Biophysical Journal
Article

Biophysical Society

The Molecular Mechanism of Domain Swapping of the C-Terminal Domain of the SARS-Coronavirus Main Protease

Vishram L. Terse[1] and Shachi Gosavi[1,*]
[1]Simons Centre for the Study of Living Machines, National Centre for Biological Sciences, Tata Institute of Fundamental Research, Bangalore, India

ABSTRACT In three-dimensional domain swapping, two protein monomers exchange a part of their structures to form an intertwined homodimer, whose subunits resemble the monomer. Several viral proteins domain swap to increase their structural complexity or functional avidity. The main protease (M^{pro}) of the severe acute respiratory syndrome (SARS) coronavirus proteolyzes viral polyproteins and has been a target for anti-SARS drug design. Domain swapping in the α-helical C-terminal domain of M^{pro} ($M^{pro}C$) locks M^{pro} into a hyperactive octameric form that is hypothesized to promote the early stages of viral replication. However, in the absence of a complete molecular understanding of the mechanism of domain swapping, investigations into the biological relevance of this octameric M^{pro} have stalled. Isolated $M^{pro}C$ can exist as a monomer or a domain-swapped dimer. Here, we investigate the mechanism of domain swapping of $M^{pro}C$ using coarse-grained structure-based models and molecular dynamics simulations. Our simulations recapitulate several experimental features of $M^{pro}C$ folding. Further, we find that a contact between a tryptophan in the $M^{pro}C$ domain-swapping hinge and an arginine elsewhere forms early during folding, modulates the folding route, and promotes domain swapping to the native structure. An examination of the sequence and the structure of the tryptophan containing hinge loop shows that it has a propensity to form multiple secondary structures and contacts, indicating that it could be stabilized into either the monomer- or dimer-promoting conformations by mutations or ligand binding. Finally, because all residues in the tryptophan loop are identical in SARS-CoV and SARS-CoV-2, mutations that modulate domain swapping may provide insights into the role of octameric M^{pro} in the early-stage viral replication of both viruses.

SIGNIFICANCE The main protease (M^{pro}) of the severe acute respiratory syndrome coronavirus cleaves single-stranded viral polyproteins into functional proteins. M^{pro} has two dimerization interfaces: a canonical one, which enables the formation of the active protease dimer, and a noncanonical domain swapping one present in its C-terminal domain ($M^{pro}C$). Together, these interfaces enable the formation of a highly active M^{pro} octamer, which has been hypothesized to play a key role in early-stage viral replication. Here, we use computer simulations to understand the mechanism of domain swapping of $M^{pro}C$ and suggest mutations that could change the domain-swapped dimer population. These mutations may also tune the octamer population in M^{pro} and help obtain experimental evidence for the role of the octamer during viral replication.

INTRODUCTION

The severe acute respiratory syndrome (SARS) outbreak of 2003 was attributed to a novel coronavirus named SARS-CoV (1–3). It is a single-stranded positive sense RNA virus with a genome of ~30 kb (4,5). The virus encodes two polyproteins namely pp1a and pp1ab that are proteolyzed to give 16 nonstructural proteins (nsps 1–16). This proteolysis is performed by two viral proteinases: a papain-like proteinase and a 3C-like proteinase (3CLpro). 3CLpro, also known as the main protease (M^{pro}), is involved in the cleavage of 11 nsps (6,7). Hence, it has been an attractive target for anti-SARS drug design (8–10).

M^{pro} is a 33.8 kDa protein of 306 residues (11). The N-terminal domain of M^{pro} ($M^{pro}N$) has a fold similar to trypsin-like serine proteases but with the catalytic residues being Cys145 and His41 instead of the usual Ser-His-Asp found in serine proteases (9). $M^{pro}N$ is further split into domain I

Submitted July 14, 2020, and accepted for publication November 24, 2020.
*Correspondence: shachi@ncbs.res.in
Editor: Alan Grossfield.
https://doi.org/10.1016/j.bpj.2020.11.2277

(residues 8–101) and domain II (residues 102–184) and has the substrate-binding cleft. The N-finger formed by the N-terminal residues 1–7 interacts with the C-terminal domain of M^{pro} ($M^{pro}C$; residues 201–303) and is important for M^{pro} dimerization and active site formation (9). $M^{pro}C$, also called domain III, is unique to coronaviruses and is not found in other cysteine proteases with a chymotrypsin fold (12,13). It has a globular fold with five α-helices (Fig. 1) and is connected to $M^{pro}N$ by a long loop (residues 185–200) (9).

M^{pro} exists in solution in an equilibrium between a monomer and a non-domain-swapped side-by-side dimer (9,15). Several experiments (15–17) show that only the dimeric form is catalytically active with the $M^{pro}C$ domain also being a potential drug target (18) because it contributes to the dimerization of M^{pro}, switching the enzyme from the inactive monomeric form to the active dimeric form (15). Specifically, residues of $M^{pro}C$ that are in close contact with $M^{pro}N$ as well as the N-finger in the dimer can affect dimerization as well as catalysis through allostery (19–21).

M^{pro} shows reversible unfolding in guanidinium chloride at 30°C and pH 7.7 (22). These equilibrium unfolding

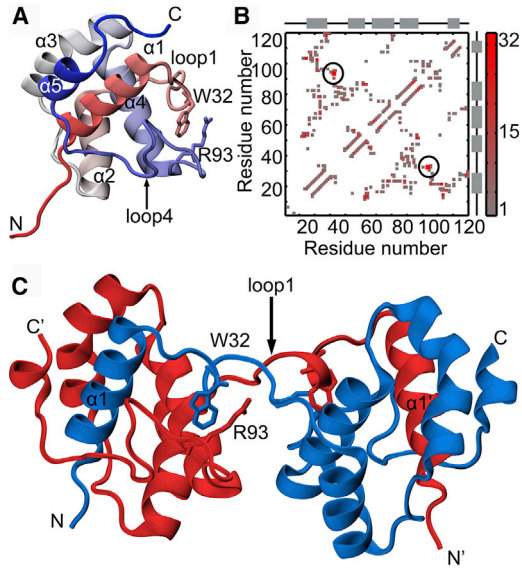

FIGURE 1 Monomer and domain-swapped dimer of $M^{pro}C$. (*A*) The structure of the C-terminal domain ($M^{pro}C$) of the M^{pro} of SARS-CoV (PDB: 2H2Z, residues 187–306) is shown colored from red through white to blue from N- to C-termini. Secondary structural elements α1–α5, N- and C-termini, and loops 1 (L1) and 4 (L4) are marked. (*B*) The 4.5-Å heavy atom cutoff contact map of $M^{pro}C$ projected onto the Cα-atoms is shown. A contact between residues *i* and *j* is shown as a square at (*i*, *j*) and (*j*, *i*). The color of each contact represents the number of all atom contacts between the residues. Color bar is shown on the right. The contact with the highest number of atomic contacts (W32-R93) is circled in black. These residues are also shown in (*A*) and (*C*). Helices, represented by gray rectangles, and loops, represented by black lines, are marked on the axes. (*C*) Domain-swapped dimer of $M^{pro}C$ (PDB: 3EBN) is shown. The two identical protein chains involved in domain swapping are colored red and blue, and their termini are labeled. L1, the hinge region, and α1, the swapped helix, are marked on the structure. All protein structures were drawn using VMD (14). To see this figure in color, go online.

studies show that M^{pro} unfolds in a sequential manner with $M^{pro}N$ unfolding at lower guanidinium chloride concentrations followed by $M^{pro}C$. The stability of full-length M^{pro} calculated from these experiments is ~12 kcal/mol whereas the stability of $M^{pro}C$ within the context of full-length M^{pro} is ~10 kcal/mol. This indicates that a large part of the M^{pro} stability derives from $M^{pro}C$ (22). The equilibrium unfolding of $M^{pro}C$ has been described as a two state process (22–24). Isolated $M^{pro}C$ exists as a monomer and dimer in solution and no interconversion was observed between these species in gel filtration (25). This $M^{pro}C$ dimer, unlike the previously identified side-by-side dimer of M^{pro} (9), is a three-dimensional domain-swapped dimer (26).

Three-dimensional domain swapping, often called only domain swapping, is the process by which two identical protein chains exchange a part of their structure to form an intertwined dimer or higher-order oligomer (27,28). The piece of structure that is exchanged is called a "swapped-domain," whereas the remaining protein is called the "core protein." The fragment joining these parts is called the hinge or the hinge loop. Only the hinge undergoes a conformational change when the protein domain swaps. In the monomer, this hinge is in a closed or loop form, whereas in the domain-swapped form, it is in an extended conformation (29). Domain swapping in diverse viral proteins plays a role in functional regulation (30) as well as structural assembly (31,32). The domain swapping of $M^{pro}C$ when present within the context of M^{pro} creates a second dimerization interface in addition to that present in the side-by-side dimer. The geometry and orientation of the two interfaces lock M^{pro} into a stable octamer (AO-M^{pro}) that has been crystallized (30). This octamer has been termed "super active" because all eight of its units are active and this activity is not protein concentration dependent (30). AO-M^{pro} is expected to have much higher proteolytic activity at low protein concentrations than the non-domain-swapped M^{pro}, which is in dynamic equilibrium between a catalytically active side-by-side dimer and an inactive monomer. Thus, it has been hypothesized that AO-M^{pro} may be useful during the early stages of viral infection when the concentration of M^{pro} is low (30). An octamer has also been observed in a mutant of the MERS coronavirus M^{pro} (33), providing further evidence that this form may be biologically relevant. As stated earlier, $M^{pro}C$ is composed of five α-helices (Fig. 1 *A*). The experimentally observed domain-swapped dimer is formed by the exchange of the α1 helices between the two monomers (Fig. 1 *C*). The hinge between α1 and α2, here termed loop 1 or L1, is in a loop conformation in the monomer and an extended structure with a helical turn in the domain-swapped dimer (26).

