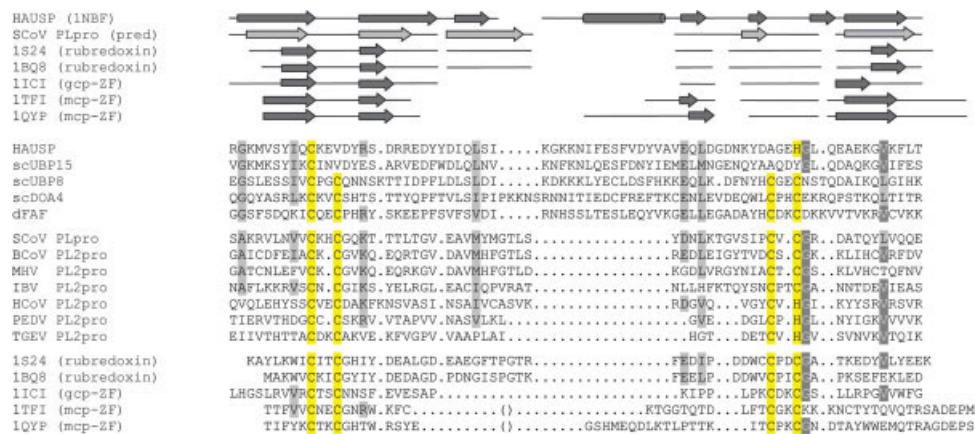
Fig. A1. Sequence comparison of the putative circularly permuted Zn-ribbon domains of SCoV PLpro and selected coronaviral PLpros, the finger domain inserts of several USPs, and members of the rubredoxin-like fold in the SCOP database,[55] including classical and circularly permuted Zn-ribbons. gcp-ZF denotes a genuine circularly permuted Zn finger. mcp-ZF denotes manually circularly permuted Zn fingers, with the position of circular permutation indicated in the sequence by "( )". Secondary structure elements are rendered as in Figure 1. Metal-chelating residues are highlighted on yellow background. Residues identical in half or more of all sequences are in white on dark gray background, and those conserved in half or more of all sequences are on light gray background, based on the BOXSHADE program (http://www.ch.embnet.org). HCoV refers to HCoV-229E. See Materials and Methods for coronavirus nomenclature and sequence accession numbers. The USPs selected along with HAUSP are from the SwissProt database: scUBP15 (*Saccharomyces cerevisiae*; SW: P50101); scUBP8 (*S. cerevisiae*; SW: P50102); scDOA4 (*S. cerevisiae*; SW: P32571); and dFAF (*Drosophila melanogaster*; SW: P55824).

align onto the HAUSP residues corresponding precisely to the predicted Zn-chelating positions of related USPs.[41] Second, the predicted secondary structure elements of coronaviral PLpros correspond to those determined for the circularly permuted Zn-ribbon-like domain (previously termed finger domain) of HAUSP.[51] Notably, these two criteria also apply to a comparison with the sequence of the circularly permuted Zn-ribbon fold from the structure of the silent information regulator 2 (Sir2) homolog,[65] the only other known representative of this fold (PDB code 1ICI; Fig. A1). In contrast, coronaviral PLpros can be readily aligned (i.e., preserving the spacing between secondary structure elements after alignment of Zn-chelating residues) onto representatives of the classical C4-type Zn-ribbon fold only after assuming the appropriate circular permutation of the latter (Fig. A1).

Further fold recognition data provide a third line of evidence for a circularly permuted Zn-ribbon domain in coronaviral PLpros. By querying the sequence of SCoV PLpro intermediate domain to fold recognition servers, rubredoxins were top-ranked as structural relatives by consensus scoring, albeit below the significance threshold of the 3D-JURY method. The iron- instead of zinc-binding rubredoxins display the same overall fold topology as circularly permuted Zn-ribbons according to the SCOP database (http://scop.berkeley.edu).[55] As for USPs and naturally or manually circularly permuted Zn-ribbons, the alignment of rubredoxins onto coronaviral PLpros sequences agrees with the assignment of metal-chelating residues and secondary structure conservation (Fig. A1). Fold recognition, however, failed to signal the genuine circularly permuted Zn-ribbons in the structures of HAUSP and Sir2 homolog, which is not surprising because in these

cases the Zn-ribbon domains are part of much larger protein structures. These failures rather reflect an existing shortcoming of present fold recognition methods to correctly detect suitable template domains embedded in large multidomain structures. It does not necessarily imply that these two genuine circularly permuted Zn-ribbons are more distant structural homologs of coronaviral PLpros than rubredoxins. Importantly, classical Zn-ribbons were not detected even though they are represented in the PDB as single-domain structures.

