Homology Models and Molecular Dynamics Simulations of Main Proteinase from Coronavirus Associated with Severe Acute Respiratory Syndrome (SARS)

Hsuan-Liang Liu[a]* (劉宣良), Jin-Chung Lin[a] (林進中), Yih Ho[b] (何　意), Wei-Chan Hsieh[a] (謝偉強), Chin-Wen Chen[a] (陳錦文) and Yuan-Chen Su[a] (蘇永成)
[a]*Department of Chemical Engineering and Graduate Institute of Biotechnology, National Taipei University of Technology, Taipei 10608, Taiwan, R.O.C.*
[b]*School of Pharmacy, Taipei Medical University, Taipei 110, Taiwan, R.O.C.*

In this study, two structural models (denoted as M^{pro}ST and M^{pro}SH) of the main proteinase (M^{pro}) from the novel coronavirus associated with severe acute respiratory syndrome (SARS-CoV) were constructed based on the crystallographic structures of M^{pro} from transmissible gastroenteritis coronavirus (TGEV) (M^{pro}T) and human coronavirus HcoV-229E (M^{pro}H), respectively. Various 200 ps molecular dynamics simulations were subsequently performed to investigate the dynamics behaviors of several structural features. Both M^{pro}ST and M^{pro}SH exhibit similar folds as their respective template proteins. These structural models reveal three distinct functional domains as well as an intervening loop connecting domains II and III as found in both template proteins. In addition, domain III of these structures exhibits the least secondary structural conservation. A catalytic cleft containing the substrate binding subsites S1 and the S2 between domains I and II are also observed in these structural models. Although these structures share many common features, the most significant difference occurs at the S2 subsite, where the amino acid residues lining up this subsite are least conserved. It may be a critical challenge for designing anti-SARS drugs by simply screening the known database of proteinase inhibitors.

Keywords: Main proteinase; Coronavirus; Severe acute respiratory syndrome (SARS); Molecular dynamics simulations; Functional domain; Structural model; Inhibitor.

INTRODUCTION

An outbreak of atypical pneumonia, designated as severe acute respiratory syndrome (SARS), was first reported in Guangdong Province of China in late 2002, and rapidly spread to several countries.[1,2] Infection by SARS is usually characterized by high fever, malaise, rigor, headache, nonproductive cough and may progress to generalized, interstitial infiltrates in the lung.[3] Attempts to identify the etiology of the SARS outbreak were not successful until March 2003, when laboratories in the United States, Canada, Germany, and Hong Kong isolated a novel coronavirus (SARS-CoV) from SARS patients. The sequence of the complete genome of SARS-CoV was further determined and characterized with two different isolates.[4,5] Phylogenetic analyses and sequence comparisons reveal that SARS-CoV is not closely related to any of the three groups of coronaviruses, including two human coronaviruses, HcoV-229E (group I) and HcoV-OC43 (group II), which are responsible for about 30% of mild upper respiratory tract illnesses,[6-8] in particular, the common cold.[9]

Coronaviruses belong to a diverse group of positive-stranded RNA viruses featuring the largest viral RNA genomes known to date (27-31 kb). They share a similar genome organization and common transcriptional and translational processes as *Arteriviridae*.[10,11] The human coronavirus HcoV-229E replicase gene encodes two overlapping polyproteins, pp1a (replicase 1a, ~450 kDa) and pp1ab (replicase 1ab, ~750 kDa),[12] that mediate all the functions required for viral replication and transcription.[13] The functional polypeptides are released from the polyproteins by extensive proteolytic processing, which is primarily achieved by the 33.1-kDa HCoV-229E main proteinase (M^{pro}).[14] M^{pro} is commonly also called 3C-like proteinase ($3CL^{pro}$) to indicate a similarity of its cleavage site specificity to that observed for picornavirus 3C proteinase ($3C^{pro}$) and the identification of a Cys residue as the principle nucleophile in the context of a predicted two-β-barrel fold.[15,16] M^{pro} from HcoV-229E (M^{pro}H) has been biosynthesized in *Escherichia coli* and the enzyme

* Corresponding author. Tel: +886-2-27712171 ext. 2542; fax: +886-2-27317117; e-mail: f10894@ntut.edu.tw

properties, inhibitor profile, and substrate specificity of the purified protein have been well characterized.[14,17]

Several studies have revealed significant differences in both the active sites and domain structures of M^{pro} from coronavirus and picornavirus.[17-21] It is noteworthy that coronavirus M^{pro} possesses a large C-terminal domain of ~110 amino acid residues (domain III) that is not found in other RNA virus $3CL^{pro}$. Deletion of this domain results in dramatic losses of proteolytic activity, suggesting that the C-terminal domain III of M^{pro} contributes to proteolytic activity through undefined mechanisms. Previous experimental data have shown that the differential cleavage kinetics of sites within pp1a/pp1ab are a conserved feature of coronavirus M^{pro} and that similar processing kinetics for the replicase polyproteins of all coronaviruses can be predicted.[22] Furthermore, the cleavage pattern appears to be conserved in M^{pro} from SARS-CoV ($M^{pro}S$) and from other coronaviruses,[23] as deduced from the genome sequence.[5,24] The functional importance of M^{pro} in the viral life cycle has made this proteinase an attractive target for the development of drugs directed against SARS and other coronavirus infections. Moreover, molecular modeling has suggested that available rhinovirus $3C^{pro}$ inhibitors such as compound AG7088 may be modified to be tested for SARS therapy.[23] Therefore, screening the known proteinase inhibitor libraries may be an appreciated shortcut to discover anti-SARS drugs.[25]

Recently, crystal structures of $M^{pro}H$[23] and M^{pro} from porcine coronavirus (transmissible gastroenteritis virus, TGEV) ($M^{pro}T$) complexed with its inhibitor[26] have been determined. In addition, homology models of $M^{pro}S$ based on the crystal structures of $M^{pro}H$[23] and $M^{pro}T$[25] have been also reported. Comparison of these structures reveals a remarkable degree of conservation of the substrate binding sites, which is further supported by the cleavage of the substrate for the $M^{pro}T$ with the recombinant $M^{pro}S$.[23] In addition, $M^{pro}S$ shows 40 and 44% sequence identity to $M^{pro}H$ and $M^{pro}T$, respectively.[23] Although the results from the deduced genome sequence of SARS-CoV have indicated that it belongs to a new group of coronaviruses,[24] the significantly high sequence identity of $M^{pro}S$ to bovine coronavirus (BCoV) M^{pro} (49%) and mouse hepatitis virus (MHV) M^{pro} (50%) from group II coronaviruses has allowed Anand et al.[23] to recognize it as an outlier among group II coronaviruses.

Molecular dynamics (MD) simulations in the atomic level have been intensively applied to gain insight into the structure-function relationships of proteins. Previously, several MD simulations and molecular docking experiments have been successfully conducted towards various target proteins in our group.[27-33] In this paper, two structural models of $M^{pro}S$ (denoted as $M^{pro}SH$ and $M^{pro}ST$) were constructed based on the crystallographic structures of $M^{pro}H$[23] and $M^{pro}T$[26], respectively, by the comparative approach. In addition, MD simulations were conducted to investigate the dynamics behaviors of these structures. Beyond the continued characterization of M^{pro} from various coronaviruses, the amino acid sequence alignment and structural homology analyses of $M^{pro}S$ presented in this study provide particularly attractive targets for further structure-based studies, such as folding/unfolding mechanism and molecular docking, which are currently being carried out in our group.

METHODS

Model proteins

Structural homology to construct the structural models of $M^{pro}S$ ($M^{pro}ST$ and $M^{pro}SH$) was based on the monomer of the three-dimensional (3D) structure of $M^{pro}T$, refined to 1.96 Å resolution[26] (Fig. 1(A)), and that of $M^{pro}H$, solved at 2.54 Å resolution[23] (Fig. 1(B)), obtained from the protein data bank (PDB; accession numbers 1lvo and 1p9u, respectively). The inhibitor, a substrate analog hexapeptidyl chloromethyl ketone, was removed from the crystallographic structure of $M^{pro}T$ before being used as a template. Unfavorable nonphysical contacts in these structures were then eliminated using the Biopolymer module of the Insight II program (Accelyrs, San Diego, CA, USA) with the force field Discover CVFF (consistent valence force field)[34-36] in the SGI O200 workstation with a 64-bit HIPS RISC R12000 2 × 270 MHz CPU and PMC-Sierra RM7000A 350 MHz processor (Silicon Graphics, Inc., Mountain View, CA, USA), followed by 10,000 energy minimization calculations using steepest descent method, to yield the model proteins for further structure building.