The molecular basis for the domain swapping of $M^{pro}C$ and in turn, the formation of AO-M^{pro} is unclear (34). Web servers known to predict domain swapping either predict $M^{pro}C$ as non-domain-swapping (35) or predict the location of the hinge region incorrectly (36). Theoretical analysis of the kinetic data

for domain swapping and thermodynamic data of folding and unfolding led to the conclusion that $M^{pro}C$ undergoes domain swapping from a completely unfolded state (37). Although $M^{pro}C$ does not show any reversible exchange between the monomeric and domain-swapped forms at room temperature, increasing the temperature to 37°C or the addition of urea facilitates this process. Hydrogen exchange NMR experiments showed that this exchange is initiated by localized unfolding of $\alpha5$ (24). Amide exchange rates from this experiment helped to conclude that $\alpha1$ present in the core of $M^{pro}C$ is not exposed to water during the process of domain swapping. To explain these observations a model was proposed in which two partially unfolded monomers form a dimeric intermediate using their $\alpha5$ helices and exchange their $\alpha1$ helices in a hydrophobic environment. This led to the hypothesis that nonnative interactions may be important for $M^{pro}C$ domain swapping (24). However, molecular dynamics (MD) simulations of the monomer and the domain-swapped form of $M^{pro}C$ predict that it is necessary for the swapped $\alpha1$ helix to be exposed in the unfolded form to induce domain swapping (34). Thus, the mechanism of domain swapping inferred from experiments (24) differs from that predicted from simulations (34) and theoretical analysis (37). Here, we investigate the causes for and the mechanism of $M^{pro}C$ domain swapping using coarse-grained structure-based models.

Structured proteins fold on a biologically relevant timescale because of a funnel shaped energy landscape in which interactions not present in the native or folded state of the protein (nonnative interactions) stabilize structure far less than native interactions do (38). Thus, protein models that encode only the native structure of the protein, termed structure-based models (SBMs), can be used to simulate proteins and have been successfully used to understand the barriers to and the mechanisms of protein folding (39–42) and domain swapping (43–46). We performed MD simulations of an SBM (39,47,48) of $M^{pro}C$ to investigate both protein folding and domain swapping. Our monomer simulations of $M^{pro}C$ were able to recapitulate several experimental features (22–24) of its folding. Further, domain-swapping simulations were able to predict the correct domain-swapped dimer structure, i.e., the largest population of simulated domain-swapped dimers was similar in structure to the experimental domain-swapped dimer (26). Because these simulations did not include nonnative interactions, domain swapping is intrinsic to the $M^{pro}C$ structure and nonnative interactions are not necessary for swapping. We then simulated a variant SBM to identify interactions that promoted correct domain swapping. We found that a disruption of interactions between residues in L1 (loop between $\alpha1$ and $\alpha2$) and $\alpha1$ with residues in L4 (loop between $\alpha4$ and $\alpha5$) affects domain-swapping propensity and leads to nonnative domain-swapped structures. We then examine several interactions at the L1–L4 interface in detail and suggest mutations that could modulate $M^{pro}C$ domain swapping.

The virus causing the ongoing Coronavirus disease 2019 (COVID-19) pandemic (49), SARS-CoV-2, is similar in sequence to SARS-CoV. Although a domain-swapped structure has not yet been reported for the SARS-CoV-2 $M^{pro}C$, the $M^{pro}Cs$ from the two viruses differ at only four residues. We end the Discussion by presenting the relevance of the current simulations to the $M^{pro}C$ of SARS-CoV-2.

METHODS

SBMs

As stated earlier, interactions present in the folded state of the protein are far more stabilizing than nonnative interactions (38). Thus, SBMs encoding only native interactions in their potential energy functions have been successfully used to understand several aspects of protein folding (39–42). Here, we use a common coarse-grained SBM (39) in which each amino acid is represented by a single bead at the position of its Cα-atom. The potential energy of this SBM is:

$$E = \sum_{\text{bonds}} K_r (r - r_0)^2 + \sum_{\text{angles}} K_\theta (\theta - \theta_0)^2 + \sum_{\text{dihedrals}}^{n=1,3} K_\varphi^{(n)} (1 - \cos(n(\varphi - \varphi_0))) + \sum_{\text{contacts}} \varepsilon_1 \left[5 \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{12} - 6 \left(\frac{\sigma_{ij}}{r_{ij}} \right)^{10} \right] + \sum_{\text{non-contacts}} \varepsilon_2 \left(\frac{\sigma}{r_{ij}} \right)^{12}$$

The first two harmonic terms represent the energy of bond and angle fluctuations. The third term represents dihedral rotations. These terms have their deepest minima at r_0, θ_0, and φ_0, the values of which are calculated from the Cα-coordinates of the folded protein. The force constants for these bond, angle, and dihedral terms are $K_r = 100\varepsilon$, $K_\theta = 20\varepsilon$, $K_\varphi{}^{(1)} = \varepsilon$, and $K_\varphi{}^{(3)} = 0.5\varepsilon$, where ε is the basic energy scale in which the SBM potential energy is expressed. The final term gives the repulsive energy between those pairs of Cα-atoms, which are not in contact in the native state. This excluded volume term with σ set to 0.4 nm ensures that Cα-atoms not in contact do not pass through each other during the dynamics. The strength of all noncontacts, ε_2, is set to ε.

The fourth term, an attractive 10–12 Lennard-Jones-like potential, represents the attractive energy between residues that are in contact in the native state. σ_{ij} is the distance between the Cα-atoms of residues i and j, which are in contact. A contact exists between the two residues i and j if at least one nonhydrogen (or heavy) atom from residue i is within 4.5 Å of at least one heavy atom of residue j. Furthermore, only those pairs of residues (i, j) are present in the contact list where $|j-i| > 3$. Contact maps are plots of these contact lists, with boxes being plotted at (i, j) and (j, i) on a contact map when a contact is present between residues i and j in the contact list. The contact strength (ε_1) calculation is described in the next section.

The choice of contact weights

Several weighting schemes (47,48,50–52), including equal weights or $\varepsilon_1 = \varepsilon$ for all contacts (39), have been used for calculating contact strengths (ε_1) in the potential energy function shown above. Choosing an appropriate weighting scheme is particularly important in α-helical proteins in which heterogeneity in contact weights is more likely to matter to folding outcomes (50). We tested the folding of $M^{pro}C$ using two different weighting schemes.

In the first, the weight, ε_1, of a contact (47,48) is calculated as follows: if n_{AA} is the number of interheavy atom contacts present between the two residues i and j, then ε_1 is $(n_{AA}/\Sigma n_{AA}) \times N_{C\alpha} \times \varepsilon$, where Σn_{AA} is the total number of interheavy atom contacts in the protein and $N_{C\alpha}$ is the total number of Cα-Cαcontacts. $N_{C\alpha}$ equals 233 in $M^{pro}C$, whereas the total number of interheavy atom contacts is 1270. The list of Cα contacts and the n_{AA} for each contact are given in List S1. Here, we term this weighting scheme and the associated model SBM. In the second, more commonly used scheme (39), the weights of all contacts n_{AA} is set to 1. Thus, $\Sigma n_{AA} = N_{C\alpha}$ and $\varepsilon_1 = 1 \times \varepsilon$. We term this weighting scheme and model SBMe. All parameters other than the weights in SBMe are identical to SBM described above.

Hydrogen deuterium exchange NMR experiments on $M^{pro}C$ show that the stability of α1 is the highest among all the helices and is similar to that of the entire $M^{pro}C$ (24), implying that α1 is likely to fold early and unfold last. Average distance maps (Fig. S1) and an alignment of five randomly chosen structures (Fig. S2) indicate that α1 folds first and unfolds last in the SBM simulations, but it folds after the folding of α2-α3-α4 in SBMe simulations. Because α1 is involved in domain swapping and both its order as well as the overall order of structure formation in the weighted SBM (α1 and α4 followed by α2 and α3 followed by α5) more closely resembles the order of structure formation inferred from helix stabilities ($\alpha1 > \alpha4 > \alpha2 > \alpha3 > \alpha5$) in experiment (24), we chose to simulate the folding and domain swapping of $M^{pro}C$ using the model with weighted contacts (SBM).