The fourth line of evidence is given by the 3D structural comparison of classical versus circularly permuted Zn-ribbons (including rubredoxins, see Fig. A2). A classical Zn-ribbon fold has its chain termini forming the outer β-strands of the β-sheet. In contrast, a circularly permuted Zn-ribbon fold has its chain termini forming the inner strands of the β-sheet. In both cases, the inner strands are generally longer than the outer ones. The difference between the two folds is particularly striking at the N-terminus. In the classical Zn-ribbon fold, the N-terminus forms a very short outer β-strand, which is even absent in some of the fold representatives. In the circularly permuted Zn-ribbon fold family, including rubredoxins, on the other hand, the N-terminus forms a long inner β-strand. Our prediction that the Zn-ribbon domain of SCoV PLpro contains long β-strands at the sequence termini is compatible with a circularly permuted fold.

As mentioned earlier, there is good agreement between the secondary structures predicted for the intermediate domain in coronaviral PLpros and the one observed for the Zn-ribbon domain of HAUSP. The latter has two additional structural features outside the β-ribbon: a β-strand and an α-helix in the crossover segment connecting the
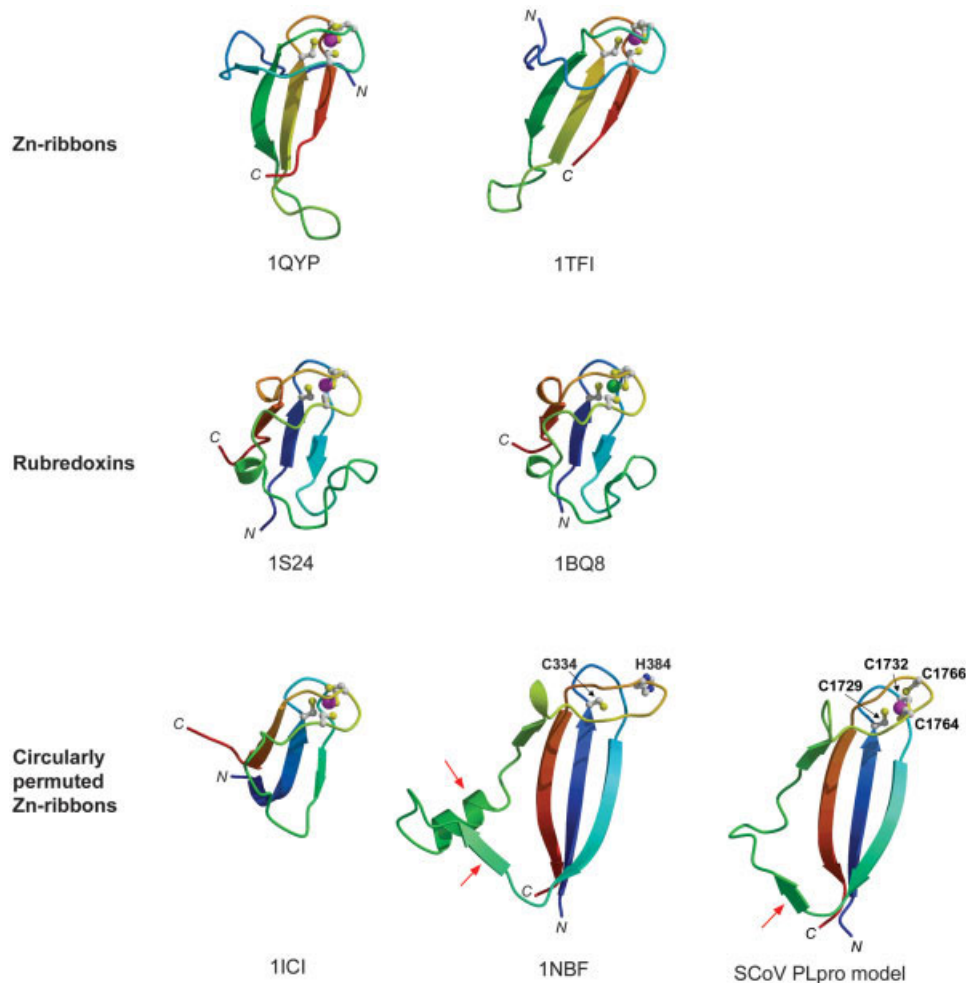
Fig. A2. Structural comparison between classical C4-type Zn-ribbons, rubredoxins, and circularly permuted Zn-ribbons. The two structures shown in the upper row are representatives of a large family of the classical C4-type Zn-ribbon fold, namely, from the RNA polymerase II subunit 9 (left, PDB code 1QYP: residues 1–57) and from the transcription elongation factor SII (right, 1TFI: 1–50). The two structures shown in the middle row are typical examples of rubredoxins, from *Pseudomonas oleovorans* (left, 1S24: 1–56), and from *Pyrococcus furiosus* (right, 1BQ8: 1–54). The structures shown in the bottom row are the currently known members of the circularly permuted C4-type Zn-ribbon fold, from the Sir2 homolog (left, 1ICI: 116–159) and from HAUSP (middle, 1NBF: 325–399) together with the Zn-ribbon in the modeled structure of SCoV PLpro (right, residues 1720–1779). Ribbons are colored using a rainbow color ramp starting from blue at the N-terminus and ending in red at the C-terminus of each domain. The Zn^{2+} ions bound to classical and circularly permuted Zn-ribbons are shown as magenta spheres. A Cd^{2+} ion and a Fe^{3+} ion bound to the exemplified rubredoxins are shown as purple and green spheres, respectively. The metal-chelating cysteine residues are also displayed, except for the circularly permuted Zn-ribbon of HAUSP that lost its Zn^{2+}-binding capability and retains two of the four metal-coordinating residues (Cys and His, also displayed). Red arrows indicate additional secondary structure elements present outside the β-ribbon in the circularly permuted Zn-ribbon domains of HAUSP and SCoV PLpro.