Structural homology

Homology utilizes structure and sequence similarities for predicting unknown protein structures. The Homology module in Insight II allows us to build the 3D models of the target protein (i.e., $M^{pro}S$) using both its amino acid sequence and the structures of known, related model proteins (i.e., $M^{pro}H$ and $M^{pro}T$). The Homology program provides simultaneous optimization of both structure and sequence homologies for multiple proteins in a 3D graphics environment, based on a method developed by Greer.[37] Smith-Waterman pairwise amino acid sequence alignments were performed based on the conserved active site and substrate binding

subsites among M^{pro} from various coronaviruses to find the location of the active site and substrate binding subsite of $M^{pro}S$. The consensus structural conserved regions (SCRs) of the target protein were generated from alignments of the target protein to the model proteins. The atomic coordinates were then transferred from the model proteins to the target

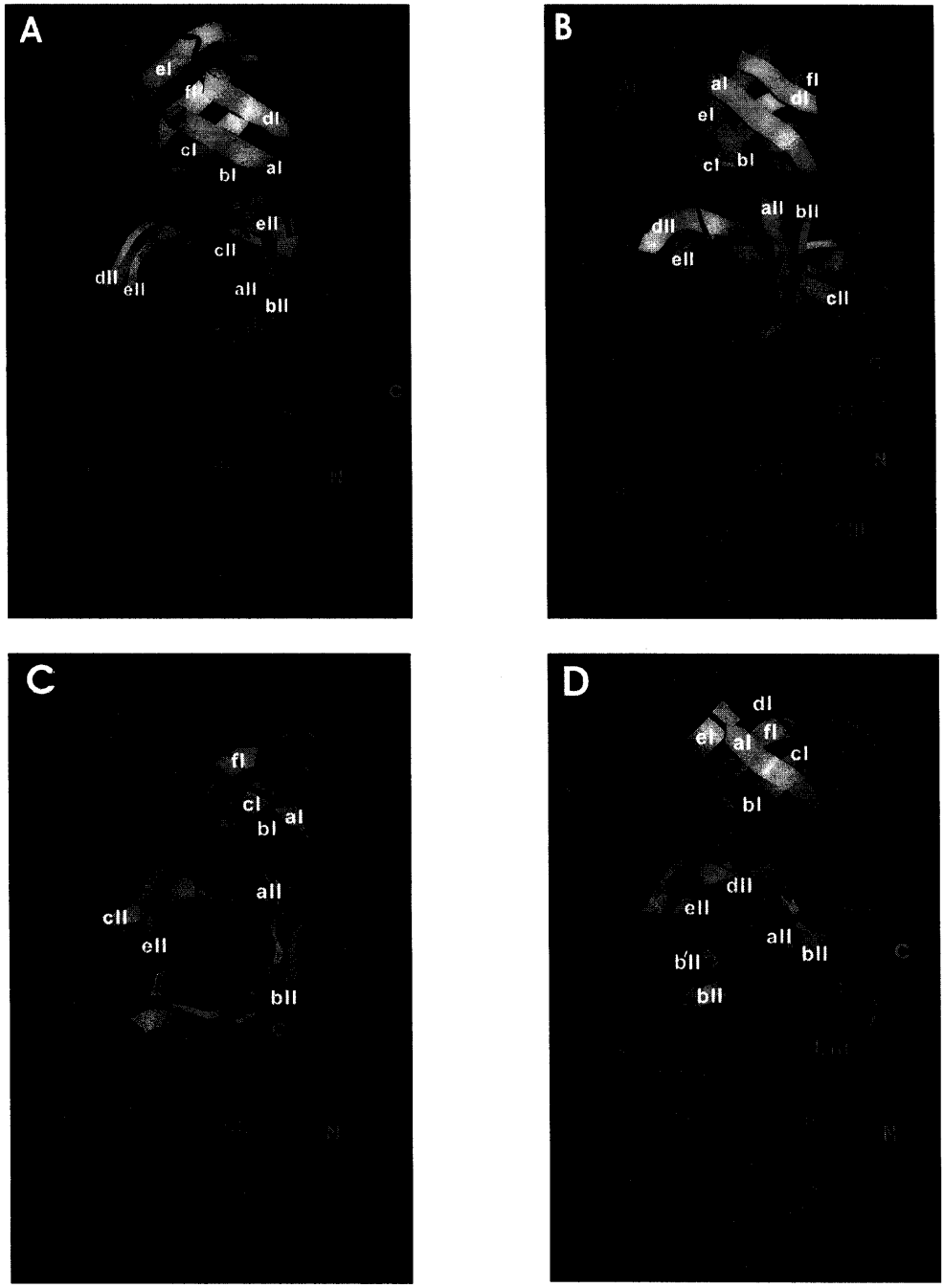


Fig. 1. The x-ray crystallographic structure of (A) $M^{pro}T$ and (B) $M^{pro}H$ and the structural model of (C) $M^{pro}ST$ and (D) $M^{pro}SH$. These structures are visualized by the Insight II program. The N- and C-termini are indicated. α-Helices are shown in red cylinders, while β-strands are illustrated in arrows pointing from N- to C-terminus. The polypeptide backbones belonging to the turn and random coil regions are shown in blue and green, respectively. The general acid-base catalyst His residue and the nucleophilic Cys residue are labeled. The locations of the putative substrate binding subsites S1 and S2 are indicated.

protein in each SCR using the Mutation Matrix module of the Insight II program. Automatic loop building was performed either by database searching[38] or generation through random conformational search.[39] The coordinates at the N- and C-termini of these loops were then automatically assigned. Side chains of the target protein were automatically replaced, preserving the conformations of the model proteins. The side chain conformations were optimized either manually or automatically using a rotamer library.[24] Similar secondary structure motifs were identified by database searching and predicted by DSSP.[40] The bond lengths and torsion angles in the SCRs and loop regions were repaired and relaxed using Homology/Refine/SpliceRepair and Homology/Refine/Relax, respectively. The newly built structures of the target protein were substantially refined to avoid van der Waals radius overlapping, unfavorable atomic distances, and undesirable torsion angles using molecular mechanics and dynamics features in the Discover module.

Molecular dynamics simulations

The crystallographic structures of $M^{pro}H$ and $M^{pro}T$ and the structural models of $M^{pro}SH$ and $M^{pro}ST$ were subjected to energy minimization calculations by steepest descent method with 3,000 iterations followed by Newton-Raphson method with 5,000 iterations to be used as the initial energy-minimized structures for further structural comparison. Each energy-minimized structure was subsequently placed in the center of a lattice with the size of $50 \times 60 \times 85$ $Å^3$ full of 6,222, 5,866, 5,836, and 5,776 water molecules for the system of $M^{pro}H$, $M^{pro}T$, $M^{pro}SH$, and $M^{pro}ST$, respectively. These systems composed of the target protein and water molecules were then equilibrated by performing 20,000 steepest descent minimization and 10 ps dynamics calculations. The explicit image periodic boundary condition (PBC) was used for solvent equilibrium. At the end of explicit image equilibrium, Discover will re-image a molecule whose center of mass has moved out of the lattice in order to maintain the integrity of the lattice with a relatively constant density. A cut-off of 14 Å was used to calculate long-range electrostatic interactions. Finally, 200 ps MD simulation was carried out for each system using the Discover module of Insight II. The temperature and pressure were maintained for each MD simulation by weak coupling the system to a heat bath at 300 K and an external pressure bath at one atmosphere with a coupling constant of 0.5 ps, according to the method described by Berendsen et al.[41] A cut-off radius of 10 Å for the non-bonded interactions was applied to each MD simulation. The time-step of the MD simulations was 1 fs. The trajectories and coordinates of these structures were recorded every 2 ps for further analysis.

