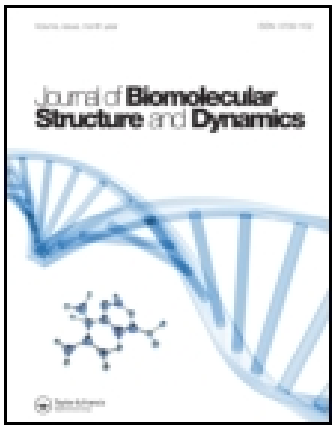

Journal of Biomolecular Structure and Dynamics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/tbsd20

Molecular Dynamics Simulations of Various Coronavirus Main Proteinases

Hsuan-Liang Liu [a] , Jin-Chung Lin [a] , Yih Ho [b] , Wei-Chan Hsieh [a] , Chin-Wen Chen [a] & Yuan-Chen Su [a]

[a] Department of Chemical Engineering , Graduate Institute of Biotechnology National Taipei University of Technology , 1 Section 3 Chung-Hsiao East Road, Taipei , Taiwan , 10608

[b] School of Pharmacy Taipei Medical University , 250 Wu-Hsing Street, Taipei , Taiwan , 110

Published online: 15 May 2012.

To cite this article: Hsuan-Liang Liu , Jin-Chung Lin , Yih Ho , Wei-Chan Hsieh , Chin-Wen Chen & Yuan-Chen Su (2004) Molecular Dynamics Simulations of Various Coronavirus Main Proteinases, Journal of Biomolecular Structure and Dynamics, 22:1, 65-77, DOI: 10.1080/07391102.2004.10506982

To link to this article: http://dx.doi.org/10.1080/07391102.2004.10506982

Journal of Biomolecular Structure & Dynamics, ISSN 0739-1102
Volume 22, Issue Number 1, (2004)

Molecular Dynamics Simulations of Various Coronavirus Main Proteinases

http://www.jbsdonline.com

Hsuan-Liang Liu[1]*
Jin-Chung Lin[1]
Yih Ho[2]
Wei-Chan Hsieh[1]
Chin-Wen Chen[1]
Yuan-Chen Su[1]

[1]Department of Chemical Engineering
and Graduate Institute of Biotechnology
National Taipei University of Technology
1 Section 3 Chung-Hsiao East Road
Taipei, Taiwan 10608
[2]School of Pharmacy
Taipei Medical University
250 Wu-Hsing Street
Taipei, Taiwan 110

Phone: +886-2-27712171 ext. 2542
Fax: +886-2-27317117
Email: f10894@ntut.edu.tw

Abstract

In this study, two homology models (denoted as $M^{pro}ST$ and $M^{pro}SH$) of main proteinase (M^{pro}) from the novel coronavirus associated with severe acute respiratory syndrome (SARS-CoV) were constructed based on the crystal structures of M^{pro} from transmissible gastroenteritis coronavirus (TGEV) ($M^{pro}T$) and human coronavirus HcoV-229E ($M^{pro}H$), respectively. Both $M^{pro}ST$ and $M^{pro}SH$ exhibit similar folds as their respective template proteins. These homology models reveal three distinct functional domains as well as an intervening loop connecting domains II and III as found in both template proteins. A catalytic cleft containing the substrate binding sites S1 and S2 between domains I and II are also observed. S2 undergoes more significant structural fluctuation than S1 during the 400 ps molecular dynamics simulations because it is located at the open mouth of the catalytic cleft, while S1 is situated in the very bottom of this cleft. The thermal unfolding of these proteins begins at domain III, where the structure is least conserved among these proteins. M^{pro} may still maintain its proteolytic activity while it is partially unfolded. The electrostatic interaction between Arg40 and Asp186 plays an important role in maintaining the structural integrity of both S1 and S2.

Key words: Homology, Main Proteinase, Coronavirus, Severe acute respiratory syndrome (SARS), Substrate binding site, Molecular dynamics simulations.

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 of SARS is usually characterized by high fever, malaise, rigor, headache, nonproductive cough and may progress to generalized, interstitial infiltrates in the lung (3). The sequence of the complete genome of the coronavirus associated with SARS (SARS-CoV) has been determined and characterized with two different isolates (4, 5). Phylogenetic analyses and sequence comparisons have further revealed that SARS-CoV is not closely related to any of the three groups of coronaviruses.

Coronaviruses belong to a diverse group of positive-stranded RNA viruses featuring the largest viral RNA genomes. They share a similar genome organization and common transcriptional and translational processes as *Arteriviridae* (6, 7). The

Abbreviations: $3CL^{pro}$: 3C-like proteinase; 3D: Three-dimensional; ASA: Accessible surface area; CVFF: Consistent valence force field; DSSP: Dictionary of secondary structure pattern; MD: Molecular dynamics; M^{pro}: Main proteinase; $M^{pro}H$: Main proteinase of human coronavirus HcoV-229E; $M^{pro}S$: Main proteinase of SARS-CoV; $M^{pro}SH$: Homology model of $M^{pro}S$ based on the crystal structure of $M^{pro}H$; $M^{pro}ST$: Homology model of $M^{pro}S$ based on the crystal structure of $M^{pro}T$; $M^{pro}T$: Main proteinase of TGEV; PCB: Periodic boundary condition; RMSD: Root-mean-square deviation; SARS: Severe acute respiratory syndrome; SARS-CoV: Coronavirus associated with SARS; SCR: Structural conserved region; TGEV: Transmissible gastroenteritis coronavirus

human coronavirus HcoV-229E replicase gene encodes two overlapping polyproteins, pp1a and pp1ab (8), that mediate all the functions required for viral replication and transcription (9). 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}) (10). 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 (11, 12). M^{pro} from HcoV-229E ($M^{pro}H$) has been biosynthesized in *Escherichia coli* and the enzyme properties, inhibitor profile, and substrate specificity of the purified protein have been well characterized (10, 13).

Recently, the crystal structures of $M^{pro}H$ (14) and M^{pro} from porcine coronavirus (transmissible gastroenteritis virus, TGEV) ($M^{pro}T$) complexed with its inhibitor (15) have been determined. In addition, homology models of $M^{pro}S$ based on the crystal structures of $M^{pro}H$ (14) and $M^{pro}T$ (16, 17) have been also constructed. 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 $M^{pro}T$ with the recombinant $M^{pro}S$ (14). In addition, $M^{pro}S$ exhibits 40 and 44% sequence identity to $M^{pro}H$ and $M^{pro}T$, respectively (14).

Molecular dynamics (MD) simulations in the atomic level have been intensively preformed to gain insight into protein unfolding from its native state induced by raising the temperature (18-20), changing the solvent (21) or increasing the pressure (22). Usually, temperatures in the range of 400 to 600K are employed. According to the Arrhenius equation, the unfolding rate is expected to be approximately 10^3-, 10^6-, 10^9-folds faster than it is observed experimentally when the temperature is increased by 100, 200, and 300 °C, respectively (23). Daggett and Levitt (24) have shown that the results obtained from the MD simulations of protein unfolding induced by increasing the temperature should be reliable by comparing their results to the pH induced denaturation of barnase (25). Previously, several MD simulations, homology modeling, and molecular docking experiments have been successfully conducted towards various target proteins in our group (26-35). In this paper, two homology models of M^{pro} from SARS-CoV ($M^{pro}S$) were constructed (denoted as $M^{pro}ST$ and $M^{pro}SH$) based on the crystal structures of $M^{pro}T$ (15) and $M^{pro}H$ (14), respectively. Subsequently, MD simulations associated with temperature jump technique were conducted to investigate the structure variations of these proteins. Beyond the continued characterization of M^{pro} from various coronaviruses, the amino acid sequence alignment, structural homology analyses, and MD simulations of $M^{pro}S$ presented in this study shall provide particularly attractive targets for further structure-based design of anti-SARS drugs.