The effect of a particular contact weight on folding and domain swapping was studied by setting the n_{AA} of that contact to 1. Once this is done, Σn_{AA} (the total number of heavy atom contacts) was recalculated with the new contact weight. $N_{C\alpha}$ remains the same as before. The weights (ε_1) were then recalculated for all the contacts using the new Σn_{AA}.

Simulation details

All simulations were performed using the GROMACS 4.5.4 (53) program suite. The basic energy (ε), time, and distance scales in GROMACS are 1 kJ/mol, 1 ps, and 1 nm, respectively, and all terms in the potential energy function are expressed in terms of these units. Cα-SBMs were constructed using the SMOG webserver (54) which requires Cα coordinates and contact maps as input. Cα-coordinates for $M^{pro}C$ were obtained from the crystal structure of M^{pro} (Protein Data Bank, PDB: 2H2Z) using residues 187–306, which are renumbered 1–120. Contact lists were obtained from this protein structure fragment as described in the previous sections. The SMOG (54) server provides GROMACS coordinate (.gro) and topology (.top) files as output. MD simulations were performed in an NVT ensemble using these output files and a leap-frog stochastic dynamics integrator with a time step of 0.0005 ps. Both folding and domain-swapping simulations were performed at T_f, the folding temperature. The folded and unfolded ensembles are equally populated in folding simulations performed at T_f, and several transitions occur between the various populated ensembles or basins. The mechanism of folding or domain swapping can be determined by studying these transitions and the underlying free-energy surface can be calculated. The values of T_f and the number of transitions in the simulations are given in Table S1. Error analyses for both the folding and domain-swapping simulations are given in the Figs. S3 and S4.

Symmetrized SBMs and domain-swapping simulations

Symmetrized SBMs (symSBMs) (43–45,55) use information present only in the monomer structure, simulate two copies of the protein using an SBM, but allow native contacts to form both within the chain and between chains. This framework allows structurally diverse domain-swapped dimer structures to be populated. However, it has been shown (43–45,56,57) that symSBMs can not only be used to understand the mechanism of domain swapping but can also be used to predict the structure of the native (crystallized or obtained using NMR spectroscopy) domain-swapped dimer. Here, we test if MD simulations of a symmetrized form of the SBM used for folding (symSBM) can be used to understand the domain swapping of $M^{pro}C$.

symSBM (43–45) simulations were performed with two chains of the protein: A and B. Each chain has all the potential energy terms of the SBM used for single-chain simulations and shown in the above equation for the potential energy. Additionally, for each intrachain contact between residues i and j present in chain A and i' and j' in chain B, the corresponding interchain contacts between residues i and j' and i' and j are also included in the simulations. A weak harmonic restraint of 1.0 ε/nm^2 with an equilibrium distance of 0.5 nm was applied between the centers of mass of the two protein chains (43,44,56). This was achieved using the pull code in GROMACS (53). The .gro and .top files obtained for the single-chain simulations from the SMOG webserver (54) were modified for the symSBM simulations. The .gro file was modified to include a renumbered second protein chain. The .top file had the following modifications: the atom section was modified to match the .gro file atoms and their numbering, the pairs section was modified to include the intrachain contacts for the new chain and interchain contacts for both chains. The exclusion section was also changed in accordance with the pairs section. The bonds, angles, and dihedral section for the second chain was repeated from the monomer .top file with renumbered atoms. All the bonded and nonbonded parameters were modified such that the corresponding values were same for both the chains.

Analysis of folding simulations

Because protein structure is encoded through contacts in SBMs, and contacts form and break during folding and unfolding transitions, the number of formed native contacts, Q, is often used as an order parameter to understand features of folding such as order of structure formation and the presence or absence of folding intermediates (58,59). Because $M^{pro}C$ is a small single-domain protein, it is not expected to have complex folding features such as backtracking (60), which can be obscured in analysis using only Q. So, we use Q to understand folding features such as barrier heights and folding cooperativity.

A contact is defined to be formed (and its value is set to 1) in a given simulation snapshot if the distance between the two Cα-atoms (which are in contact in the native state) is less than 1.2 times their distance in the native state. Unformed contacts have a value of 0. The number of native contacts formed in a given snapshot is the value, q, of Q for that snapshot. Average contact maps at partial Q-values, i.e., Q = q, with q < the total number of contacts, are calculated by extracting all the simulation frames with a value of $Q = q \pm 12$. The value of a specific contact at this Q is the fraction of extracted frames in which that contact is formed. On average contact maps at partial Q, colored boxes are plotted at (i, j) and (j, i) when a native contact is present between residues i and j. The color of each box represents the value of that contact or how formed that contact is. Thus, partial contact maps represent the average structure of a protein at a particular level of "foldedness" (the value of Q at which they are plotted).

Simulation snapshots with a given Q-value were binned to obtain a histogram. The negative logarithm of this histogram is a plot of the scaled free energy ($\Delta G/k_BT_f$) as a function of Q and we call it the one-dimensional free-energy profile (1DFEP). The baseline of the 1DFEPs was adjusted such that the lowest free energy is reset to 0 and all the other free energies are either up- or downshifted by the same number. In cases in which the folded and unfolded states were not equally populated, single-histogram reweighing was performed to the actual T_f, a temperature at which the folded and the unfolded minima were at equal depths.

Two-dimensional free-energy plots (2DFEPs), with the radius of gyration (R_g; represents the average size of the protein) and Q as the two order parameters plotted on the two axes, were calculated from the folding simulations as follows: Q for each snapshot was calculated as described earlier. R_g-values were obtained using the "g_gyrate" utility of GROMACS (53). All simulation snapshots were binned into a two-dimensional histogram based on their R_g- and Q-values. The negative logarithm of this histogram gives the 2DFEP with R_g and Q as order parameters. These 2DFEPs were also baseline corrected similar to 1DFEPs. The color of each point on the 2DFEP indicates the depth at

that point with darker colors representing a higher depth and indicating that the system is more likely to be present in that state.

Transition analysis and free-energy plots for domain-swapping simulations

symSBM (43–45) simulations are performed using two chains and consequently have two types of contacts: intrachain contacts (Q_{intra}) and interchain contacts (Q_{inter}). Both Q_{intra} and Q_{inter} are calculated for each simulation snapshot. A 1DFEP is plotted, as described for the folding simulations, using $Q_{total} = Q_{inter} + Q_{intra}$ instead of Q. This 1DFEP (Fig. S4) has three minima: the low Q_{total} minimum corresponds to the unfolded ensemble (U), the intermediate Q_{total} minimum corresponds to an ensemble in which one chain is folded whereas the other is unfolded and the high Q_{total} minimum corresponds to the folded ensemble (F). A transition was defined as a piece of the trajectory that starts from the unfolded minimum of the 1DFEP, visits the folded minimum and traverses back to the unfolded minimum. The symSBM simulations were performed for long enough that the number of transitions exceeded the number of transitions present in the folding simulations. The folded minimum contains both domain-swapped states and states where the two chains have each folded to a monomer. A transition is classified as either a domain-swapping or two-monomer transition by visually analyzing the first folded structure that it reaches and classifying it as a domain-swapped or two-monomer structure. The domain-swapping trajectories were further classified based on the loop that formed the hinge region in the domain swapping. For the SBM, this visual classification was confirmed by calculating the total intrachain and interchain contacts of helix α1 for each of the above snapshots. The intrachain contacts for all the structures were then plotted versus the interchain contacts. This plot has three obvious clusters, each from structures belonging to the M, L1, and L4 minima, and the cluster classification exactly matches the visual analysis. The number of transitions for each symSBM is given in Table S1 and the classification of transitions is given in Table S2.

2DFEPs of the symSBM simulations were calculated similar to the folding 2DFEPs described earlier with Q_{inter} and Q_{intra} as the two order parameters. Briefly, all simulation snapshots were binned into a two-dimensional histogram based on their Q_{inter}- and Q_{intra}-values. The negative logarithm of this histogram was baseline corrected (the lowest free-energy value is set to 0 and all other values are accordingly shifted up or down) to give the 2DFEP, and this is plotted with Q_{inter} and Q_{intra} on the two axes. The 2DFEP of symSBM shows the presence of an unfolded minimum (U) and three fully folded minima, the two-monomer minimum (M) with no domain swapping, a minimum with L1-swapped structures (L1), and a minimum with L4-swapped structures (L4) (see Fig. 1 for loop definitions). The connectivity of these minima can be understood by identifying transitions between them. Tight rectangular boundaries were defined for each minimum using Q_{intra}- and Q_{inter}-values (Fig. S5; Table S3) such that all simulation snapshots within those boundaries could be clearly classified as belonging to U, M, L1, or L4. A transition between any two minima A and B was defined as a piece of trajectory that exits minimum A through any of its boundaries and enters minimum B without visiting any other minimum. The number of such transitions between minima is given in Table S4.

Details of the servers and databases used for the analysis of domain-swapping propensity and the structure-based sequence alignment of diverse coronaviral M^{pro}Cs are given in the Supporting Material.