Structural analyses

Although some complicated algorithms have been proposed to measure the structural similarity between proteins,[42,43] the root-mean-square deviation (RMSD) remains the simplest one for closely related proteins.[44] For each MD simulation, the RMSDs of the trajectories recorded every 2 ps interval were calculated for the backbone C_α atom of the entire protein, domains I, II, and III, and the substrate binding subsites S1 and S2 during the course of 200 ps MD simulations with reference to the respective starting structure according to Koehi.[45] The RMSDs were calculated after optimal superimposition of the coordinates to remove translational and rotational motion.[46] Secondary structures were predicted based on DSSP,[40] in which pattern recognition of the hydrogen bond was correlated to the geometrical features. The default hydrogen bonding energy criterion of -0.5 kcal/mol was used. Accessible surface areas (ASAs) of the substrate binding subsites S1 and S2 and the linear distance between the sulfur atom of the nucleophilic Cys residue and the $N^{\epsilon 2}$ of the general acid-base catalyst His residue for each structure were also recorded as a function of MD simulation time.

RESULTS AND DISCUSSION

Amino acid sequence alignment

Point, insertion, or deletion mutations that would result in a critical loss of biological functions are less favored by evolution and consequently, functionally and structurally relevant domains tend to be highly conserved across a corresponding protein family. Such conservation can be detected as a pattern of conserved residues that would be unlikely to have occurred by chance. Therefore, an optimal amino acid sequence alignment based on the conserved residues is essential to the success of structural homology. The results of amino acid sequence alignment of $M^{pro}S$ to $M^{pro}T$ and $M^{pro}H$ are given in Fig. 2. There are 301, 300, and 306 residues in $M^{pro}T$, $M^{pro}H$, and $M^{pro}S$, respectively. The residue corresponding to Ala46 in domain I of $M^{pro}S$ and those corresponding to Asp248, Ile249, and Gln273 in domain III of $M^{pro}S$ are missing in both $M^{pro}T$ and $M^{pro}H$. In addition, there are one and two extra residues at the C-terminus of $M^{pro}S$ compared to $M^{pro}T$ and $M^{pro}H$, respectively. There are 33, 33, and 29 totally conserved residues in domains I, II, and III

among these M^{pro}, respectively, indicating that domain III exhibits higher sequence variation among these three domains. It has been predicted earlier that the active site of the coronavirus M^{pro} is similar to those of picornavirus $3C^{pro}$.[16] Both the general acid-base catalyst and the nucleophile residues of these three proteins are totally conserved, where the general acid-base catalyst His41 is located in a highly conserved signature sequence (LNGLWLXDXVXCPR**H**VI) of domain I and the nucleophilic Cys144 for $M^{pro}T$ and $M^{pro}H$ or Cys145 for $M^{pro}S$ is located in the highly conserved signature sequence (TIXGSFXXGX**C**GSXG) of domain II (i.e., Xs indicate the nonconserved residues).

Table 1 lists the percentages of amino acid identity among these proteins. $M^{pro}T$ and $M^{pro}H$ show the highest total amino acid identity (60.80%), whereas $M^{pro}H$ and $M^{pro}S$ exhibit the lowest total amino acid identity (40.19%). In addition, domain II has the highest amino acid identity, whereas domain III shows the lowest amino acid identity among these three proteins. $M^{pro}S$ shows slightly higher amino acid identity to $M^{pro}T$ than $M^{pro}H$, indicating that the structure of $M^{pro}S$ may be more similar to $M^{pro}T$ than $M^{pro}H$. The above results are in good agreement with the previous finding that $M^{pro}S$ shows 40 and 44% sequence identity to $M^{pro}H$ and $M^{pro}T$, respectively.[23] Although the significantly high sequence identity of $M^{pro}S$ to BCoV M^{pro} (49%) and MHV M^{pro} (50%) from group II coronaviruses has allowed Anand et al.[23] to recognize it as an outlier among group II coronaviruses, the low sequence identities between $M^{pro}S$ and $M^{pro}T$ and between $M^{pro}S$ and $M^{pro}H$ from the present study strongly support the results of Marra et al.,[4] in which SARS-CoV was classified as a new group of coronavirus from the analysis of the deduced genome sequence.

Table 1. The amino acid sequence identities among $M^{pro}H$, $M^{pro}T$, and $M^{pro}S$

	Identity (%)			
	Total	Domain I	Domain II	Domain III
$M^{pro}H$ and $M^{pro}T$	60.80	63.44	65.06	55.45
$M^{pro}H$ and $M^{pro}S$	40.19	41.94	45.78	35.64
$M^{pro}T$ and $M^{pro}S$	43.85	44.09	49.40	39.22

The Structural Models of $M^{pro}ST$ and $M^{pro}SH$

Recently, two 3D models of $M^{pro}S$ have been constructed based on the crystallographic structure of $M^{pro}H$[23] and that of $M^{pro}T$,[25] using the homology modeling technique. Virtual screening was further performed employing molecular docking towards both constructed models to identify possible $3CL^{pro}$ inhibitors from small molecular databases for SARS therapy. The level of similarity between $M^{pro}S$ and $M^{pro}T$ as well as $M^{pro}H$ allowed us to construct two structural models for $M^{pro}S$ (denoted as $M^{pro}ST$ and $M^{pro}SH$) by the comparative approach and the results are illustrated in Figs. 1(C) and 1(D). There are three 1- and 2-residue insertions in $M^{pro}S$, relative to both structural templates (Fig. 2); as to be expected, these are all located in loops and do not present a problem in model building. Both $M^{pro}ST$ and $M^{pro}SH$ exhibit three distinct domains, indicating that they adopt similar folds as $M^{pro}T$ and $M^{pro}H$, respectively. However, the secondary structures of both $M^{pro}ST$ and $M^{pro}SH$ predicted according to DSSP[40] were less conserved compared to those of $M^{pro}T$ (Fig. 1(A)) and $M^{pro}H$ (Fig. 1(B)), particularly in do-

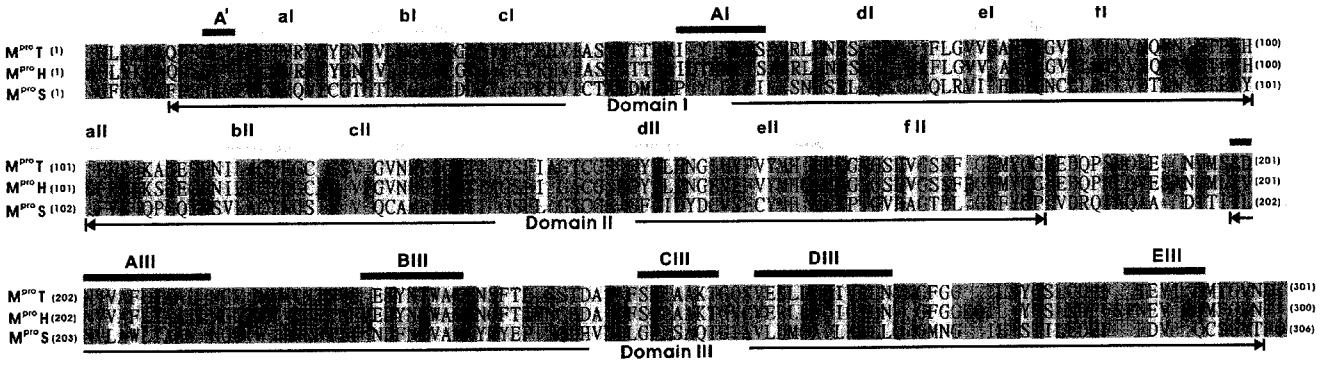


Fig. 2. Amino acid sequence alignment of $M^{pro}T$, $M^{pro}H$, and $M^{pro}S$. Secondary structures as defined in the crystallographic structure of $M^{pro}T$ are shown on top. The start and end amino acid residues are numbered in the brackets on the left and right of each sequence, respectively. Residues totally conserved in all sequences are indicated in red letters with green background. Residues conserved in $M^{pro}T$ and $M^{pro}H$ but different from those in $M^{pro}S$ are represented in black letters with yellow background. Residues where variations occur are given in blue or brown letters with grey background. The amino acid residues missing in both $M^{pro}T$ and $M^{pro}H$ are shown as dashed lines.

main III. The result is consistent with that of amino acid sequence alignment, showing that domain III exhibits the least sequence identity compared to domains I and II among these proteins. It further implies that some of the main-chain or side-chain hydrogen bonds in the constructed homology models may be destroyed in order to maintain folds similar to the model proteins.