Methods

Model Proteins

Two homology models of $M^{pro}S$ ($M^{pro}ST$ and $M^{pro}SH$) were constructed based on the monomer of the three-dimensional (3D) structure of $M^{pro}T$ refined to 1.96 Å resolution (15) (Fig. 1A) and $M^{pro}H$ solved at 2.54 Å resolution (14) (Fig. 1B), which were 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 crystal structure of $M^{pro}T$ before being used as the template. Unfavorable nonphysical contacts in these structures were then eliminated using Biopolymer module of Insight II program (Accelyrs, San Diego, CA, USA) with the force field Discover CVFF (consistent valence force field) (36-38) in the SGI O200 workstation with 64-bit HIPS RISC R12000 2 × 270 MHz CPU and PMC-Sierra RM7000A 350MHz processor (Silicon Graphics, Inc., Mountain View, CA, USA), followed by 10,000 energy mini-

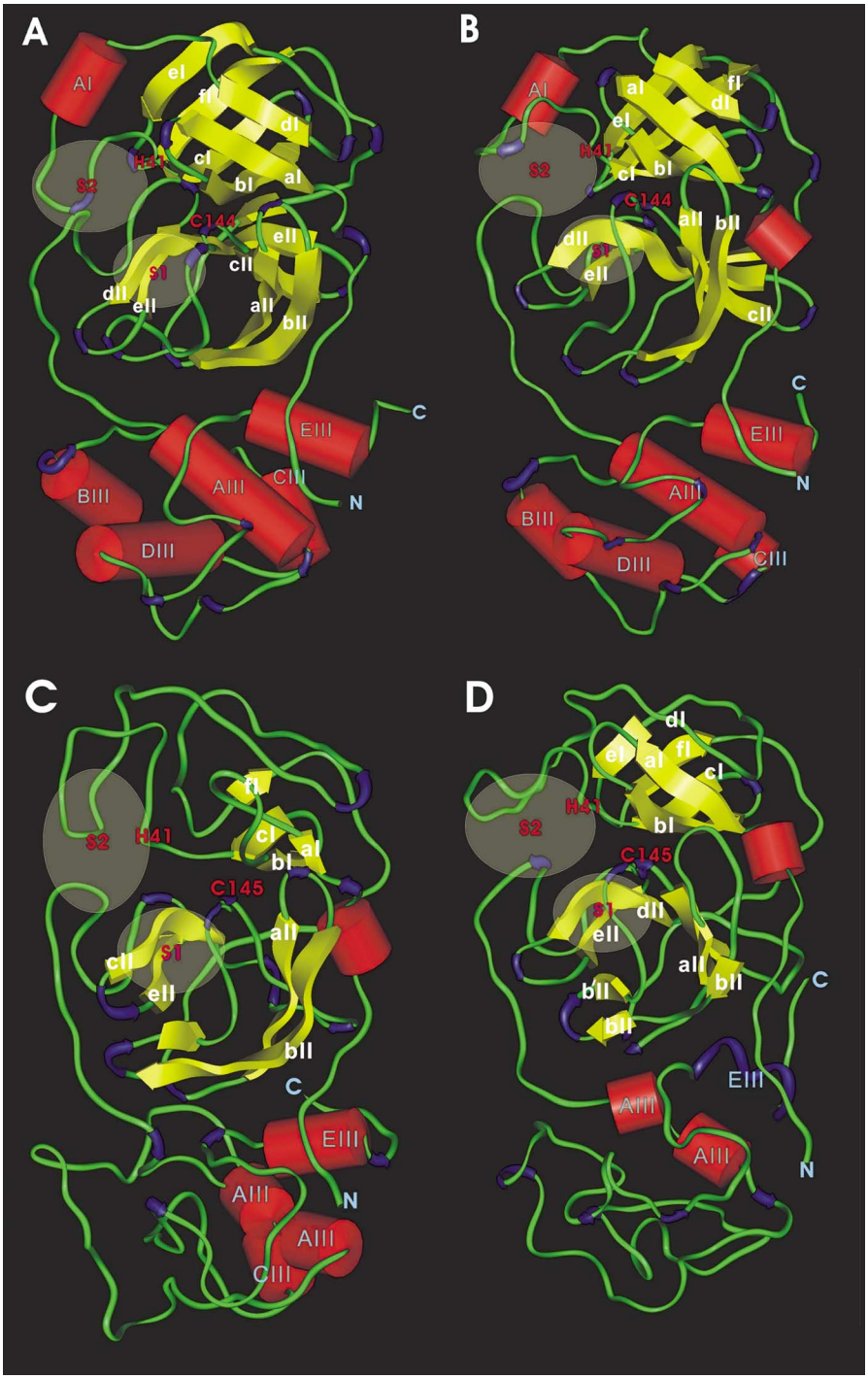

Figure 1: The crystal structure of (**A**) M^proT (15) and (**B**) M^proH (14) and the homology model of (**C**) M^proST and (**D**) M^proSH. These structures are visualized by Insight II program. The N- and C-termini are indicated. Secondary structure elements are labeled as in Table I. α-Helices are 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 sites S1 and S2 are indicated.

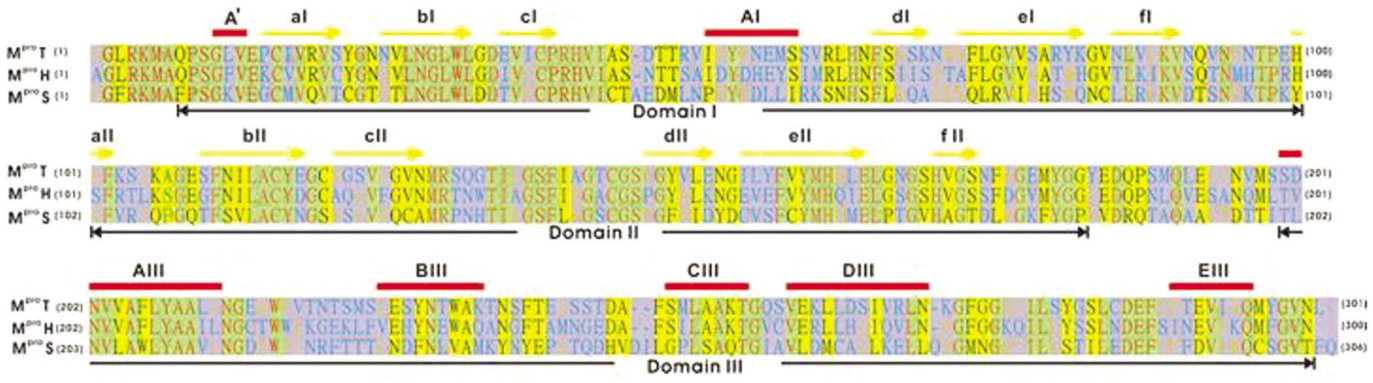

Figure 2: Amino acid sequence alignment of M^proT, M^proH, and M^proS. Secondary structures as defined in the crystallographic structure of M^proT (15) are shown on top. The start and end amino acid residues are numbered in the brackets on the left and right of each sequence, respectively. Residues totally conserved in all sequences are indicated in red letters with green background. Residues conserved in M^proT and M^proH but different from those in M^proS 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^proT and M^proH are shown as dashed lines.