RESULTS

Simulations recapitulate experimental features of M^{pro}C folding

The C-terminal fragment of M^{pro} consisting of residues 187–306 (Fig. 1 *A*) was used as the M^{pro} C-terminal domain in earlier experiments of domain swapping (22–24,26). We use the same boundaries for M^{pro}C in all of our simulations. The 233 contacts used for defining the SBM are marked on the contact map in Fig. 1 *B*, with the locations of the five helices α1–α5 and four loops connecting the helices (L1–L4)

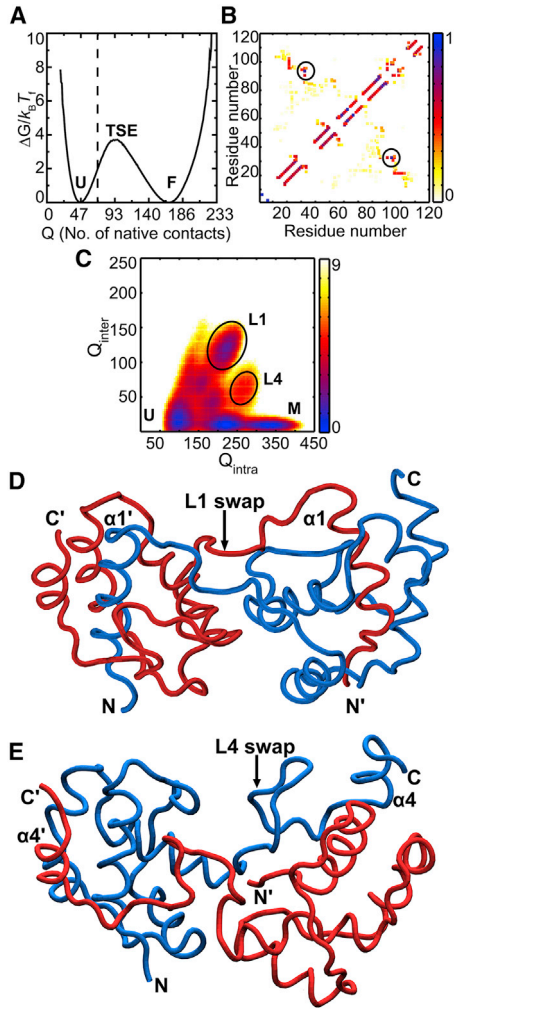

FIGURE 2 The folding and domain swapping of M^{pro}C. (*A*) Free energy plotted as a function of Q, the number of native contacts, shows the population of only two ensembles. The unfolded minimum (U; with few native contacts formed) and the folded minimum (F; many native contacts formed) are separated by a single barrier with a transition state ensemble (TSE) at Q $\approx$ 93. The dashed line marks Q = 70, where the protein is ~30% folded. The average contact map for Q = 70 is shown in (*B*). The colors indicate the probability of contact formation, and the color bar is shown on the right. When the probability is 1, the contact is formed in all snapshots of the simulation in which Q = 70. The highest weight contact, W32-R93, is mostly formed at Q = 70 and is circled in black. (*C*) The two-dimensional free-energy profile for domain-swapping simulations of M^{pro}C is shown. The minima marked are U for unfolded and M for two separate folded monomers. L1 and L4 correspond to domain-swapped structures with domain swapping at the L1 and the L4 hinges, respectively. The colors indicate the population at that position with the color scale on the right and low numbers indicating a large population. (*D*) and (*E*) Representative structures from the L1 and L4 minima viewed from angles similar to Fig. 1 *C*. To see this figure in color, go online.

shown on the axis. The $M^{pro}C$ 1DFEP (Fig. 2 *A*) plotted using folding simulations of SBM performed at T_f has two basins at the unfolded (Q near 0) and the folded (Q near the total number of contacts) ensembles with a single barrier of $\sim 4\ k_B T_f$ separating them. The presence of a single barrier and no other minima between the two basins implies that $M^{pro}C$ folds cooperatively in an all-or-nothing manner. This absence of intermediates, which is in agreement with experiments (22–24), is corroborated in a later section by using a 2DFEP with Q as well as the R_g as an order parameter. The order of structure formation of $M^{pro}C$ is mapped by plotting average contact maps at partial Q-values. Here, we plot the average contact map when $M^{pro}C$ is ~30% folded ($Q = 70$; Fig. 2 *B*; also see Fig. S1 A) because this map shows the formation of the earliest helix-packing interactions, those between $\alpha 1$ and $\alpha 4$, and can thus provide an overall order of structure formation. Contact maps at higher Q-values show more homogeneous contact formation. This contact map also shows that contacts between loops L1 and L4 and contacts between $\alpha 1$ and L4 form early. As an example, one contact that forms early and is also present at times in the unfolded ensemble ($Q \approx 47$; probability of contact formation ~0.44) is the contact (circled in Fig. 2 *B*) between a tryptophan (W32) in L1 and an arginine (R93) in L4. Overall, there is an order to contact formation, i.e., folding is polarized and does not occur homogeneously. Furthermore, the order of helix formation seen in simulations is similar to individual helix stability ($\alpha 1 > \alpha 4 > \alpha 2 > \alpha 3 > \alpha 5$) derived from protection factor data from hydrogen exchange NMR experiments (24). Specifically, helices $\alpha 1$ and $\alpha 4$ form first in simulations, followed by helices $\alpha 2$ and $\alpha 3$, with helix $\alpha 5$ being only partially formed in the folded ensemble.

It has been seen in several proteins that the order of structure formation in domain swapping is similar to that which is seen in single-chain folding (44,61–64). Because the simulated $M^{pro}C$ SBM is able to capture features of its folding dynamics, we tested if a symmetrized version would also be predictive for $M^{pro}C$ domain swapping.

The main domain-swapped ensemble in simulations resembles the experimental domain-swapped structure

As stated earlier, our symSBM simulations have two identical protein chains and do not include any extra information from the domain-swapped dimer structure. During the symSBM simulations, the chains transition from unfolded structures to folded structures and back several times. The folded structures can be either monomeric or domain-swapped. The formed contacts within each chain (Q_{intra}; contains contacts of both chains) as well as the formed contacts between chains (Q_{inter}) are calculated to determine whether the structures are unfolded (low Q_{intra} and Q_{inter}), monomeric (high Q_{intra} but low Q_{inter}), or domain-swapped (intermediate Q_{intra} and Q_{inter} with interchain contacts replacing the intrachain contacts lost because of domain swapping). To understand the overall free-energy landscape of domain swapping, the symSBM simulations are binned and plotted as a 2DFEP with Q_{intra} on the *x* axis and Q_{inter} on the *y* axis (Fig. 2 *C*). Basins with low Q_{inter}-values (*dark regions* present close to the *x* axis) have few formed interchain contacts and show no domain swapping. There are three such basins: the unfolded basin (U) in which both chains are unfolded, the monomer basin with only one chain folded and the other unfolded, and the two-monomer basin (M) in which both chains are folded as monomers but no domain swapping occurs. Several basins are also observed at larger values of Q_{inter}. Of these, two basins have fully folded chains, and these are marked L1 and L4 (Fig. 2 *C*). The structures present in the most populated domain-swapped basin, the L1 basin (Fig. 2 *D*), are similar to the experimentally observed domain-swapped structure (26) of $M^{pro}C$ (Fig. 1 *C*) and domain swap by extending their L1 loops (loop between $\alpha 1$ and $\alpha 2$, with $\alpha 1$ being exchanged between the structures; Figs. 1 *C* and 2 *D*). The structures present in the L4 basin domain swap at the L4 hinge (loop between $\alpha 4$ and $\alpha 5$, with $\alpha 5$ exchanged; Fig. 2 *E*). A transition is defined as a piece of the trajectory that begins in the unfolded basin (U), visits one of the fully folded basins (e.g., M, L1, or L4), and returns to U. Slightly more than half of the transitions (58%) fold to the two-monomer M basin and 42% fold to one of the domain-swapped basins, with 26% folding to the L1 basin and 16% folding to the L4 basin. In agreement with room temperature experiments (26), a quantitative analysis of the trajectories also shows that there is no direct conversion of the L1-swapped structures into two monomers without complete unfolding (Table S4). However, the L4-swapped structures easily convert to two monomers without unfolding, and this dynamic nature may be why they have not been detected in experiment.

Overall, not only is the basin representing the experimentally observed domain-swapped structure (26) more populated (*larger and darker* in Fig. 2 *C*) than other domain-swapped basins, but it is also visited more often. So information present in the monomer structure is sufficient to predict the domain-swapped structure of $M^{pro}C$. Thus, L1 swapping is intrinsic to the $M^{pro}C$ monomer structure and does not require nonnative interactions not present in the symSBM. Because some of the early forming contacts in the folding simulations are present between the L1–L4 loops, we next investigate the role of these contacts in the folding and the domain swapping of $M^{pro}C$ using a modified SBM.