The putative substrate binding subsites S1 and S2 of M^{pro}ST and M^{pro}SH are located in a cleft between domains I and II, which are nearly identical to those of M^{pro}T and M^{pro}H (Fig. 1). It indicates that M^{pro}S may follow the similar substrate binding mechanisms of M^{pro}T and M^{pro}H, allowing us to design anti-SARS drugs by screening the known proteinase inhibitors. Instead of separating domains I and II with the catalytic cleft, a long loop (residues 184 to 199 in both M^{pro}T and M^{pro}H and residues 185 to 200 in M^{pro}S) loosely connect domains II and III in all structures. Domain III, a globular cluster of 5, 5, 4, and 2 helices for M^{pro}T, M^{pro}H, M^{pro}ST, and M^{pro}SH, respectively (Fig. 1), has been implicated in the proteolytic activity of M^{pro}.[17] Indeed, there should be only 3 and 1 helices in domain III of M^{pro}ST and M^{pro}SH, respectively, both with helix AIII broken into two parts. Comparing the two crystallographic structures, M^{pro}T and M^{pro}H, and the two homology models, M^{pro}ST and M^{pro}SH, we found that domain I of M^{pro}S is more similar to that of M^{pro}H, while domains II and III of M^{pro}S are more similar to those of M^{pro}T. The low sequence identity and secondary structure similarity in domain III among these proteins presented in the present study, as well as the previous findings showing that the characterization of recombinant proteins, in which 33, 28, and 34 C-terminal amino acid residues of M^{pro} from IBV, MHV, and HCoV, respectively, were deleted resulted in dramatic losses of proteolytic activity, suggest that domain III may play a minor role in proteolytic activity through an undefined mechanism.

The analysis of RMSD (Table 2) shows that the structure of M^{pro}H is very similar to that of M^{pro}T,[26] with the RMSD between these two structures being 2.01 Å for all 300 C_α positions of the molecule. M^{pro}H used in this work lacks two amino acid residues from the C-terminus. Nevertheless, it has the same enzymatic properties as full-length M^{pro}H but yields much superior crystals.[23] In the structure of full-length M^{pro}H, residues 301 and 302 are disordered and not seen in the electron density.[23] With both HCoV 229E and TGEV being group I coronavirus,[47] their M^{pro} share 60.8% sequence identity (Table 1). As shown in Table 2, the RMSDs of the constructed models, M^{pro}SH and M^{pro}ST, are 4.84 and 3.94 Å, compared to their corresponding templates, M^{pro}H and M^{pro}T,

Table 2. The RMSDs between the template proteins, M^{pro}H and M^{pro}T, and the structural models, M^{pro}SH and M^{pro}ST

	RMSD (Å)		
	M^{pro}H	M^{pro}T	M^{pro}SH
M^{pro}T	2.01	-	-
M^{pro}SH	4.51	3.94	-
M^{pro}ST	4.84	4.37	5.78

respectively, while the RMSD between M^{pro}SH and M^{pro}ST is 5.78 Å. It indicates that the structure of M^{pro}S is more similar to that of M^{pro}T than that of M^{pro}H. It further implies that the choice of a more closely related template protein yields a more accurate structural model of M^{pro}S.

Molecular dynamics simulations

The six monomers of M^{pro}T[26] and M^{pro}H[23] presented the asymmetric unit are arranged in three dimmers. Each monomer is folded into three distinct domains, the first two of which are antiparallel β-barrels reminiscent of those found in serine proteinases of the chymotrypsin family. Residues 8-100 form domain I, and residues 101-183 make up domain II (Fig. 2). The connection to domain III is formed by a long loop comprising residues 184-199. Domain III, composed of residues 200-302 (Fig. 2), contains a novel arrangement of 5 α-helices (Fig. 1(A) and (B)). A deep cleft between domains I and II, lined up by hydrophobic residues, constitutes the substrate binding subsites. The catalytic site is situated at the center of the cleft. In order to investigate the dynamics behaviors of M^{pro}T, M^{pro}H, M^{pro}ST, and M^{pro}SH, various MD simulations of these proteins in explicit water were conducted at 300 K. The overall structural changes were evaluated by plotting the RMSDs of these proteins relative to the original positions in the corresponding starting structures as a function of running time as shown in Fig. 3. During the MD time course, these structures remained considerably stable, with the RMSDs remaining within 3 Å. It is obvious that domain III exhibited more structural variations than the other two domains in all cases. The substrate binding subsite S1 of M^{pro}ST was found to maintain its structural integrity during the entire MD time course, whereas the substrate binding subsite S2 of M^{pro}ST exhibited more structural variations. The higher structural variation of S2 makes it flexible enough to accommodate a bulky hydrophobic residue from the substrate.

The secondary structure propensity of these proteins was predicted according to DSSP[40] during the entire MD courses and the results are shown in Fig. 4. As expected, both M^{pro}T and M^{pro}H exhibited higher secondary structure stabil-

ity than M^{pro}ST and M^{pro}SH. The interior of the β-barrel of domain I consists entirely of hydrophobic residues. A short α-helix (Helix AI; Tyr53-Ser58) closes the barrel like a lid.[23,26] However, this short α-helix in both M^{pro}T and M^{pro}H is unstable during the 200 ps MD simulations (Fig. 4(A) and (B)). Furthermore, this short α-helix is missing in M^{pro}ST and M^{pro}SH (Fig. 4(C) and (D)). Domain II is smaller than domain I and also smaller than the homologous domain II of shymotrypsin and hepatitis A virus (HAV) $3C^{pro}$.[48-50] Several secondary structure elements of HAV $3C^{pro}$ are missing in both M^{pro}T and M^{pro}H. Domain III is composed of 5, mostly antiparallel, α-helices and the loops connecting them. Interhelical contacts are mediated by hydrophobic side chains.[26] Database searches[7,51] did not reveal other proteins or protein domains with the same topology as domain III. The present homology models showed that some of the secondary structures of M^{pro}T and M^{pro}H were missing in domain III. It is possible that in order to maintain folds similar to the template proteins, some of the main-chain and side-chain hydrogen bonding patterns of both M^{pro}ST and M^{pro}SH were missing, resulting in the loss of the secondary structure content. Our results again suggest that domain III of these M^{pro} play a role in proteolytic activity through an undefined mechanism regardless of its structural integrity.

Active site

The active site of the coronavirus M^{pro} is similar to those of the piconavirus $3C^{pro}$, as had been predicted previously.[16] The mutual arrangement of the nucleophilic Cys144 and the general acid-base catalyst His41 of M^{pro}T is identical to that of the HAV $3C^{pro}$ Cys172 and His44 and the Ser195 and His57 residues of chymotrypsin.[26] The distance between the sulfur atom of Cys144 and the $N^{\varepsilon 2}$ of His41 in M^{pro}T is 4.05 Å,[26] longer than the corresponding Cys-His distances in HAV $3C^{pro}$ (3.92 Å),[49] poliovirus (PV) $3C^{pro}$ (3.4 Å),[52] and papain (3.65 Å).[53] From a dynamics point of view (Fig. 5), the Csy144-His41 distance of M^{pro}H fluctuated more rapidly than that of M^{pro}T. In addition, the Cys145-His41 distances of