Table I
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

mization 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 template 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 (39). Amino acid sequences of $M^{pro}H$ (Accession Q05002) (40), $M^{pro}T$ (NC_002306.2) (41), and $M^{pro}S$ (NC_004718.3) (4) were obtained either from Swiss-Prot or NCBI database. Smith-Waterman pairwise amino acid sequence alignments were performed based on the conserved structural features among M^{pro} from various coronaviruses to find the location of the active site and substrate binding sites S1 and S2 of $M^{pro}S$. The consensus structural conserved regions (SCRs) of $M^{pro}S$ were generated from alignments of the target protein to the template proteins. The atomic coordinates were then transferred from the template proteins to $M^{pro}S$ in each SCR using Mutation Matrix module of the Insight II program. Automatic loop building was performed either by database searching (42) or generation through random conformational search (43). The coordinates at the N- and C-termini of these loops were then automatically assigned. Side chains of $M^{pro}S$ were automatically replaced, preserving the conformations of the template proteins. The side chain conformations were optimized either manually or automatically using a rotamer library (44). Secondary structure motifs were identified by database searching and defined by DSSP (45). 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 $M^{pro}S$ were substantially refined to avoid van der Waals radius overlapping, unfavorable atomic distances, and undesirable torsion angles using molecular mechanics and dynamics features in Discover module.

Molecular Dynamics Simulations

The crystal structures of $M^{pro}H$ and $M^{pro}T$ and the homology 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 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 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. Finally, 400 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, 400, and 600 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*. (46). Cut-off radius of 13 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 analyses.

Structural Analyses

Although some complicated algorithms have been proposed to measure the structural similarity between proteins (47, 48), the root-mean-square deviation (RMSD) remains the simplest one for closely related proteins (49). 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, the substrate binding sites S1 and S2, and domains I, II, and III during the course of 400 ps MD simulations with reference to the respective starting structure according to Koehi (50). The RMSDs were calculated after optimal superimposition of the coordinates to remove translational and rotational motion (51). Secondary structures were assigned based on DSSP (45), in which pattern recognition of hydrogen bond was correlated to the geomet-

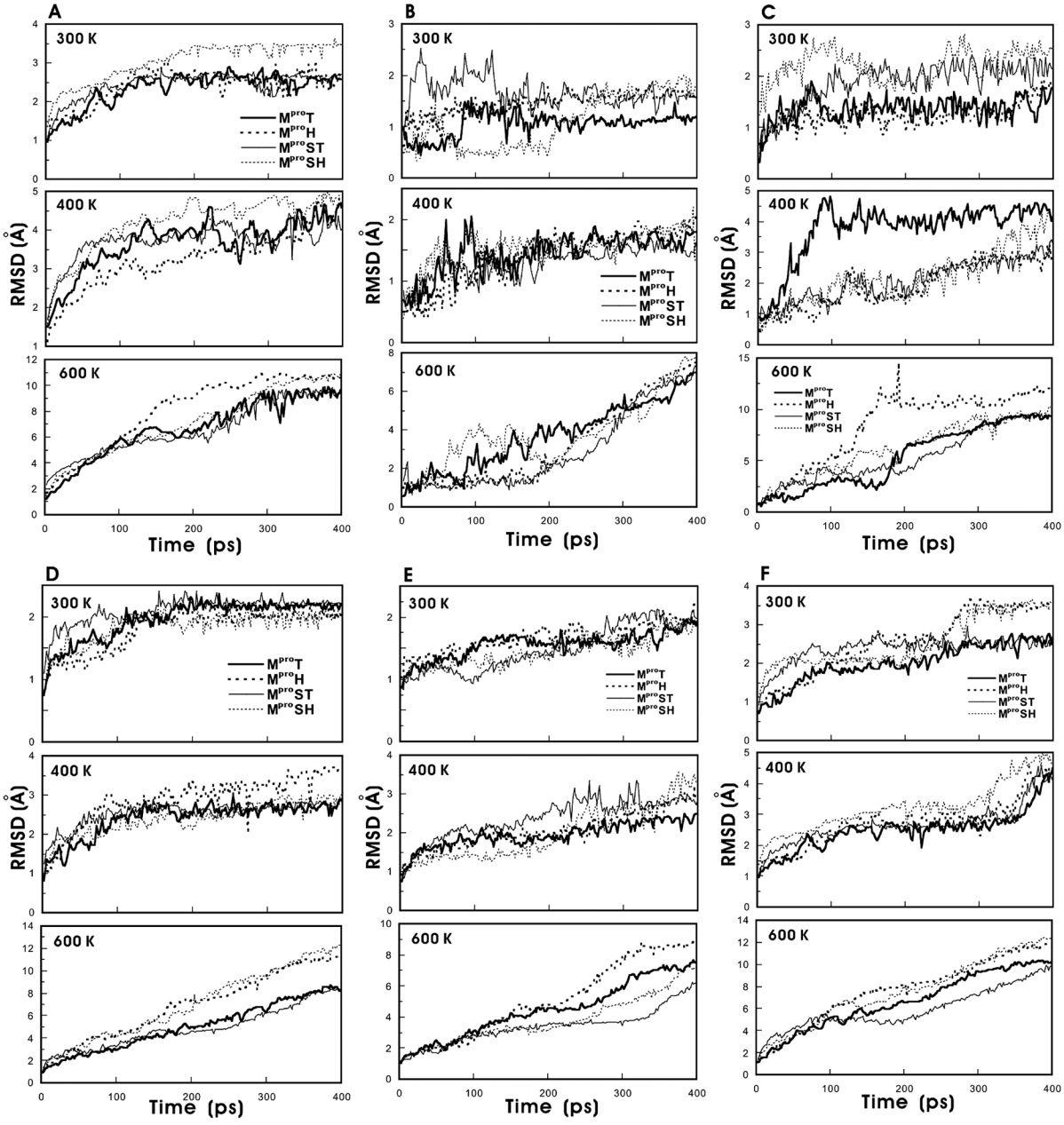

Figure 3: The RMSDs of the backbone C_α for (**A**) the entire protein, (**B**) substrate binding site S1, (**C**) substrate binding site S1, (**D**) domain I, (**E**) domain II, and (**F**) domain III of M^{pro}T, M^{pro}H, M^{pro}ST, and M^{pro}SH with reference to their respective starting structure during the 400 ps MD simulations at 300, 400, and 600 K.