The strengths of specific contacts make the unfolded ensemble more compact and modulate the folding mechanism of $M^{pro}C$

The coarse-grained $M^{pro}C$ SBM used so far encodes structure through weighted contacts between two Cα-beads. These contact weights or strengths are proportional to the

number of interatomic contacts between the two residues represented by their Cα-beads (47,48). Some of the early forming contacts (e.g., the W32-R93) in the folding simulations are between large residues and have large weights. To examine the effect of these weights on the folding mechanism, we also perform folding simulations of $M^{pro}C$ with a commonly used model, here termed SBMe (39), with equally or homogeneously weighted contacts. The only difference between SBM and SBMe, is the weighting schemes (see Methods and compare Fig. 3 *A* with Fig. 3 *B*). The folding mechanism of $M^{pro}C$ changes when simulated with SBMe (Fig. 3 *D*; compare with the contact map at partial Q in Fig. 3 *C*) with only short-ranged intrahelical secondary structural contacts being formed near the unfolded basin and few interhelix or interloop tertiary contacts being present. In contrast, the presence of weights in the SBM promotes the early formation of long-range contacts in the L1–L4 region (Fig. 3 *C*). This allows the compaction of the protein in the unfolded basin, which is less broad and more concentrated at lower values of the R_g in the 2DFEP plotted with Q and R_g as the order parameters (Fig. 3 *E*). R_g gives the average size of the protein. In comparison, the unfolded ensemble of the $M^{pro}C$ SBMe is less compact, with the 2DFEP having a broader unfolded basin whose minimum shifts to larger values of R_g (Fig. 3 *F*). We note in passing that only the folded and the unfolded basins are populated in these 2DFEPs, confirming that $M^{pro}C$ folds cooperatively in both models.

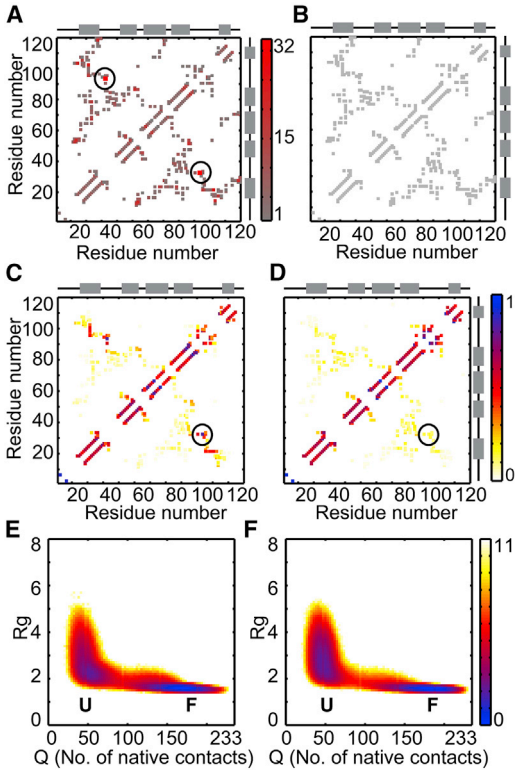

FIGURE 3 Contact weights modulate the nature of the unfolded state. (*A*)The Cα-contact map of $M^{pro}C$ reproduced from Fig. 1 *A* is shown. As before, the color represents the number of all atom contacts between two residues, with the color bar shown on the right. (*B*) Control simulations (SBMe) were also performed using the same contact map but with equally weighted contacts as shown in the contact map. Helices are marked on the axes in gray, whereas loops are in black. (*C*) and (*D*) show average contact maps at Q = 70 for simulations with the original weighted model (*C*; reproduced from Fig. 2 *B*) and the model with homogeneous weights (*D*). A few contacts (*circled*) between L1 and L4 and α1 and L4 form early in the original model (*C*) but do not form in the homogeneous model (*D*). (*E* and *F*) 2DFEPs of R_g versus Q for the original and homogeneous models are shown with the color bar on the right. The unfolded (U) and folded (F) ensembles seen in the 1DFEP (See Fig. 1 *A*) are marked. (*E*) The unfolded ensemble (Q ≈ 47) shows a small spread in R_g-values, whereas the folded ensemble (Q ≈ 186) is compact. (*F*) The unfolded ensemble (Q ≈ 47) has a larger spread in R_g than that seen in the original model. The presence of high weight contacts between L1 and L4 causes these elements to interact even in the unfolded ensemble in the original model reducing the R_g spread in U in Fig. 3 *E*. Also see Fig. S1. To see this figure in color, go online.

Domain swapping requires the swapping of the same specific structural elements between two protein chains (27). The L1 domain swapping seen in the experimentally observed structure (26) of $M^{pro}C$ occurs from the unfolded basin in the SBM simulations and could require a specifically structured, unfolded ensemble. The more homogeneously unstructured unfolded ensemble present in the $M^{pro}C$ SBMe simulations (see Fig. S1) could either reduce the amount of domain swapping or induce incorrect domain swapping. We examine the domain-swapping populations in several variant symmetrized-structure-based simulations in the next section.

Modulating contact weights reduces domain swapping and increases the population of nonnative domain-swapped structures

As stated earlier, the $M^{pro}C$ symSBM includes information only from the monomer structure in the form of weighted contacts. Simulations of this symSBM showed that the experimentally observed L1-swapped structure is the most populated domain-swapped dimer (Figs. 2 *C* and 4 *A*) and are thus predictive. The symmetrized form of SBMe (43,44), termed symSBMe, contains equally weighted contacts (Fig. 3 *B*). A 2DFEP derived from the symSBMe simulations with Q_{intra} and Q_{inter} as the order parameters is plotted in Fig. 4 *B*. The color of the high Q_{inter} basins indicates that little domain swapping exists. Additionally, a basin with domain swapping occurring at the L2/L3 hinges is populated. An analysis of the individual transitions shows that most transitions (78%) fold to the two-monomer (M) basin with only 22% leading to domain-swapped structures. Of the transitions that show domain swapping, 13% fold to the experimentally observed L1-swapped structure (26), whereas 6% fold to an L2-swapped structure, which is not seen in the symSBM simulations. The remaining 3% transitions fold to structures that swap at either the L3 or the L4 hinges. Thus, weighted contacts present in the symSBM simulations not only increase the amount of domain

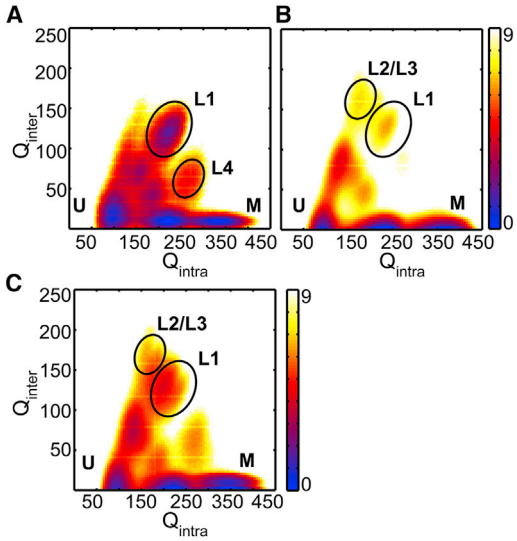

FIGURE 4 Contact weights modulate the nature and the population of domain-swapped structures. Two-dimensional free-energy profiles calculated from domain-swapping simulations that use the original weighted model (*A*) (reproduced from Fig. 2 *C*), the model with homogeneous weights (*B*), and a control model that reduces the weight of the highest weighted contact (*C*) are shown. Colors indicate the population at each position, with the color scale shown on the right. U is the unfolded ensemble, and M is the ensemble in which the two monomers are folded separately. The ensembles labeled Lx represent domain-swapped structures that are swapped at loop Lx. For instance, L1 has domain swapping at L1 (see Fig. 2 *D* for a representative structure). (*A*) The plot from the original model shows two primary domain-swapped minima with L1 being more prominent than L4. (*B*) The model with homogeneous weights has very little domain swapping with a minor L1 population. Additionally, an L2/L3 population mostly absent in (*A*) is also observed. (*C*) When the weight of the highest weighted contact (W32-R93) is reduced in the original weighted model, the population of the incorrect domain-swapped L2/L3 structures increases. The population of the L1 minimum reduces and this minimum also gets more diffuse. Similar results were obtained in domain-swapping simulations in which the weights of two and six of the highest weighted contacts are reduced (see Fig. S6). To see this figure in color, go online.

swapping but also funnel $M^{pro}C$ into the experimentally observed domain-swapped structure.