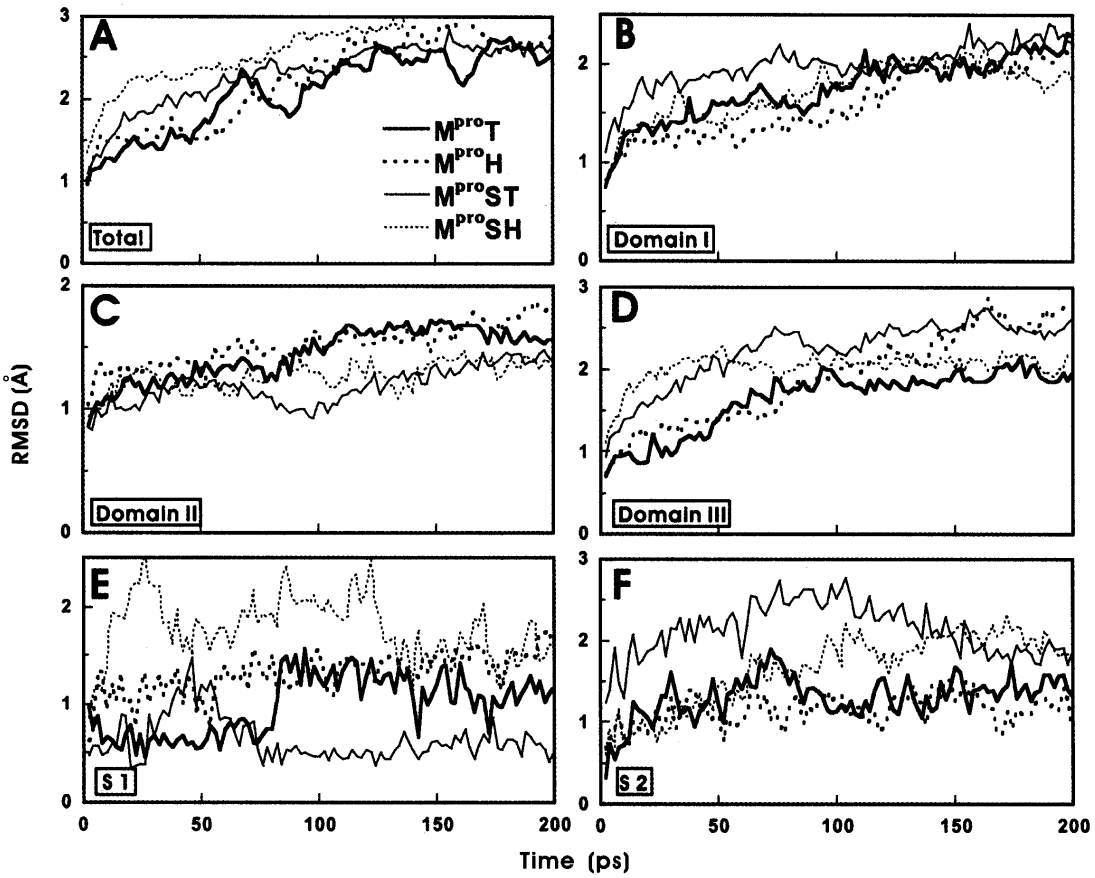


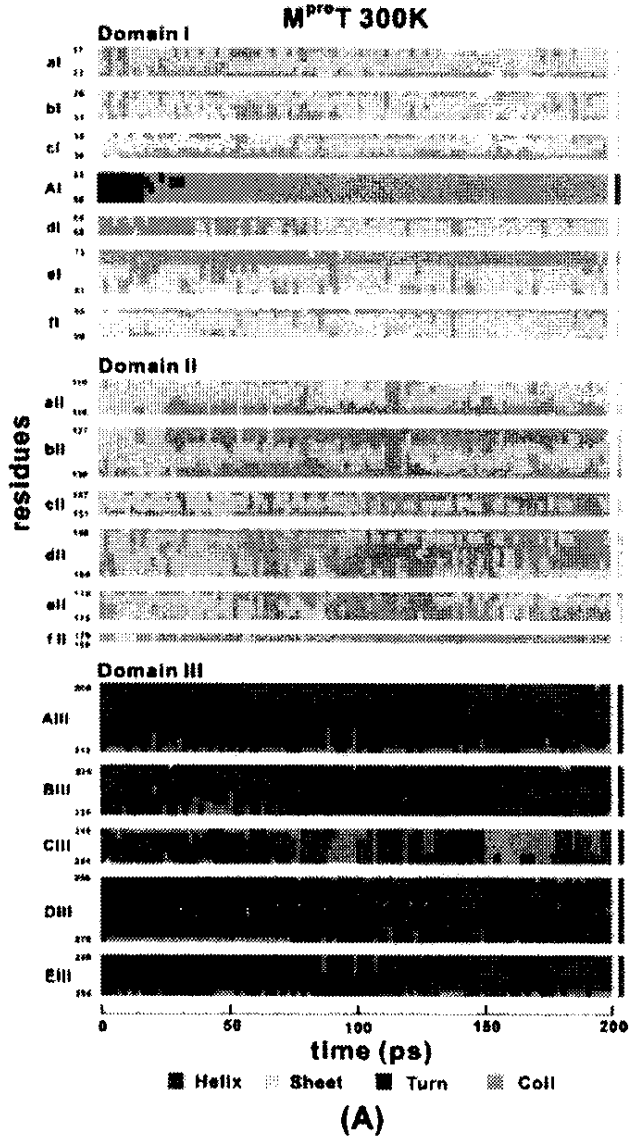
Fig. 3. The RMSDs of the backbone C_α for (A) the whole protein, (B) domain I, (C) domain II, (D) domain III, (E) substrate binding subsite S1, and (F) substrate binding subsite S2 of M^{pro}T, M^{pro}H, M^{pro}ST, and M^{pro}SH with reference to their respective starting structure during the 200 ps MD simulations at 300 K.

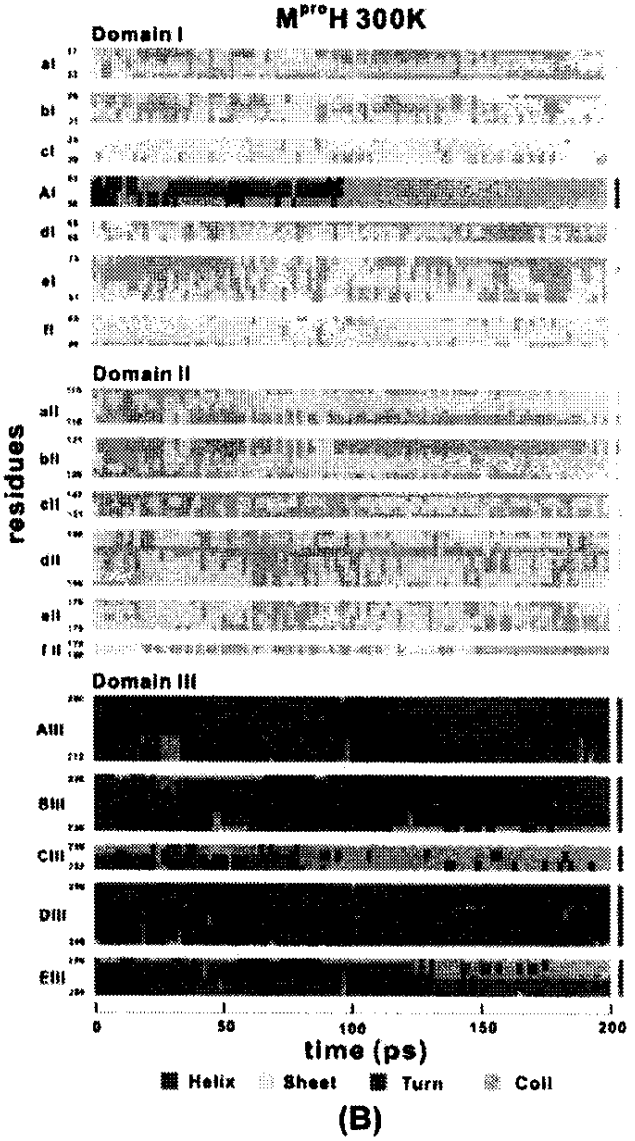
M^{pro}SH fluctuated more rapidly than that of M^{pro}ST beyond 150 ps. These results indicate that both M^{pro}T and M^{pro}ST may exhibit more stable active site configurations than those of M^{pro}S and M^{pro}SH.