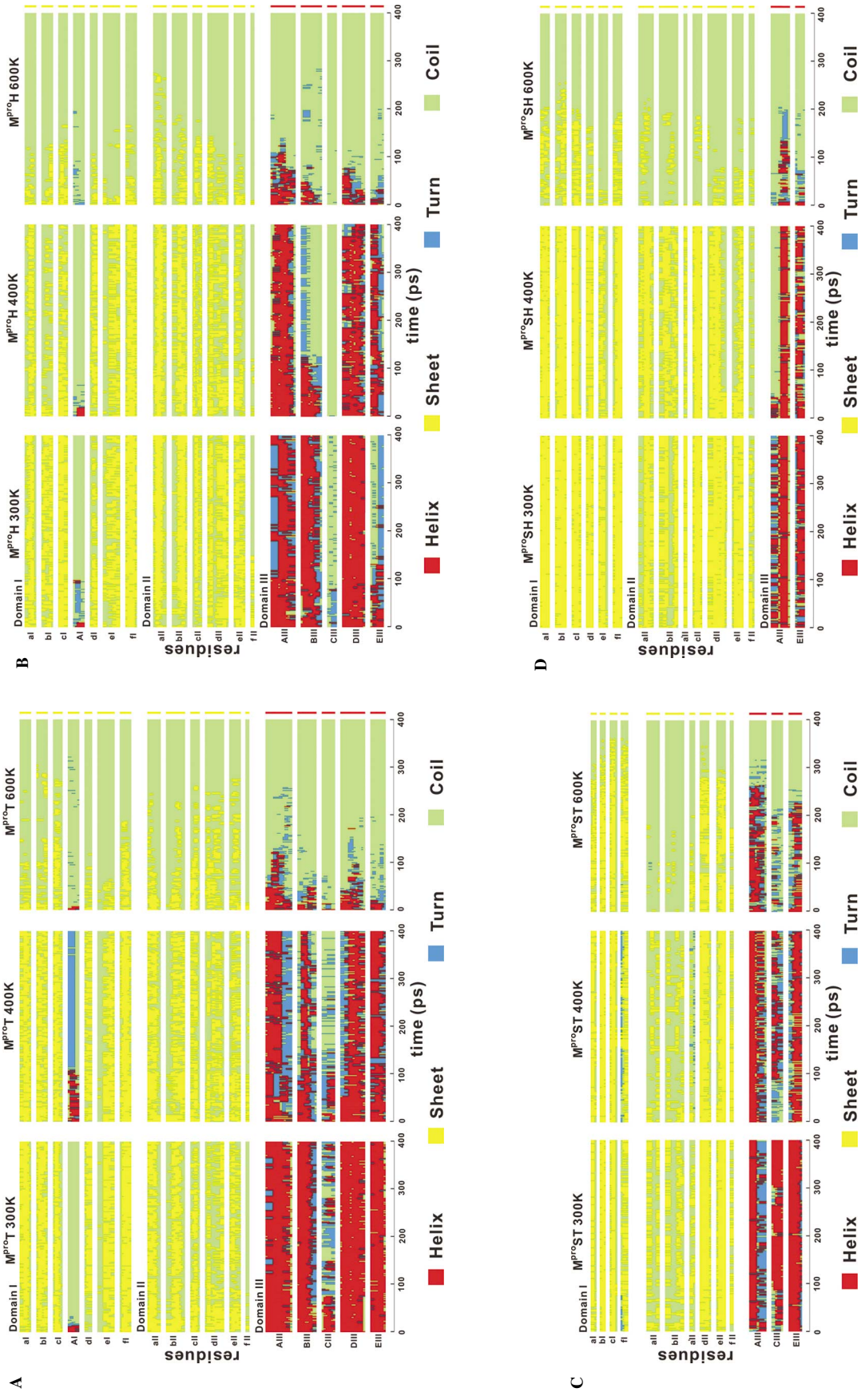

Figure 4: Secondary structures predicted according to DSSP (45) as a function of MD simulation time for (**A**) $M^{pro}T$, (**B**) $M^{pro}H$, (**C**) $M^{pro}ST$, and (**D**) $M^{pro}SH$. α-Helix, β-sheet, turn, and coil are shown in red, light yellow, blue, and green, respectively.

rical features. The default hydrogen bonding energy criterion of -0.5 kcal/mol was used. Accessible surface areas (ASAs) of the substrate binding sites S1 and S2 and the distances between the sulfur atom of the nucleophilic Cys residue and the $N^{\varepsilon 2}$ of the general acid-base catalyst His residue and between the C^{ε} atom of the totally conserved Arg40 from S2 and the C^{γ} atom of the totally conserved Asp186 from the extended loop connecting domains II and III (numbered as in M^{pro}T) for each structure were also recorded as a function of MD simulation time. The average secondary structure content was defined as the ratio of the number of the residual H bonds at time t to the number of the total H bonds in the starting structure.

Results and Discussion

The Homology Models of M^{pro}ST and M^{pro}SH

Usually, an optimal amino acid sequence alignment based on the conserved structural regions is essential to the success of homology modeling. The results of pairwise amino acid sequence alignment of M^{pro}T, M^{pro}H, and M^{pro}S are given in Figure 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 comparing to M^{pro}T and M^{pro}H, respectively. Both the general acid-base catalyst and the nucleophile residue of these three proteins are totally conserved, with the general acid-base catalyst His41 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 in the highly conserved signature sequence (TIXGSFXXGX**C**GSXG) of domain II (i.e., Xs indicate the nonconserved residues). The results of amino acid sequence identity among these three proteins are summarized in Table I. M^{pro}T and M^{pro}H show the highest amino acid identity (60.80 %), whereas M^{pro}H and M^{pro}S exhibit the lowest amino acid identity (40.19 %). 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. Comparing the three domains among these three proteins, domain II has the highest amino acid identity, whereas domain III shows the lowest amino acid identity.

The level of similarity between M^{pro}S and M^{pro}T as well as between M^{pro}S and M^{pro}H allowed us to construct two homology models for M^{pro}S (denoted as M^{pro}ST and M^{pro}SH) by comparative approach and the results are illustrated in Figure 1C and D. The quality of the geometry and of the stereochemistry of the protein structures was validated using Homology/ProStat/Struct_Check commend of Insight II program. A total of 97 and 96% of the backbone dihedral angle (φ and ϕ) densities are located within the structurally favorable regions in Ramachandran plot for M^{pro}ST and M^{pro}SH, respectively (data not shown). The calculation of main chain torsion angles (χ_1 and χ_2) of these proteins showed no severe distorsion of the backbone geometry. In addition, all bond lengths and angles for both homology models are located within the reasonable regions. Besides, the homology models of M^{pro}ST and M^{pro}SH constructed in this work are very similar to the 3D models proposed by Lee *et al*. (16) and Aland *et al*. (14), respectively. The above evidences indicate that the quality of these homology models should be reliable.