outer β-strands of the circularly permuted β-ribbon (Fig. A2). Both these structural features are utilized in HAUSP to anchor the Zn-ribbon domain to the protease domain. The isolated β-strand is preserved in the modeled circularly permuted Zn-ribbon of SCoV PLpro. Similar to the HAUSP structure, it establishes a parallel β-strand interaction to the protease domain (Fig. 2). Contrary to HAUSP, the α-helix insertion is, however, predicted to be absent from the circularly permuted Zn-ribbon of SCoV PLpro, as is its interacting loop from the protease domain. This suggests that the relative orientation of the protease and

Zn-ribbon domains in the coronaviral enzyme is less rigid than in HAUSP. This may have implications for ubiquitin binding and enzyme regulation.

Although Zn^{2+}, not Fe^{2+}, has been established as an essential cofactor of HCoV-229E PL1pro,[39] it has to be considered that rubredoxins may also represent viable templates for the modeling of the putative Zn-ribbon domain of coronaviral PLpros. In fact, sequence similarities of the intermediate domain of coronaviral PLpros to some rubredoxins appear to be even stronger than to HAUSP. In this regard, it is also interesting to note that

Herold and coworkers determined an increased amount of Fe^{2+} instead of Zn^{2+} bound to the protein when during recombinant expression supplementary zinc acetate was omitted from the bacterial growth medium.[39] However, secondary structure similarities are predicted to be more pronounced when SCoV PLpro is compared with HAUSP (Fig. A1). The principal difference between the rubredoxin and HAUSP Zn-ribbon structures rests in the crossover segment outside of the β-ribbon (Fig. A2), which in the HAUSP structure mediates the attachment to the protease domain and participates in direct interactions with Ubal.[51] In rubredoxins on the other hand, the crossover segment is a loop that folds against the β-sheet and does not reach to the opposite edge of the β-sheet. Because several of the conserved residues that stabilize the crossover segment in rubredoxins are different in SCoV PLpro, the crossover loop of coronaviral PLpros may have a decreased propensity to fold against the β-sheet of the Zn-ribbon domain, and may, therefore, become available for direct interaction with the protease domain, as modeled in this study for SCoV PLpro (Fig. 2). Such a loop conformation would further be expected to affect interdomain flexibility and ligand binding as seen in the HAUSP-ubiquitin complex.[51] Curiously, many of the HAUSP-related C19-family USPs such as UBP4, UBP15, and UBC11 from higher eukaryotes, feature an uncharacterized sequence insertion of ~290 residues between the β-strand and the α-helix within the crossover loop of the Zn-ribbon domain, as judged by a sequence family alignment that can be accessed through the MEROPS database (http://merops.sanger.ac.uk).

Supplementary Material

Supplementary Materials Details on the energy refinement procedure and the atomic coordinates of the modeled SCoV PLpro–Ubal complex are available via the Internet at http://www.interscience.wiley.com/.