It is generally assumed that the native state of the active site of papain-like cystein proteinases is a thiolate-imidazolium ion pair formed by Cys and His residues.[54] In proteinases of the papain family, an Asn residue is the third member of the catalytic triad. Chymotrypsin and other members of this serine proteinase family have a catalytic triad consisting of Ser195-His57-Asp102. In HAV $3C^{pro}$, Asp84 is present at the required position.[49,55] PV $3C^{pro}$, human rhinovirus (HRV) $3C^{pro}$, and HRV $2A^{pro}$ have a Glu or Asp residue in the proper orientation to accept a hydrogen bond from the active site His residue.[52,56,57] Both M^{pro}T and M^{pro}H have Val84 in the corresponding position, with its side chain pointing away from the catalytic site. The corresponding residue in M^{pro}S is Cys85 (Fig. 2). In both M^{pro}T and M^{pro}H, the polypeptide segment 184-199, which connects domains II and III and is probably involved in substrate binding, is held in the proper position during catalysis. The corresponding segment was also found in M^{pro}S, although its amino acid sequence is not conserved compared to those of M^{pro}T and M^{pro}H (Fig. 2). A direct involvement of His163 or Asp186 of M^{pro}T, Gln163 or Asp186 of M^{pro}H, and His164 or Asp187 of M^{pro}S in catalysis, makes them a clear case of viral cystein proteinase employing only a catalytic dyad.[26]

Substrate-binding subsites

It has been shown previously that, similarly to





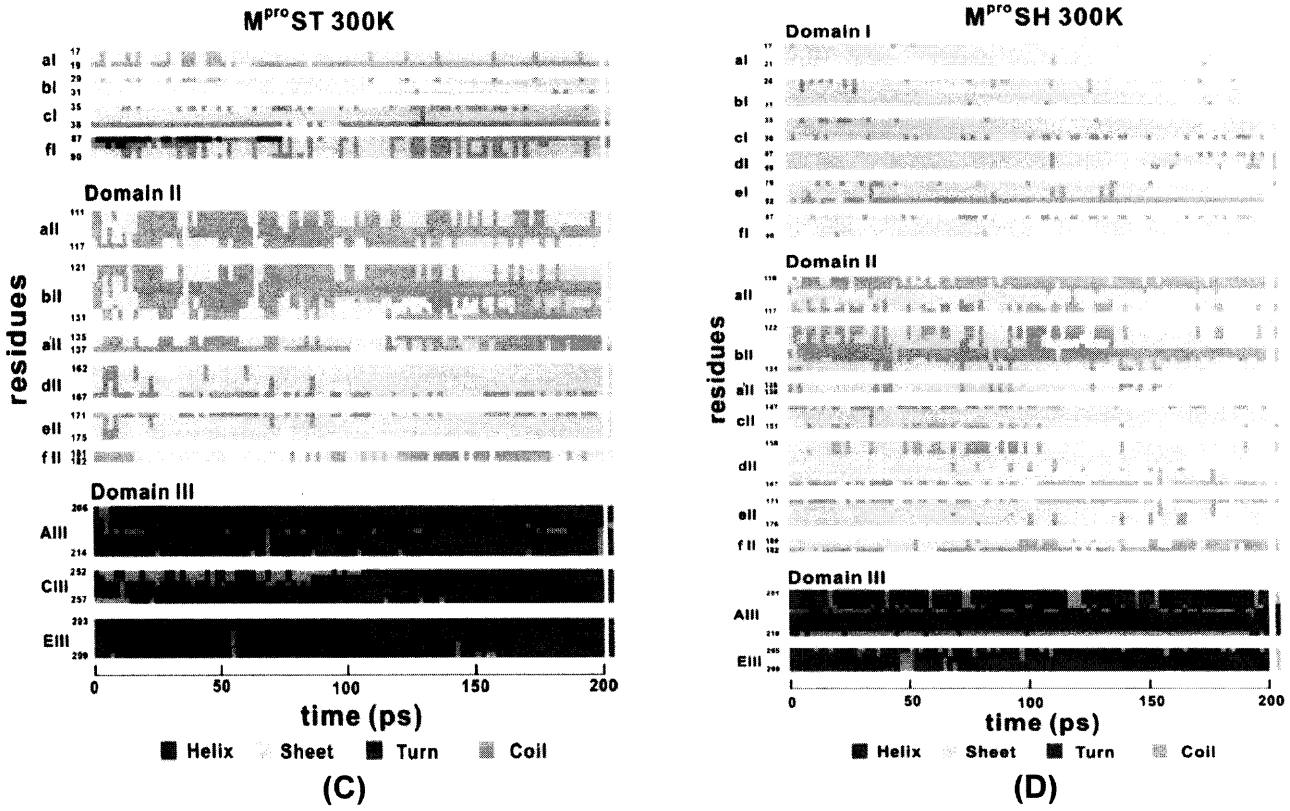


Fig. 4. Secondary structures predicted according to DSSP[40] as a function of MD simulation time for (A) $M^{pro}T$, (B) $M^{pro}H$, (C) $M^{pro}ST$, and (D) $M^{pro}SH$. α-Helix, β-sheet, turn, and coil are shown in red, light yellow, blue, and green, respectively.

$3C^{pro}$,[49,52,56] specific substrate binding by M^{pro} is ensured by well-defined S1 and S2 binding pockets.[26] In addition, it has also been shown that the imidazole side chain of a conserved His residue, which is located in the center of a hydrophobic pocket, interacts with the P1 carboxamide side chain of the substrate. This specific interaction is generally considered to determine the piconavirus $3C^{pro}$ specificity for Gln residue at P1.[49,52,56] The totally conserved His162 of both $M^{pro}T$ and $M^{pro}H$ or His163 of $M^{pro}S$ is located at the very bottom of this hydrophobic pocket which is formed by the totally conserved residues Phe139 of both $M^{pro}T$ and $M^{pro}H$ or Phe140 of $M^{pro}S$ and the main-chain atoms of Ile140, Leu164, Glu165, and His171 of $M^{pro}T$, Ile140, Ile164, Glu165, His171 of $M^{pro}H$, or Leu141, Met165, Glu166, and His172 of $M^{pro}S$. The totally conserved Glu165 of $M^{pro}T$ and $M^{pro}H$ or Glu166 of $M^{pro}S$ forms an ion pair with the totally conserved His171 of $M^{pro}T$ and $M^{pro}H$ or His172 of $M^{pro}S$.[26] This salt bridge is itself on the periphery of these molecules, forming part of the outer wall of the S1 subsite.

Coronavirus M^{pro} has a strong preference for Leu residue at the P2 position.[21] Similar to S1 subsite, the putative S2 subsute identified in the structure is also a hydrophobic pocket that is suitably positioned and large enough to accommodate a Leu side chain easily. In both $M^{pro}T$ and $M^{pro}H$, the S2 pocket is lined by the side chains of His41, Thr47, Ile51, Leu164, and Pro188, despite residue Leu164 in $M^{pro}T$ being

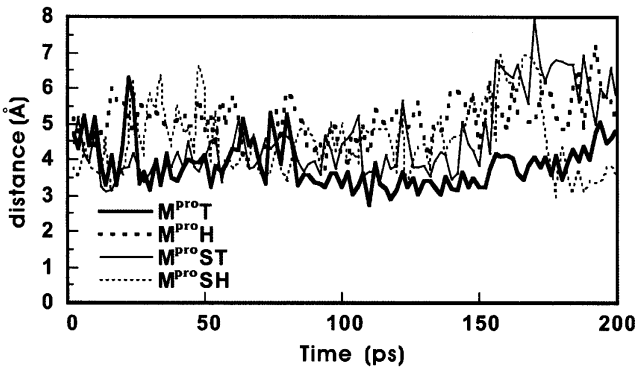


Fig. 5. The linear distance between the sulfur atom of the nucleophilic Cys residue and the $N^{\varepsilon 2}$ of the general acid-base catalyst His residue as a function of MD simulation time for $M^{pro}T$, $M^{pro}H$, $M^{pro}ST$, and $M^{pro}SH$.

replaced by Ile. In $M^{pro}S$, the S2 pocket is lined by the side chains of His41, Asp48, Pro52, Met165, and Gln189. It indicates that the S2 pocket is not as conserved as the S1 pocket, probably because the S2 subsite is located at the open mouth of the catalytic cleft formed by domains I and II, while the S1 subsite is situated in the very bottom of this cleft. Thus, the structural conservation of the substrate binding subsite S1 is necessary to maintain the structural integrity of both domains I and II. It is worth mentioning that the main chain of Leu164 of $M^{pro}T$ (or Ile164 of $M^{pro}H$ or Met165 of $M^{pro}S$) forms part of the S1 subsite while its side chain is involved in the S2 subsite. It indicates that these two subsites are somewhat influenced by each other towards substrate binding.