The results of homology modeling show that 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 (45) are less conserved comparing to those of M^{pro}T (Fig. 1A) and M^{pro}H (Fig. 1B), particularly in domain III. It is consistent with the results of amino acid sequence alignment, showing that domain III exhibits the least sequence identity comparing to domains I and II among these proteins. Instead of separating domains I and II with a catalytic cleft, domains II and III are loosely con-

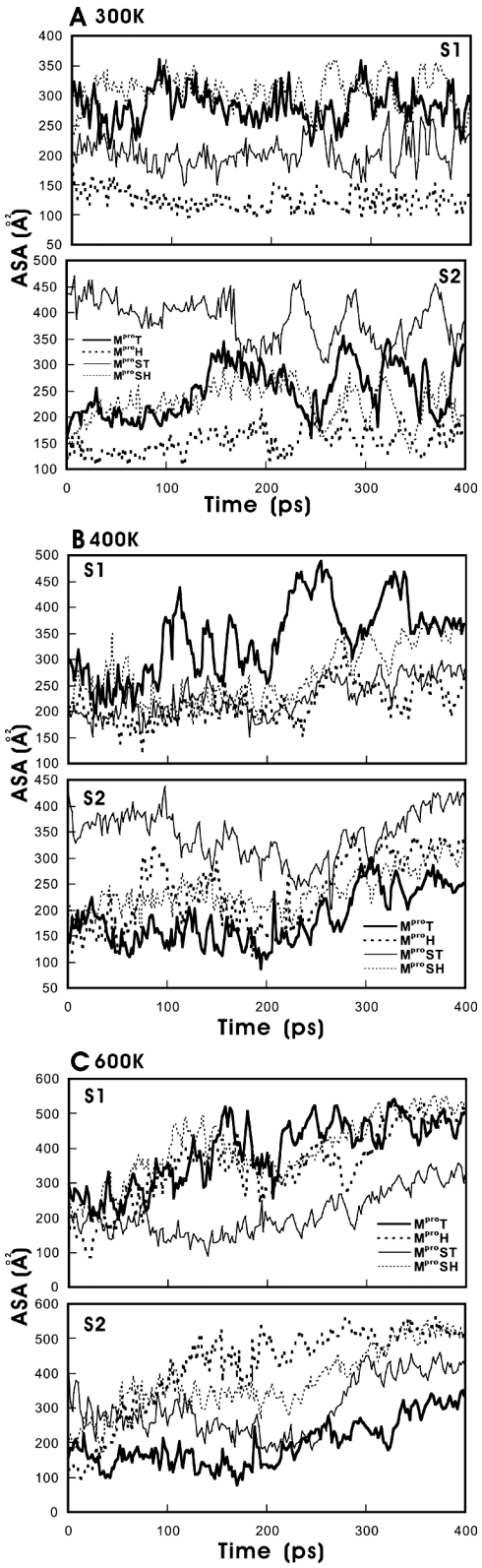

Figure 5: The ASAs of the substrate binding sites S1 and S2 at (**A**) 300, (**B**) 400, and (**C**) 600 K as a function of MD simulation time for M^{pro}T, M^{pro}H, M^{pro}ST, and M^{pro}SH.

nected by a long loop (residues 184-199 in both M^{pro}T and M^{pro}H and residues 185-200 in M^{pro}S) in all structures. Although showing the least structural identity, 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} (13). Comparing the two crystal 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 (13). The putative substrate binding sites S1 and S2 of M^{pro}ST and M^{pro}SH are also located in a catalytic cleft between domains I and II (Fig. 1C and D), which are nearly identical to those of M^{pro}T and M^{pro}H (Fig. 1A and B). 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. A good example has been given by Anand *et al*. (14). They proposed a 3D structural model of M^{pro}S based on the crystal structure of M^{pro}H and further recommended the use a rhinovirus inhibitor (codename AG7088), which is already in clinical trials as anti-common cold drug, as the potential model compound for the design of anti-SARS drugs. In addition, Lee *et al*. (16) have docked 16 available antiviral drugs from the NCI database to the structural model of M^{pro}S and detected that four of them with trade-names Nevirapine, Glycovir, Virazole, and Calanolide A fit well at the substrate binding cleft of there 3D model of M^{pro}S.

Molecular Dynamics Simulations

The structural changes of the whole protein, substrate binding sites S1 and S2, and domains I, II and III for M^{pro}T, M^{pro}H, M^{pro}ST, and M^{pro}SH were evaluated by plotting the main-chain C_α RMSDs at different temperatures as a function of running time and the results are shown in Figure 3A-F, respectively. At 300 K, the overall RMSDs for these proteins all converged below 3 Å, which is in good agreement with the results from previous MD simulations (16). In addition, the increases of the overall RMSDs for these proteins at 400 and 600 K followed the similar pattern, except for M^{pro}H, whose overall RMSD reached 9 Å at 600 K; whereas those of the other three proteins reached 6 Å only. It indicates that M^{pro}H may undergo an overall structural change more dramatically at high temperature. By comparing the RMSDs of the substrate binding sites S1 and S2 at various temperatures (Fig. 3B and C), we found that S1 exhibits higher structural integrity than S2. It is attributed to that S2 is located on the open mouth of the catalytic cleft between domains I and II and is fully solvent-exposure, whereas S1 is situated in the very bottom of this cleft and is well protected from the hydrophobic core. The higher structural variation of S2 makes it flexible enough to accommodate a bulky hydrophobic residue from the substrate. Furthermore, S2 of M^{pro}H undergoes a more dramatic structural change at higher temperatures than S2 of the other proteins, indicating that M^{pro}H may lose its binding affinity towards various substrates or inhibitors more easily than the other three M^{pro}.

Comparing the RMSD values in Figure 3D-F, we found that domains I and II of M^{pro}T, M^{pro}H, M^{pro}ST, and M^{pro}SH follow the similar dynamics behaviors; whereas domain III of these proteins shows different structural variations during the entire simulation time courses. This result is in good agreement with results of amino acid sequence alignment and homology modeling, showing that domain III of these proteins exhibit least structural similarity among these three domains. The secondary structure propensity of these proteins was predicted according to DSSP (45) during the entire MD courses at various temperatures and the results are shown in Figure 4.

The values of the average secondary structure content for each secondary structure element in these proteins are summarized in Table II. As expected, it is faster for domain III to lose its helical content than for domains I and II to lose their sheet content in all cases. The high dielectric constant of the explicit water system may increase the opportunity of hydrogen bonding between amide protons and surrounding solvent molecules and simultaneously promotes the intermolecular hydrogen bonding and therefore destabilizes the structural integrity of these helices in domain III. From the analyses of the average secondary structure contents (Table II) and the secondary structure propensities during the MD time courses (Fig. 4), we estimated that the thermal unfolding of the helices in domain III of both M^{pro}T and M^{pro}H follows the order of CIII→EIII→BIII→DIII→AIII. Helix AIII is mainly composed of nonpolar residues and forms an interior hydrophobic core in domain III, which is in turn restricted to solvent exposure and thus maintains higher helical content than the other helices. The ASA for each residue in helix AIII is nearly zero (data not shown), again indicating that the hydrophobic environment around helix AIII may protect it from forming intermolecular hydrogen bonding with water molecules. Furthermore, the result of amino acid sequence alignment shows that helix AIII exhibits higher sequence identity than the other helices in domain III among these proteins, which may also emphasize the importance of helix AIII in maintaining the structural integrity of the globular domain III in M^{pro}.

Table II
Average secondary structure content for each secondary structure element in M^{pro}T, M^{pro}H, M^{pro}ST, and M^{pro}SH.