On closer examination of contact weights or the number of atomic contacts contributing to each contact of $M^{pro}C$, we find that two contacts, namely W32-R93 (between L1 and L4) and W21-E102 (between α1 and L4), have more than 20 interatomic contacts contributing to their weights (see List S1 for a list of contacts and weights). Of these two, W32-R93 (Fig. 1 *A*) is by far the highest weighted (Fig. 1 *B*) with 32 interatomic contacts contributing to it. It also forms early in the folding simulations (Fig. 2 *B*) and we decided to tease apart the effect of this specific contact on domain swapping by resetting only its weight to 1 in symSBM (termed symSBMΔ1; see Methods for details of contact calculations) simulations. Because W32 and, especially, R93 have only a few other low weight contacts (List S1), some of which are interactions of the residue backbones, resetting the W32-R93 contact weight is likely to have similar effects as mutating either the tryptophan or the arginine to a smaller residue like alanine in experiment. The 2DFEP (Fig. 4 *C*) calculated from the symSBM⊿1 simulations shows a reduction in the domain-swapped population with an increase in the population of the L2/L3-swapped minimum. Additionally, the L1 basin gets broader and merges with the L2/L3 minimum. Simulations of symSBM⊿2 (with the weights of both W32-R93 and W21-E102 set to 1) and symSBM⊿6 (with the weights of the top six-highest weighted contacts set to 1) show a similar effect with a further decrease in the domain-swapped population and an increase in the heterogeneity of the observed domain-swapped structures (see Fig. S6 for the 2DFEPs). Because symSBMe is equivalent to symSBMΔN, where N is the total number of contacts and the weights of all contacts are set to 1, setting the weights of increasing numbers of contacts to 1 will lead to decreasing domain swapping and a 2DFEP that is more similar to that of symSBMe (Fig. 4 *B*). Overall, a few high weight contacts promote domain swapping in $M^{pro}C$.

DISCUSSION

L4-swapped structures may be transiently stable even in experiments

Our structure-based simulations recapitulate several experimental features of the folding and domain swapping of $M^{pro}C$ (22–24). In particular, $M^{pro}C$ folds cooperatively, its α1 region gets structured early and it domain swaps primarily at the L1-hinge. However, we also see a minor population of structures swapped at the L4-hinge with an exchange of α5. Analysis of the transitions shows that the L4-swapped structures can transition to the two-monomer minimum without crossing the barrier to the unfolded state. L4 and the following helix α5 fold last in both simulations and experiments (24). Consequently, this region can undergo transient local unfolding and, in the presence of a similarly structured second protein chain, lead to L4-swapped structures in simulations. The domain-swapping hinge predictor, H-predictor (36), also predicts residues 90–95 in L4 as the domain-swapping hinge (34). Collating these observations, we predict that L4-swapped dimers may be transiently populated in experiment but difficult to detect because they can exist in equilibrium with the monomers of $M^{pro}C$.

Although there is no exchange between the monomer and the domain-swapped dimer in conditions that promote protein folding (25), an exchange between monomers and L1-swapped dimers has been experimentally observed at temperatures and urea concentrations in which α5 becomes dynamic (24). It is possible that these conditions will also lead to L4-swapped structures that have an exchange of α5. In fact, a dimeric intermediate that involves the α5 helix has been proposed to enable the formation of L1-swapped dimers without the exposure of the α1 helix to solvent (24). However, in our simulations, L4-

swapped structures do not lead to L1-swapped structures without unfolding. So a transition to L1-swapped structures could require nonnative interactions. Because a W-R interaction mediates L1-swapping, the nearby W21, present on α1, may present a nonnative interaction site for R93 to create an intermediate state that leads to L1-swapping from L4-swapping. However, simulations of SBMs that include nonnative interactions (65–70) will be required to test the existence of this L4-swap to L1-swap route.

A comparison of the sequences of $M^{pro}C$ from other coronaviruses

Our $M^{pro}C$ simulations show that the presence of high weight contacts in the L1–L4 and the α1-L4 regions determines the order of structure formation and, consequently, the structures present in the partially folded (including the unfolded) ensembles in both folding and domain swapping. Because the encounter of chains in these partially folded ensembles leads to domain swapping, the topology of the structures present in these ensembles can promote domain swapping at a particular hinge, in this case the L1-hinge, and suppress swapping at other hinges (the L2 and L3 hinges). Because resetting the weight of just the highest weighted W32-R93 contact is sufficient to both increase the diversity of domain-swapped structures and reduce overall domain-swapping population in simulations (Fig. 4 *C*), this contact could be important for specific ordering in the $M^{pro}C$ unfolded ensemble leading to the L1 domain swapping.

Cation-π interactions have been predicted to stabilize the domain-swapped forms of various proteins (71) with tryptophan being the most common aromatic residue and arginine being the most common cationic residue participating in these interactions (72). We examined how conserved the residues that form the W32-R93 contact are by using a structure-based sequence alignment of the $M^{pro}Cs$ from eight coronaviruses (Fig. S7; details of the viruses and the alignment are given in Supporting Materials and Methods and the $M^{pro}C$ PDB identification numbers in List S2). Tryptophan is present in the equivalent position of residue number 32 in all the chosen coronaviruses except the infectious bronchitis virus (IBV). IBV is also different from the other coronaviruses in that it has a large insertion in L1. Either an arginine or a lysine is present at residue number 93 except in IBV. Thus, the 32–93 contact is a cation-π interaction in all the considered coronaviruses except IBV. The 32–93 contact is made between a lysine and an aspartic acid in IBV. Because both cation-π (W-R/K) and Coulomb (K-D) interactions are stabilizing, the conservation of this attractive contact across coronaviruses may point at its importance for the order of structure formation during the folding of coronavirus $M^{pro}Cs$. To further examine this hypothesis, we calculated the weights of the equivalent of the 32–93 contact in the other coronavirus $M^{pro}C$ structures. We find that the W-R contact is the highest weighted contact in all the structures in which it is present. The W-K contact, although not the highest weighted contact, is still quite highly weighted (see List S2 for contact weights). The K-D contact present in IBV, on the other hand, has a low weight. However, its contact strength may not be directly determined by the number of interatomic contacts that contribute to it but may be increased by the presence of the Coulomb interaction. Overall, the contact structurally equivalent to the W32-R93 is not completely conserved but could still form early in other coronaviruses, determine the order of structure formation, and potentially promote domain swapping. However, W32 is almost completely conserved and another possibility is that although the contact determines the order of structure formation, it is this residue identity that is important for $M^{pro}C$ domain swapping. We examine this hypothesis further in the next section.

The role of W32 in the domain swapping of $M^{pro}C$

In addition to the formation of the correct partially folded ensemble, increased domain swapping also requires that monomer folding stall sufficiently so that interchain contacts form instead of intrachain contacts. Both native-like dihedrals and native contacts are encoded in the symSBM (43,44) and loop formation can stall when the stabilization energy of nonlocal contacts is much higher than that of the dihedrals, which are local in sequence and promote loop formation. Domain swapping can thus occur at loops whose intrachain folding has stalled or become frustrated (56) because of an energetic mismatch between high weight contacts and local dihedrals. In $M^{pro}C$, the high contact weights of the residues in L1 seem to increase the domain-swapping propensity. This loop "frustration" (73) can arise directly from the secondary structural propensity of the L1 loop sequence, RWFLN, surrounding the tryptophan or indirectly through the presence of other residues in the L1 loop that form frustrated interactions in the monomer.

We first examine the secondary structural propensity of RWFLN, using ChSeq (74). ChSeq is a searchable database of chameleon sequences: sequences that adopt different conformations in different proteins. We find that RWFLN is known to exist in both helical and extended conformations within the context of the RWFLNV sequence. Additionally, the secondary structure prediction server, Jpred4 (75), predicts RWFLN to be a helix in the context of a shorter fragment (LYAA…VAMKY) and as an extended β-strand in the context of the full $M^{pro}C$. Thus, it is possible that the sequence of L1 with the conserved W32 can adopt multiple conformations and promote stalling during folding and, consequently, domain swapping.

We next examine the structural and contact (calculated using a 4.5 Å cutoff like before) differences between the entire L1 loops of the monomer (Fig. 1 *A*) and the experimentally crystallized domain-swapped dimer (Fig. 1 *C*). A cation-π

interaction is present between R36 and F37 in the monomer that gets disrupted in the dimer. In the dimer, R36 forms a salt bridge with D77 of the same chain, whereas F37 interacts with both W32 of the same chain as well as R93 of the other chain, which are in the W32-R93 contact. This implies that R36 may be frustrated (73) because it has a choice of two interactions in the monomer, both of which cannot be satisfied at the same time. The F37-W32-R93 interaction is not accessible to F37 in the monomer L1-loop conformation, and thus, only the F37-R93 interaction may form in the monomer to shield the hydrophobic F37 from the solvent.

Mutations that could affect the domain-swapping propensity of $M^{pro}C$

The simulations and the sequence and structure analysis indicate that the domain swapping in $M^{pro}C$ could be promoted by the cation-π contact W32-R93 and the sequence frustration present in the RWFLN sequence of the L1 loop. We propose several W32 and R93 mutations that could reduce domain swapping. However, W32-R93 is the highest weighted contact and such mutations may reduce domain swapping at the expense of protein foldability and stability.

The residues A, S, and Y present at position 32 in the context of the RW32FLN sequence continue to show chameleon-like behavior according to the ChSeq (74) database. So a W32A or a W32S mutation, which abrogates the cation-π interaction, could be used to separate the individual contributions of the contact and sequence frustration to domain swapping. A W32F mutation, on the other hand, is likely to reduce the chameleon character while maintaining the cation-π interaction. A way to potentially increase domain swapping in $M^{pro}C$ is to increase the helical tendency in L1 by introducing a W32E mutation. This is predicted to force the RE pair of the resulting REFLN sequence into a helical conformation (76). The E32 may also be stabilized by a salt bridge with R93.