Fig. 6 shows the ASAs of both S1 and S2 subsites for $M^{pro}T$, $M^{pro}H$, $M^{pro}ST$, and $M^{pro}SH$ as a function of MD simulation time. Both subsites are flexible enough to accommodate the substrates. In order to gain a clearer look of these structures, the snapshots of both S1 and S2 subsites for these proteins with the smallest and largest ASAs sampled from the 200 ps MD simulations are illustrated in Fig. 7. The smallest ASA of S1 is 212, 95.4, 150, and 233 $Å^2$ sampled at 38, 116, 146, and 2 ps, while the largest ASA of S1 is 360, 158, 286, and 361 $Å^2$ sampled at 88, 30, 28, and 94 ps for $M^{pro}T$, $M^{pro}H$, $M^{pro}ST$, and $M^{pro}SH$, respectively. The smallest ASA of S2 is 117, 107, 290, and 143 $Å^2$ sampled at 2, 118, 176, and 4 ps, while the largest ASA of S2 is 344, 217, 461, and 296 $Å^2$ sampled at 158, 196, 26, and 158 ps for $M^{pro}T$, $M^{pro}H$, $M^{pro}ST$, and $M^{pro}SH$, respectively. It is interesting that the sizes and conformations of the smallest and the largest S1 pocket of $M^{pro}SH$ are very similar to those of $M^{pro}T$. The variation of the size and conformation of S2 subsite for these proteins is more significant than the S1 subsite during the MD simulations, probably because S2 is fully exposed to the solvent and is easy to be subjected to structural change. The structural variation of these two subsites allows them to accommodate the specific recognition residues of the substrates upon binding.

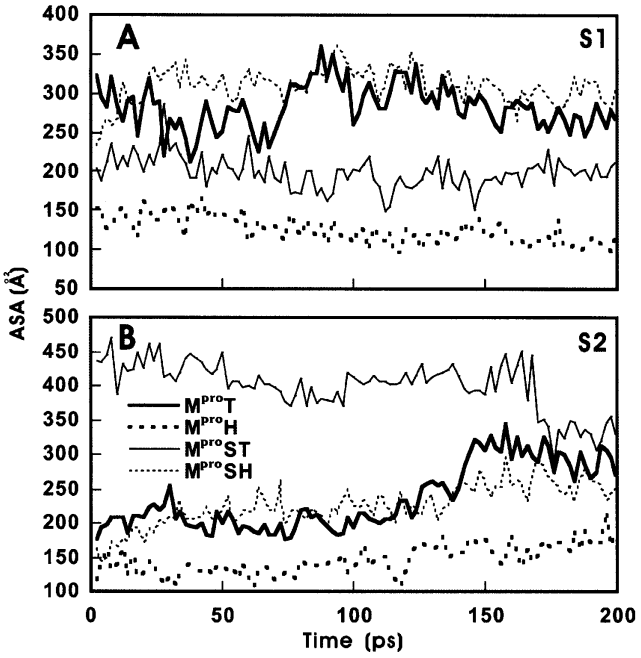


Fig. 6. The ASAs of the substrate binding subsites (A) S1 and (B) S2 as a function of MD simulation time for $M^{pro}T$, $M^{pro}H$, $M^{pro}ST$, and $M^{pro}SH$.

In conclusion, the technique of the comparative approach was successfully applied to construct the homology models of $M^{pro}ST$ and $M^{pro}SH$ based on the crystal structures of $M^{pro}T$ and $M^{pro}H$, respectively, in this study. Molecular dynamics simulations were subsequently conducted to investigate the dynamics behaviors of the structural elements of these structures. Although these structures share many com-

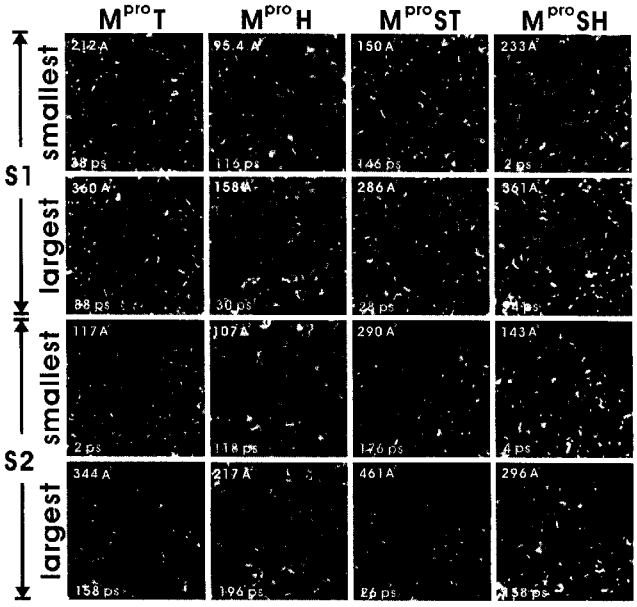


Fig. 7. The snapshots of the substrate binding subsites S1 and S2 for $M^{pro}T$, $M^{pro}H$, $M^{pro}ST$, and $M^{pro}SH$ with the smallest and the largest ASAs during the 200 ps MD simulations. The protein residues are illustrated in CPK with the residues forming these subsites being shown in red. The residues lining up the hydrophobic pockets of subsites S1 and S2 are labeled. The value of the smallest and the largest ASAs for each protein and the time point the structure was sampled are given at the upper-right and the lower-right corners of each frame, respectively.

mon features, the most significant difference occurs at the S2 subsite, where the amino acid residues lining up this subsite are least conserved. It may be a critical challenge for designing anti-SARS drugs by simply screening the known database of proteinase inhibitors.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the financial support from the National Science Council of Taiwan (NSC-92-2214-E-027-001).

Received March 1, 2004.