Secondary structure element	Average secondary structure content (%)											
	M^{pro}T			M^{pro}H			M^{pro}ST			M^{pro}SH		
	300 K	400 K	600 K	300 K	400 K	600 K	300 K	400 K	600 K	300 K	400 K	600 K
aI	75	65	10	55	43	7	60	73	43	99	90	20
bI	72	64	14	64	18	9	93	85	62	95	84	22
cI	55	67	26	85	48	18	65	62	53	90	75	27
AI	3	16	2	3	3	0	-	-	-	90	85	12
dI	59	61	8	40	53	8	-	-	-	68	46	3
eI	53	45	5	50	41	5	-	-	-	91	74	23
fI	77	76	20	83	72	12	50	47	25	55	65	8
aII	60	50	15	61	58	17	37	21	1	56	46	5
bII	45	36	10	54	53	10	36	19	2	86	55	1
cII	45	39	22	44	39	7	33	15	14	87	76	5
dII	42	46	19	60	45	10	88	76	52	82	60	5
eII	49	37	18	65	34	3	86	76	53	81	44	8
fII	22	20	3	22	5	1	52	48	19	49	8	5
AIII	85	56	13	69	63	14	56	44	34	61	42	8
BIII	78	45	5	71	39	6	-	-	-	-	-	-
CIII	37	13	1	1	0	0	78	36	6	-	-	-
DIII	93	63	9	96	72	9	-	-	-	-	-	-
EIII	92	76	3	35	53	3	90	59	22	66	58	4

In contrast to the specific unfolding order of the helices in domain III, there is no particular unfolding order of the sheets in domains I and II (Fig. 4 and Table II). The packing of the sheets in domains I and II is similar to a sandwich and the catalytic cleft is located in the middle of this well organized structure. The nucleophilic Cys144 is located in the center of this catalytic cleft and some of the residues forming the substrate binding site S1 is distributed in some of the sheets in domains I and II. Thus, in order to maintain the proteolytic activity, these sheets have to preserve their secondary structural integrity. Most of the structural variations in domains I and II at high temperatures are resulted from the fluctuation of outer loops, which are fully exposed to the solvent. Previous study has shown that the region around residues 10-20 (corresponding to sheet bI in domain I) is relatively rigid and the region around residues 265-287 (corresponding to the loop connecting helices DIII and EIII in domain III) is relatively flexible than the other regions of M^{pro}ST (16). The present results also indicate that the structural network formed by the sheets in domains I and II is relatively stable during the MD simulation courses comparing to the network formed by the helices in domain III. A short helix AI is observed in the outer surfaces of domain I in the crystal structures of M^{pro}T and M^{pro}H (Fig. 1A and B), whereas this helix is missing in the homology models of M^{pro}ST and M^{pro}SH

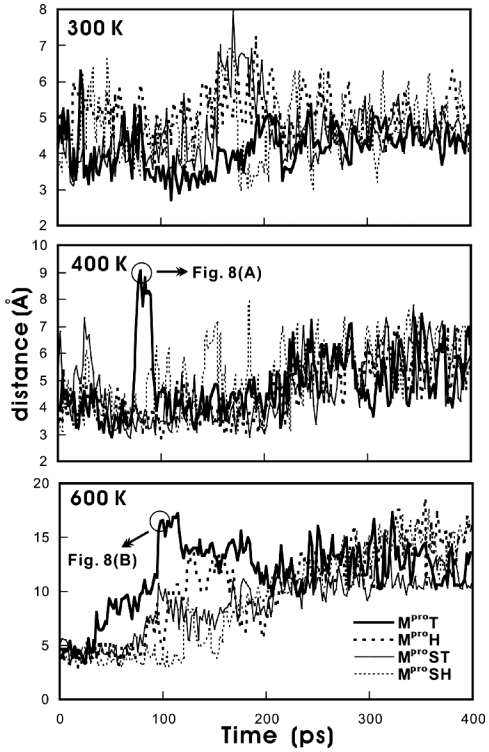

Figure 7: The distance between the sulfur atom of the nucleophilic Cys residue and the $N^{\varepsilon 2}$ of the general acid-base catalyst His residue as a function of MD simulation time for $M^{pro}T$, $M^{pro}H$, $M^{pro}ST$, and $M^{pro}SH$ at 300, 400, and 600 K. The snapshots of $M^{pro}T$ taken at 82 ps, 400 K and at 100 ps, 600 K are shown in Figure 8A and B, respectively.

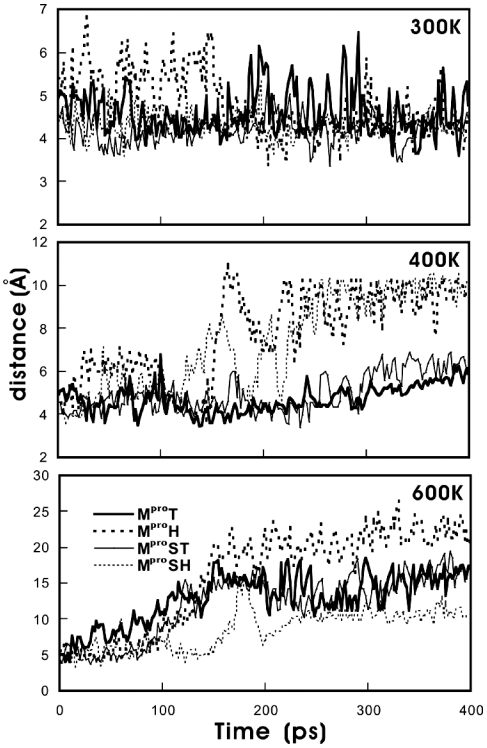

Figure 6: The distance between the C^{ε} atom of the totally conserved Arg40 from the substrate binding site S2 and the C^{γ} atom of the totally conserved Asp186 from the intervening loop connecting domains II and III (residue numbered as in $M^{pro}T$) as a function of MD simulation time for $M^{pro}T$, $M^{pro}H$, $M^{pro}ST$, and $M^{pro}SH$ at 300, 400, and 600 K.

(Fig. 1C and D). During the MD simulation courses, this helix disappeared very quickly due to the contact with the surrounding water molecules.

It has been shown previously that, similarly to $3C^{pro}$ (52-54), specific substrate binding by M^{pro} is ensured by the well-defined S1 and S2 substrate binding pockets (15). In addition, it has also been shown that the imidazole side chain of the 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 (52-54). 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 $M^{pro}S$ (15). This salt bridge is itself on the periphery of these molecules, forming part of the outer wall of the substrate binding site S1. Figure 5 shows the ASAs of the substrate binding sites S1 and S2 of $M^{pro}T$, $M^{pro}H$, $M^{pro}ST$, and $M^{pro}SH$ at various temperatures. In general, S2 exhibits higher ASAs than S1 during the MD simulation courses. In addition, S1 was found to maintain its structural integrity, whereas S2 exhibits more structural variations during the MD simulations. It is attributed to that S2 is located on the open mouth of the catalytic cleft between domains I and II and thus is fully exposed to the surrounding solvent, whereas S1 is situated in the very bottom of this cleft and is subsequently protected by the hydrophobic core. The higher structural variation of S2 is necessary for the proteolytic activity of M^{pro} because it is flexible enough to accommodate a bulky hydrophobic residue from the substrate and further allows the substrate to form close contact with the substrate binding cleft formed by S1 and S2 of this enzyme.