A double mutation of W32K and R93D, suggested by the comparison of the $M^{pro}C$ sequences of SARS-CoV and IBV (Fig. S7), will retain the attractive contact while removing the chameleon-like nature of the L1 and the L4 sequence stretches (74). These mutations may also help in distinguishing between the effects of contact weights and loop frustration on domain swapping. It should be noted that the predicted mutations that change the local secondary structure propensity are based on either the ChSeq database (74) or secondary structure predictors such as SOPMA (76) and Jpred4 (75). The outcome of these mutations is subject to the accuracy of these predictors and the environment of the mutated residue. The predictors SOPMA and Jpred4 have been reported to have an accuracy of 73.2% (76) and 82.0% (75), respectively.

The symSBM simulations of $M^{pro}C$ also showed a minor population of L4-swapped structures, and we hypothesized in an earlier section that these structures may not only be present in experiment but may also promote the transitions to the L1-swapped state seen experimentally in destabilizing (e.g., high temperature) conditions (24). So, an alternate method to modulate L1 domain swapping could be to reduce the population of the L4-swapped structures. One way to achieve this is to pin α5 to the rest of the protein by strengthening the contacts between them. One of the highest weighted contacts of the human coronavirus NL63 (HCoV NL63 in Fig. S7) is a contact between K68 in α5 and Y112 in α3. $M^{pro}C$ has Q70 and S115 at the equivalent positions. These residues form a contact that is not highly weighted. Furthermore, both pairs of residues (Q, K and S, Y) are surface exposed and form similar numbers of contacts with surrounding residues. Thus, a double mutation of Q70K and S115Y in $M^{pro}C$ may strengthen the interaction between α3 and α5 and reduce the population of L4-swapped structures.

Finally, we suggest mutations that change the interactions in the secondary interface formed by R36 and F37. Specifically, an R36N mutation may increase the solvent exposure of F37 in the monomer, stabilize the dimer, and potentially increase its population. In contrast, an F37N mutation might allow R36 to form a salt bridge with D77, stabilize the monomer, and potentially lead to a reduced population of the domain-swapped dimer.

Biological significance, potential sites for drug design, and application to COVID-19

The domain swapping of $M^{pro}C$ locks M^{pro} into a super active octamer, which may promote the initial phases of viral replication (30). If true, then reducing the domain-swapping propensity of $M^{pro}C$ using mutations suggested in the previous section, may suppress octamer formation and in turn viral replication. The following series of experiments will be required to assess the mutations and understand the biological relevance of the octamer: 1) mutation of $M^{pro}C$ followed by size exclusion chromatography to test if the mutations modulate domain swapping (26), 2) introduction of those mutations that reduce domain swapping into M^{pro} followed by tests for octamer formation in a manner analogous to previous experiments (30), and 3) introduction of mutations that suppress octamer formation into the virus to understand viral replication and growth (77).

Residues of $M^{pro}C$ also participate in the non-domain-swapping dimerization (15,17) of M^{pro}. Previous mutation and drug binding studies have focused on both this non-domain-swapping dimerization interface and residues that are important for enzyme function (16,19–21,25,78–81). Using our simulations, we suggest that binding competent sites near the L1–L4 interaction region could be used to lock $M^{pro}C$ and, in turn, M^{pro} into a non-domain-swapping monomer. Specifically, a small molecule that binds the L1 hinge and stabilizes it in the loop conformation or binds near the L1–L4 interface and reduces the dynamics of α5 could reduce domain swapping in $M^{pro}C$.

A coronavirus similar in sequence to SARS-CoV, SARS-CoV-2, is the cause of the current COVID-19 pandemic

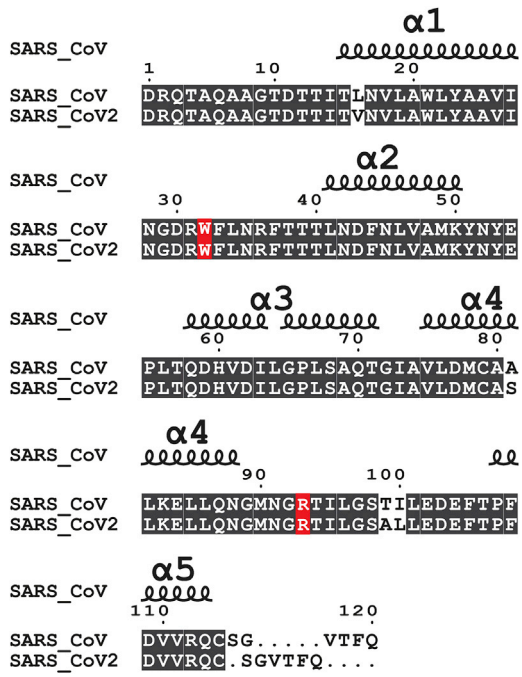

FIGURE 5 Structure-based sequence alignment of M^{pro}Cs from SARS-CoV and SARS-CoV-2. The sequence alignment of the M^{pro}C structures of both SARS-CoV and SARS-CoV-2 (PDB: 6Y2E, residues 187–306) were visualized using ESPript (83). The structures align well except in the C-terminal region after α5, with a mean Cα-root mean-square deviation of 0.85 Å. Conserved residues in both the M^{pro}Cs are shown in white with a black background. Mutations (L16V, A81S, T99A, and I100L) are shown in black with a white background. Residues W32 and R93 are highlighted in red and are conserved in SARS-CoV-2. To see this figure in color, go online.

(49). A sequence alignment (82) of the M^{pro}s from SARS-CoV and SARS-CoV-2 shows that they are 96% identical, i.e., there are 12 mutations with 294 out of the 306 residues being identical. After a structure-based sequence alignment of the two M^{pro}s (SARS-CoV-2 M^{pro} PDB: 6Y2E), residues 187–306 were chosen to represent the M^{pro}C of SARS-CoV-2 (Fig. 5). The M^{pro}Cs of the two viruses are 97% identical, with 116 of 120 residues being identical. Furthermore, a superimposition of the two M^{pro}Cs shows that they align with a Cα-root mean-square deviation of 0.85 Å. Of the four mutations (L16V, A81S, T99A, and I100L) in SARS-CoV-2, only L16V is not present on the surface of the protein. A domain-swapped structure has not been reported for the SARS-CoV-2 M^{pro}C. However, the structure-based sequence alignment (Fig. 5) shows that both W32 and R93 are conserved in SARS-CoV-2 M^{pro}C. Overall, based on the high structure and sequence similarity between the two proteins, we expect that our observations and inferences about SARS-CoV M^{pro}C will also hold true for the M^{pro}C of SARS-CoV-2.

CONCLUSIONS

Domain swapping in the C-terminal domain of the main protease of the SARS coronavirus (M^{pro}C) enables the formation of an octameric assembly of M^{pro}, which has high protease activity at low concentrations of M^{pro} and thus has been hypothesized to be relevant in the early stages of SARS-CoV replication (30). We studied the folding of M^{pro}C using MD simulations of SBMs and found that the simulations could reproduce experimentally observed features such as the folding cooperativity and the overall order of structure formation during folding (22–24). Furthermore, domain-swapping simulations of the same model use information present only in the monomer structure of M^{pro}C and can predict the structure of the experimentally observed domain-swapped dimer (26). By comparing these simulations to those of a variant model, we find that the strengths of contacts in the α1-L1–L4 region of M^{pro}C enable these contacts to form early, direct structure formation during folding and domain swapping and promote domain swapping to the experimentally observed L1-swapped dimer. Specifically, the highest weighted W32-R93 contact between L1 and L4 is important for L1 domain swapping because reducing its weight reduces domain swapping overall while increasing domain swapping at hinges other than L1. This contact is present in most SARS-like coronaviruses implying that the order of structure formation during M^{pro}C folding may be conserved across coronaviruses. An analysis of the sequence and the structure of the W32-containing L1 loop shows that it may be able to fold into multiple conformations, likely promoting swapping at the L1 hinge. Using these observations, we suggest mutations that could reduce the amount of M^{pro}C domain swapping. Such mutations could be used to test the biological relevance of the M^{pro} octamer (30) in viral replication assays (77).

SUPPORTING MATERIAL

Supporting Material can be found online at https://doi.org/10.1016/j.bpj.2020.11.2277.

AUTHOR CONTRIBUTIONS

S.G. designed the study. V.L.T. performed the simulations and analyzed the data. V.L.T. and S.G. wrote the manuscript.

ACKNOWLEDGMENTS

This work was funded by a grant from the Government of India SERB (EMR/2016/003885). We acknowledge the support of the Simons Foundation (Grant No. 287975), the Tata Institute of Fundamental Research and the Department of Atomic Energy, Government of India (12-R&D-TFR-5.04-0900 for funding NCBS core computational facilities and 12-R&D-TFR-5.04-0800).

SUPPORTING CITATIONS

References (84–89) appear in the Supporting Material.