REFERENCES

1. Drosten, C.; Günther, S.; Preiser, W.; van der Werf, S.; Brodt, H.-R.; Becker, S.; Rabenau, H.; Panning, M.; Kolesnikova, L.; Fouchier, R. A. M.; Berger, A.; Burguière, A.-M.; Cinatl, J.; Eickmann, M.; Escriou, N.; Grywna, K.; Kramme, S.; Manuguerra, J.-C.; Müller, S.; Rickerts, V.; Stürmer, M.; Vieth, S.; Klenk, H.-D.; Osterhaus, A. D. M. E.; Schmitz, H. M. D.; Doer, H. W. *N. Engl. J. Med.* **2003**, *348*, 1967.
2. Ksiazek, T. G.; Erdman, D.; Goldsmith, C. S.; Zaki, S. R.; Peret, T.; Emery, S.; Tong, S.; Urbani, C.; Comer, J. A.; Lim, W.; Rollin, P. E.; Dowell, S. F.; Ling, A.-E.; Humphrey, C. D.; Shieh, W.-J.; Guarner, J.; Paddock, C. D.; Rota, P.; Fields, B.; DeRisi, J.; Yang, J.-Y.; Cox, N.; Hughes, J. M.; LeDuc, J. W.; Bellini, W. J.; Anderson, L. J. *N. Engl. J. Med*. **2003**, *348*, 1953.
3. Lee, N.; Hui, D.; Wu, A.; Chan, P.; Cameron, P.; Joynt, G. M.; Ahuja, A.; Yung, M. Y.; Leung, C. B.; To, K. F.; Lui, S. F.; Szeto, C. C.; Chung, S.; Sung, J. J. Y. *N. Engl. J. Med.* **2003**, *348*, 1986.
4. Marra, M. A.; Jones, S. J. M.; Astell, C. R.; Holt, R. A.; Brooks-Wilson, A.; Butterfield, Y. S. N.; Khattra, J.; Asano, J. K.; Barber, S. A.; Chan, S. Y.; Cloutier, A.; Coughlin, S. M.; Freeman, D.; Girn, N.; Griffith, O. L.; Leach, S. R.; Mayo, M.; McDonald, H.; Montgomery, S. B.; Pandoh, P. K.; Petrescu, A. S.; Robertson, A. G.; Schein, J. E.; Siddiqui, A.; Smailus, D. E.; Stott, J. M.; Yang, G. S. *Science* **2003**, *300*, 1399.
5. Rota, P. A.; Oberste, M. S.; Monroe, S. S.; Nix, W. A.; Compagnoli, R.; Icenogle, J. P.; Peñaranda, S.; Bankamp, B.; Maher, K.; Chen, M.-H.; Tong, S.; Tamin, A.; Lowe, L.; Frace, M.; DeRisi, J. L.; Chen, Q.; Wang, D.; Erdman, D. D.; Peret, T. C. T.; Burns, C.; Ksiazek, T. G.; Rollin, P. E.; Sanchez, A.; Liffick, S.; Holloway, B.; Limor, J.; McCaustland, K.; Olsen-Rassmussen, M.; Fouchier, R.; Günther, S.; Osterhaus, A. D. M. E.; Drosten, C.; Pallansch, M. P.; Anderson, L. J.; Bellini, W. J. *Science* **2003**, *300*, 1394.
6. Enjuanes, L. et al.; van Regenmortal, M. H. V.; Fauqet, C. M.; Bishop, D. H. L.; Carstens, E. B.; Rstes, M. K.; Lemon, S. M.; Mayo, M. A.; McGeoch, D. J.; Pringle, C. R. In *Virus Taxonomy*; Wickner, R. B., Ed.; Academic Press: New York, 2000; pp 835-849.
7. Holm, L.; Sander, C. *J. Mol. Biol*. **1993**, *233*, 123.
8. Lai, M. M. C.; Holmes, K. V. In *Fields Virology*; Knipe, D. M.; Howley, P. M., Ed.; Lippincott Williams and Wilkins: New York, 2001; Chapter 35.
9. Myint, S. H. In *The Coronavirdae*; Siddell, S. G., Ed.; Plenum Press: New York, 1995; p 389.
10. Cavanagh, D. *Arch. Virol.* **1997**, *142*, 629.
11. Den Boon, J. A.; Snijder, E. J.; Chirnside, E. D.; de Vries, A. A.; Horzinek, M. C.; Spaan, W. J. *J. Virol.* **1991**, *65*, 2910.
12. Herold, J.; Raabe, T.; Schelle-Prinz, B.; Siddell, S. G. *Virology* **1993**, *195*, 680.
13. Thiel, V.; Herold, J.; Schelle, B.; Siddell, S. G. *J. Virol.* **2001**, *75*, 6676.
14. Ziebuhr, J.; Herold, J.; Siddell, S. G. *J. Virol.* **1995**, *69*, 4331.
15. Gornalenya, A. E.; Donchenko, A. P.; Blinov, V. M.; Koonin, E. V. *FEBS Lett*. **1989**, *243*, 103.
16. Gornalenya, A. E.; Koonin, E. V.; Donchenko, A. P.; Blinov, V. M. *Nucleic Acids Res.* **1989**, *17*, 4847.
17. Ziebuhr, J.; Heusipp, G.; Siddell, S. G. *J. Virol*. **1997**, *71*, 3992.
18. Hegyi, A.; Friebe, A.; Gorbalenya, A. E.; Ziebuhr, J. *J. Gen. Virol.* **2002**, *83*, 581.
19. Liu, D. X.; Brown, T. D. *Virology* **1995**, *209*, 420.
20. Lu, Y.; Denison, M. R. *Virology* **1997**, *230*, 335.
21. Ziebuhr, J.; Snijder, E. J.; Gorbalenya, A. E. *J. Gen. Virol.* **2000**, *81*, 853.
22. Hegyi, A.; Ziebuhr, J. *J. Gen. Virol*. **2002**, *83*, 595.
23. Anand, K.; Ziebuhr, J.; Wadhwani, P.; Mesters, J. R.; Hilgenfeld, R. *Science* **2003**, *300*, 1763.
24. Ponder, J. W.; Richards, F. M. *J. Mol. Biol*. **1987**, *193*, 775.
25. Xiong, B.; Gui, C.-S.; Xu, X.-Y.; Luo, C.; Chen, J.; Luo, H.-B.; Chen, L.-L.; Li, G.-W.; Sun, T.; Yu, C.-Y.; Yue, L.-D.; Duan, W.-H.; Shen, J.-K.; Qin, L.; Shi, T.-L.; Li, Y.-X.; Chen, K.-X.; Luo, X.-M.; Shen, X.; Shen, J.-H.; Jiang, H.-L. *Acta Pharmacol. Sin*. **2003**, *24*, 497.
26. Anand, K.; Palm, G. J.; Mesters, J. R.; Siddell, S. G.; Ziebuhr, J.; Hilgenfeld, R. *EMBO J.* **2002**, *21*, 3213.
27. Liu, H.-L.; Wang, W.-C. *Chem. Phys. Lett.* **2002**, *366*, 284.
28. Liu, H.-L.; Wang, W.-C. *Protein Eng*. **2003**, *16*, 19.
29. Liu, H.-L.; Wang, W.-C. *J. Biomol. Struct. Dyn.* **2003**, *20*, 615.
30. Liu, H.-L.; Lin, Y.-M. *J. Chin. Chem. Soc.* **2003**, *50*, 799.
31. Liu, H.-L.; Hsu, C.-M. *J. Chin. Chem. Soc.* **2003**, *50*, 1235.
32. Liu, H.-L.; Lin, J.-C. *Chem. Phys. Lett*. **2003**, *381*, 592.
33. Liu, H.-L.; Lin, J.-C. *J. Biomol. Struct. Dyn.* **2004**, *21*, 639.

34. Hwang, M.-J.; Ni, X.; Waldman, M.; Ewig, C. S.; Hagler, A. T. *Biopolymers* **1998**, *45*, 435.
35. Maple, J. R.; Hwang, M.-J.; Jalkanen, K. J.; Stockfisch, T. P.; Hagler, A. T. *J. Comp. Chem*. **1998**, *19*, 430.
36. Peng, Z.; Ewig, C. S.; Hwang, M.-J.; Waldman, M.; Hagler, A. T. *J. Phys. Chem. A* **1997**, *101*, 7243.
37. Greer, J. *Proteins* **1990**, *7*, 317.
38. Schuler, G. D.; Altschul, S. F.; Lipman, D. J. *Proteins: Struct. Funct. Genet.* **1991**, *9*, 180.
39. Shenkin, P. S.; Yarmush, D. L.; Fine, R. M.; Wang, H.; Levinthal, C. *Biopolymers* **1987**, *26*, 2053.
40. Kabsch, W.; Sander, C. *Biopolymers* **1983**, *22*, 2577.
41. Berendsen, H. J. C.; Postma, J. P. M.; van Gunsteren, W. F.; DiNola, A.; Haak, J. R. *J. Comp. Phys*. **1984**, *81*, 3684.
42. Lesk, A. M. *Proteins* **1998**, *33*, 320.
43. Levitt, M.; Gerstein, M. *Proc. Natl. Acad. Sci. USA* **1998**, *95*, 5913.
44. Mizuguchi, K.; Go, N. *Curr. Opin. Struct. Biol.* **1995**, *5*, 377.
45. Koehi, P. *Curr. Opin. Struct. Biol.* **2001**, *11*, 348.
46. Kabsch, W. *Acta Cryst. A* **1976**, *32*, 922.
47. Siddell, S. G. In *The Coronavirdae*; Siddell, S. G., Ed.; Plenum Press: New York, 1995; p 1.
48. Allaire, M.; Chernaia, M. M.; Malcolm, B. A.; James, M. N. *Nature* **1994**, *369*, 72.
49. Bergmann, E. M.; Mosimann, S. C.; Chernaia, M. M.; Malcolm, B. A.; James, M. N. *J. Virol.* **1997**, *72*, 2436.
50. Tsukada, H.; Blow, D. M. *J. Mol. Biol.* **1985**, *184*, 703.
51. Gilbert, D.; Westhead, D.; Nagano, N.; Thornton, J. *Bioinformatics* **1999**, *15*, 317.
52. Mosimann, S. C.; Cherney, M. M.; Sia, S.; Plotch, S.; James, M. N. *J. Mol. Biol.* **1997**, *273*, 1032.
53. Kamphuis, I. G.; Kalk, K. H.; Swarte, M. B.; Drenth, J. *J. Mol. Biol*. **1984**, *179*, 233.
54. Polgár, L. *FEBS Lett*. **1974**, *47*, 15.
55. Malcolm, B. A. *Protein Sci.* **1995**, *4*, 1439.
56. Matthews, D. A. et al. *Cell* **1994**, *77*, 761.
57. Petersen, J. F.; Cherney, M. M.; Liebig, H. D.; Skern, T.; Kuechler, E.; James, M. N. *EMBO J.* **1999**, *18*, 5463.