Previous study has indicated that the loop connecting domains II and III (residues 184-199) plays an important role in maintaining the proteolytic activity of $M^{pro}T$ (15). This intervening loop is located in adjacent to the substrate binding site S2. Moreover, the totally conserved Arg40 from S2 forms an electrostatic interaction

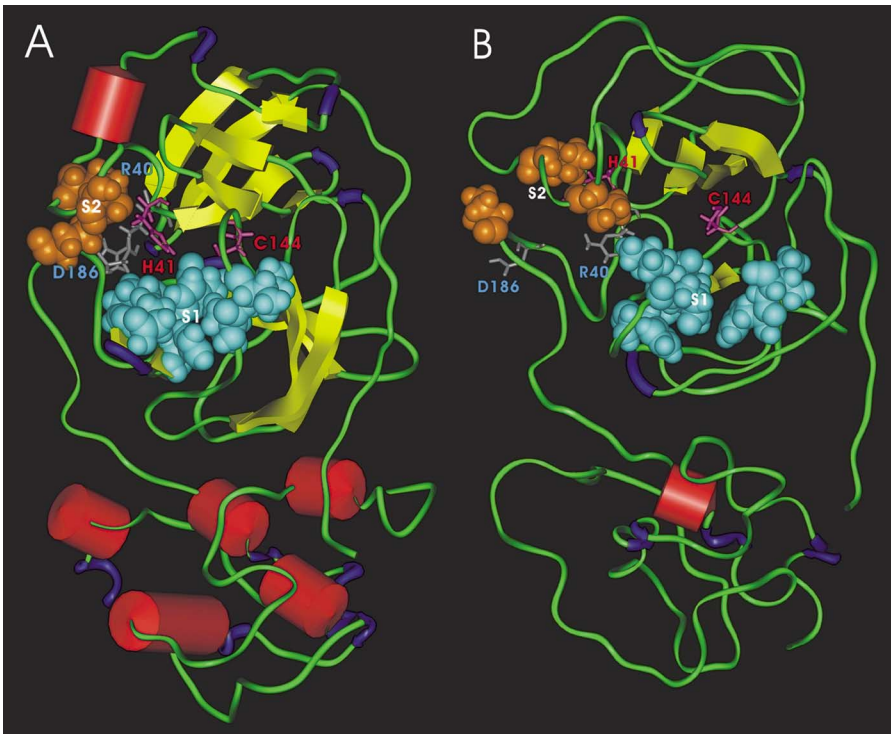

Figure 8: The snapshots of M^proT taken at (**A**) 82 ps, 400 K and (**B**) 100 ps, 600 K. Secondary structures predicted according to DSSP (43) are shown as in Figure 1. Substrate binding sites S1 and S2 are represented as CPK and colored in indigo and brown, respectively. The nucleophilic Cys144 and the general acid-base catalyst His41 are shown in purple as sticks. The totally conserved residues Arg40 and Asp186 forming the electrostatic interaction in the native structure of M^proT are shown in grey as stick. These structures are visualized by Insight II program.

with the totally conserved Asp186 from this extended loop (15). In order to investigate the importance of this electrostatic interaction in maintaining the structural integrity of S2 during the MD simulations, the distance between the C^{ε} atom of Arg40 and the C^{γ} atom of Asp186 (residues numbered as in M^proT) for each protein was measured and the results are shown in Figure 6. In addition, the distance between the S atom of the nucleophilic Cys residue and the $N^{\varepsilon 2}$ of the general acid-base catalyst His residue for each structure was also recorded as a function of MD simulation time and the results are given in Figure 7. At higher temperatures, these distances all increase significantly, indicating that the electrostatic interaction formed by Arg40 and Asp186 and the catalytic activity formed by the Cys-His pair are destroyed towards heating. Interestingly, the distance between Cys-His pair increases dramatically for M^proT between 75 and 90 ps at 400 K. In order to compare the partially unfolded structure with the totally unfolded one, the snapshots of M^proT at 82 ps, 400 K (the distances between Arg40 and Asp186 and between His41 and Cys144 are 7.79 and 9.04 Å, respectively) and at 100 ps, 600 K (the distances between Arg40 and Asp186 and between His41 and Cys144 are 11.18 and 16.08 Å, respectively) were generated as in Figure 8A and B, respectively. According to these snapshots, the former structure still maintains most of its secondary structural integrity, whereas most of the secondary structures disappear in the later one. In addition, the substrate binding sites S1 and S2 still maintain their structural integrity in the partially unfolded M^proT, whereas these two binding sites are shifted and destroyed in the totally unfolded M^proT. It indicates that the electrostatic interaction between Arg40 and Asp186 plays an important role in maintaining the packing of S1 and S2, thus preserving the proteolytic activity of this enzyme (15). From the above results, we may conclude that M^proT may still maintain its proteolytic activity while it is partially unfolded and that the electrostatic interaction between Arg40 and Asp186 functions as a gate controlling the open and close states of the substrate binding sites S2. Previous MD simulations have shown that M^proST complexed with inhibitor is, in average, less flexible than the free enzyme either in the monomeric or dimeric form (16). Our simulation results also indicates that water molecules may enter S2 and further penetrate S1 without the protection from the bound inhibitor when the electrostatic interaction between Arg40 and Asp186 is destroyed at high temperatures, resulting in the distortion and destroy of the packing of these two sites, which are mainly lined up by hydrophobic residues.

In conclusion, the homology models of M^{pro}S were successfully constructed based on the crystal structures of both M^{pro}T (15) and M^{pro}H (14) by comparative approach. Both M^{pro}ST and M^{pro}SH exhibit similar folds as their respective templates M^{pro}T and M^{pro}H. Three distinct functional domains as well as an intervening loop from residues 184 to 199 connecting domains II and III are also obtained in these homology models as in the template proteins. A catalytic cleft containing the substrate binding sites S1 and S2 between domains I and II are also observed in these homology models. S2 undergoes more dramatic structural changes than S1 because it is located at the open mouth of the catalytic cleft and is fully exposed to the solvent, whereas S1 is situated in the very bottom of this cleft and is well protected from the hydrophobic core. The unfolding of these proteins begins at domain III, where the structure is least conserved among these proteins. M^{pro} may still maintain its proteolytic activity while it is partially unfolded. The electrostatic interaction between the totally conserved Arg40 from the substrate binding site S2 and the totally conserved Asp186 from the intervening loop between domains II and III (residues numbered as in M^{pro}T) plays an important role in maintaining the structural integrity of both S1 and S2.

Acknowledgements

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

References and Footnotes

1. C. Drosten *et al*. *N*. *Engl*. *J*. *Med*. *348*, 1967-1976 (2003).
2. T. G. Ksiazek *et al*. *N*. *Engl*. *J*. *Med*. *348*, 1953-1966 (2003).
3. N. Lee *et al*. *N*. *Engl*. *J*. *Med*. *348*, 1986-1994 (2003).
4. M. A. Marra *et al*. *Science 300*, 1399-1404 (2003).
5. P. A. Rota *et al*. *Science 300*, 1394-1399 (2003).
6. J. A. Den Boon, E. J. Snijder, E. D. Chirnside, A. A. de Vries, M. C. Horzinek, and W. J. Spaan. *J*. *Virol*. *65*, 2910-2920 (1991).
7. D. Cavanagh. *Arch*. *Virol*. *142*, 629-633 (1997).
8. J. Herold, T. Raabe, B. Schelle-Prinz, and S. G. Siddell. *Virology 195*, 680-691 (1993).
9. V. Thiel, J. Herold, B. Schelle, and S. G. Siddell. *J*. *Virol*. *75*, 6676-6681 (2001).
10. J. Ziebuhr, J. Herold, and S. G. Siddell. *J*. *Virol*. *69*, 4331-4338 (1995).
11. A. E. Gornalenya, A. P. Donchenko, V. M. Blinov, and E. V. Koonin. *FEBS Lett*. *243*, 103-114 (1989).
12. A. E. Gornalenya, E. V. Koonin, A. P. Donchenko, and V. M. Blinov. *Nucleic Acids Res*. *17*, 4847-4861 (1989).
13. J. Ziebuhr, G. Heusipp, and S. G. Siddell. *J*. *Virol*. *71*, 3992-3997 (1997).
14. K. Anand, J. Ziebuhr, P. Wadhwani, J. R. Mesters, and R. Hilgenfeld. *Science 300*, 1763-1767 (2003).
15. K. Anand, G. J. Palm, J. R. Mesters, S. G. Siddell, J. Ziebuhr, and R. Hilgenfeld. *EMBO J*. *21*, 3213-3224 (2002).
16. V. S. Lee *et al*. *Science Asia 29*, 181-188 (2003).
17. B. Xiong *et al*. *Acta Pharmacol*. *Sin*. *24*, 497-504 (2003).
18. V. Muñoz, E. R. Henry, J. Hofrichter, and W. A. Eaton. *Proc*. *Natl*. *Acad*. *Sci*. *USA 95*, 5872-5879 (1998).
19. P. A. Thompson, W. A. Eaton, and J. Hofrichter. *J*. *Laser Biochemistry 36*, 9200-9210 (1997).
20. S. Williams, T. P. Causgrove, R. Gilmanshin, K. S. Fang, R. H. Callender, W. H. Woodruff, and R. B. Dyer. *Biochemistry 35*, 691-697 (1996).
21. A. E. Mark and W. F. van Gunsteren. *Biochemistry 31*, 7745-7748 (1992).
22. G. Hummer, S. Garde, A. E. Garcia, M. E. Paulaitis, and L. R. Pratt. *Proc*. *Natl*. *Acad*. *Sci*. *USA 95*, 1552-1555 (1998).
23. M. Karplus and A. Šali. *Curr*. *Opin*. *Struct*. *Biol*. *5*, 58-73 (1995).
24. V. Daggett and M. Levitt. *Curr*. *Opin*. *Struct*. *Biol*. *4*, 291-295 (1994).
25. C. N. Pace, D. V. Laurentz, and R. E. Erickson. *Biochemistry 31*, 1351-1361 (1992).
26. H.-L. Liu and W.-C. Wang. *Chem.Phys*. *Lett*. *366*, 284-290 (2002).
27. H.-L. Liu and W.-C. Wang. *J*. *Biomol*. *Struct*. *Dyn*. *20*, 615-622 (2003).
28. H.-L. Liu and W.-C. Wang. *Biotechnol*. *Progr. 19*, 1583-1590 (2003).
29. H.-L. Liu, W.-C. Wang, and C.-M. Hsu. *J*. *Biomol*. *Struct*. *Dyn*. *20*, 567-574 (2003).
30. H.-L. Liu, Y. Ho, and C.-M. Hsu. *Chem*. *Phys*. *Lett*. *372*, 249-254 (2003).
31. H.-L. Liu, Y. Ho, and C.-M. Hsu. *J*. *Biomed*. *Sci*. *10*, 302-312 (2003).
32. H.-L. Liu, Y.-C. Shu, and Y.-H. Wu. *J*. *Biomol*. *Struct*. *Dyn*. *20*, 741-746 (2003).
33. H.-L. Liu and Y.-M. Lin. *J*. *Chin*. *Chem*. *Soc*. *50*, 799-808 (2003).

34. H.-L. Liu and J.-C. Lin. *Chem. Phys. Lett. 381*, 592-597 (2003).
35. H.-L. Liu and J.-C. Lin. *J. Biomol. Struct. Dyn. 21*, 639-650 (2004).
36. Z. Peng, C. S. Ewig, M.-J. Hwang, M. Waldman, and A. T. Hagler. *J. Phys. Chem. A 101*, 7243-7252 (1997).
37. M.-J. Hwang, X. Ni, M. Waldman, C. S. Ewig, and A. T. Hagler. *Biopolymers 45*, 435-468 (1998).
38. J. R. Maple, M.-J. Hwang, K. J. Jalkanen, T. P. Stockfisch, and A. T. Hagler. *J. Comp. Chem. 19*, 430-458 (1998).
39. J. Greer. *Proteins 7*, 317-334 (1990).
40. V. Thiel, J. Herold, B. Schelle, and S. G. Siddell. *J. Gen. Virol. 82*, 1273-1281 (2001).
41. F. Almazan, J. M. Gonzalez, Z. Penzes, A. Izeta, and E. Calvo. *Proc. Natl. Acad. Sci. USA 97*, 5516-5521 (2000).
42. G. D. Schuler, S. F. Altschul, and D. J. Lipman. *Proteins: Struct. Funct. Genet. 9*, 180-189 (1991).
43. P. S. Shenkin, D. L. Yarmush, R. M. Fine, H. Wang, and C. Levinthal. *Biopolymers 26*, 2053-2085 (1987).
44. J. W. Ponder and F. M. Richards. *J. Mol. Biol. 193*, 775-791 (1987).
45. W. Kabsch and C. Sander. *Biopolymers 22*, 2577-2637 (1983).
46. H. J. C. Berendsen, J. P. M. Postma, W. F. van Gunsteren, A. DiNola, and J. R. Haak. *J. Comp. Phys. 81*, 3684-3690 (1984).
47. A. M. Lesk. *Proteins 33*, 320-328 (1998).
48. M. Levitt and M. Gerstein. *Proc. Natl. Acad. Sci. USA 95*, 5913-5920 (1998).
49. K. Mizuguchi and N. Go. *Curr. Opin. Struct. Biol. 5*, 377-382 (1995).
50. P. Koehi. *Curr. Opin. Struct. Biol. 11*, 348-353 (2001).
51. W. Kabsch. *Acta Cryst. A 32*, 922-923 (1976).
52. D. A. Matthews *et al. Cell 77*, 761-771 (1994).
53. E. M. Bergmann, S. C. Mosimann, M. M. Chernaia, B. A. Malcolm, and M. N. James. *J. Virol. 72*, 2436-2448 (1997).
54. S. C. Mosimann, M. M. Cherney, S. Sia, S. Plotch, and M. N. James. *J. Mol. Biol. 273*, 1032-1047 (1997).

Date Received: February 23, 2004

Communicated by the Editor Ramaswamy H Sarma