REFERENCES

1. Peiris, J. S., S. T. Lai, …, K. Y. Yuen; SARS Study Group. 2003. Coronavirus as a possible cause of severe acute respiratory syndrome. *Lancet.* 361:1319–1325.
2. Drosten, C., S. Günther, …, H. W. Doerr. 2003. Identification of a novel coronavirus in patients with severe acute respiratory syndrome. *N. Engl. J. Med.* 348:1967–1976.
3. Ksiazek, T. G., D. Erdman, …, L. J. Anderson; SARS Working Group. 2003. A novel coronavirus associated with severe acute respiratory syndrome. *N. Engl. J. Med.* 348:1953–1966.
4. Rota, P. A., M. S. Oberste, …, W. J. Bellini. 2003. Characterization of a novel coronavirus associated with severe acute respiratory syndrome. *Science.* 300:1394–1399.
5. Marra, M. A., S. J. M. Jones, …, R. L. Roper. 2003. The genome sequence of the SARS-associated coronavirus. *Science.* 300:1399–1404.
6. Snijder, E. J., P. J. Bredenbeek, …, A. E. Gorbalenya. 2003. Unique and conserved features of genome and proteome of SARS-coronavirus, an early split-off from the coronavirus group 2 lineage. *J. Mol. Biol.* 331:991–1004.
7. Thiel, V., K. A. Ivanov, …, J. Ziebuhr. 2003. Mechanisms and enzymes involved in SARS coronavirus genome expression. *J. Gen. Virol.* 84:2305–2315.
8. Anand, K., J. Ziebuhr, …, R. Hilgenfeld. 2003. Coronavirus main proteinase (3CLpro) structure: basis for design of anti-SARS drugs. *Science.* 300:1763–1767.
9. Yang, H., M. Yang, …, Z. Rao. 2003. The crystal structures of severe acute respiratory syndrome virus main protease and its complex with an inhibitor. *Proc. Natl. Acad. Sci. USA.* 100:13190–13195.
10. Yang, H., W. Xie, …, Z. Rao. 2005. Design of wide-spectrum inhibitors targeting coronavirus main proteases. *PLoS Biol.* 3:e324.
11. Xue, X., H. Yang, …, Z. Rao. 2007. Production of authentic SARS-CoV M(pro) with enhanced activity: application as a novel tag-cleavage endopeptidase for protein overproduction. *J. Mol. Biol.* 366:965–975.
12. Anand, K., G. J. Palm, …, R. Hilgenfeld. 2002. Structure of coronavirus main proteinase reveals combination of a chymotrypsin fold with an extra alpha-helical domain. *EMBO J.* 21:3213–3224.
13. Zeitler, C. E., M. K. Estes, and B. V. Venkataram Prasad. 2006. X-ray crystallographic structure of the Norwalk virus protease at 1.5-A resolution. *J. Virol.* 80:5050–5058.
14. Humphrey, W., A. Dalke, and K. Schulten. 1996. VMD: visual molecular dynamics. *J. Mol. Graph.* 14:33–38, 27–28..
15. Shi, J., Z. Wei, and J. Song. 2004. Dissection study on the severe acute respiratory syndrome 3C-like protease reveals the critical role of the extra domain in dimerization of the enzyme: defining the extra domain as a new target for design of highly specific protease inhibitors. *J. Biol. Chem.* 279:24765–24773.
16. Chou, C. Y., H.-C. Chang, …, G. G. Chang. 2004. Quaternary structure of the severe acute respiratory syndrome (SARS) coronavirus main protease. *Biochemistry.* 43:14958–14970.
17. Hsu, W. C., H. C. Chang, …, G. G. Chang. 2005. Critical assessment of important regions in the subunit association and catalytic action of the severe acute respiratory syndrome coronavirus main protease. *J. Biol. Chem.* 280:22741–22748.
18. Fan, K., P. Wei, …, L. Lai. 2004. Biosynthesis, purification, and substrate specificity of severe acute respiratory syndrome coronavirus 3C-like proteinase. *J. Biol. Chem.* 279:1637–1642.
19. Shi, J., and J. Song. 2006. The catalysis of the SARS 3C-like protease is under extensive regulation by its extra domain. *FEBS J.* 273:1035–1045.
20. Shi, J., J. Sivaraman, and J. Song. 2008. Mechanism for controlling the dimer-monomer switch and coupling dimerization to catalysis of the severe acute respiratory syndrome coronavirus 3C-like protease. *J. Virol.* 82:4620–4629.
21. Lim, L., J. Shi, …, J. Song. 2014. Dynamically-driven enhancement of the catalytic machinery of the SARS 3C-like protease by the S284-T285-I286/A mutations on the extra domain. *PLoS One.* 9:e101941.
22. Chang, H.-P., C.-Y. Chou, and G.-G. Chang. 2007. Reversible unfolding of the severe acute respiratory syndrome coronavirus main protease in guanidinium chloride. *Biophys. J.* 92:1374–1383.
23. Tsai, M.-Y., W.-H. Chang, …, H.-P. Chang. 2010. Essential covalent linkage between the chymotrypsin-like domain and the extra domain of the SARS-CoV main protease. *J. Biochem.* 148:349–358.
24. Kang, X., N. Zhong, …, B. Xia. 2012. Foldon unfolding mediates the interconversion between M(pro)-C monomer and 3D domain-swapped dimer. *Proc. Natl. Acad. Sci. USA.* 109:14900–14905.
25. Zhong, N., S. Zhang, …, B. Xia. 2008. Without its N-finger, the main protease of severe acute respiratory syndrome coronavirus can form a novel dimer through its C-terminal domain. *J. Virol.* 82:4227–4234.
26. Zhong, N., S. Zhang, …, B. Xia. 2009. C-terminal domain of SARS-CoV main protease can form a 3D domain-swapped dimer. *Protein Sci.* 18:839–844.
27. Bennett, M. J., S. Choe, and D. Eisenberg. 1994. Domain swapping: entangling alliances between proteins. *Proc. Natl. Acad. Sci. USA.* 91:3127–3131.
28. Bennett, M. J., M. P. Schlunegger, and D. Eisenberg. 1995. 3D domain swapping: a mechanism for oligomer assembly. *Protein Sci.* 4:2455–2468.
29. Rousseau, F., J. Schymkowitz, and L. S. Itzhaki. 2012. Implications of 3D domain swapping for protein folding, misfolding and function. *Adv. Exp. Med. Biol.* 747:137–152.
30. Zhang, S., N. Zhong, …, B. Xia. 2010. Three-dimensional domain swapping as a mechanism to lock the active conformation in a super-active octamer of SARS-CoV main protease. *Protein Cell.* 1:371–383.
31. Sun, Z., K. El Omari, …, J. T. Huiskonen. 2017. Double-stranded RNA virus outer shell assembly by bona fide domain-swapping. *Nat. Commun.* 8:14814.
32. Ivanov, D., O. V. Tsodikov, …, T. Collins. 2007. Domain-swapped dimerization of the HIV-1 capsid C-terminal domain. *Proc. Natl. Acad. Sci. USA.* 104:4353–4358.
33. Ho, B.-L., S.-C. Cheng, …, C.-Y. Chou. 2015. Critical assessment of the important residues involved in the dimerization and catalysis of MERS coronavirus main protease. *PLoS One.* 10:e0144865.
34. Huang, Y. Q., X. Kang, …, Z. Liu. 2012. Mechanism of 3D domain swapping for Mpro-C: clues from molecular simulations. *Wuli Huaxue Xuebao.* 28:2411–2417.
35. Shameer, K., G. Pugalenthi, …, R. Sowdhamini. 2011. 3dswap-pred: prediction of 3D domain swapping from protein sequence using Random Forest approach. *Protein Pept. Lett.* 18:1010–1020.
36. Ding, F., K. C. Prutzman, …, N. V. Dokholyan. 2006. Topological determinants of protein domain swapping. *Structure.* 14:5–14.
37. Liu, Z., and Y. Huang. 2013. Evidences for the unfolding mechanism of three-dimensional domain swapping. *Protein Sci.* 22:280–286.
38. Bryngelson, J. D., J. N. Onuchic, …, P. G. Wolynes. 1995. Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins.* 21:167–195.
39. Clementi, C., H. Nymeyer, and J. N. Onuchic. 2000. Topological and energetic factors: what determines the structural details of the transition state ensemble and "en-route" intermediates for protein folding? An investigation for small globular proteins. *J. Mol. Biol.* 298:937–953.
40. Hills, R. D., Jr., and C. L. Brooks, III. 2009. Insights from coarse-grained Gō models for protein folding and dynamics. *Int. J. Mol. Sci.* 10:889–905.
41. Hyeon, C., and D. Thirumalai. 2011. Capturing the essence of folding and functions of biomolecules using coarse-grained models. *Nat. Commun.* 2:487.
42. Gershenson, A., S. Gosavi, …, P. L. Wintrode. 2020. Successes and challenges in simulating the folding of large proteins. *J. Biol. Chem.* 295:15–33.

43. Yang, S., S. S. Cho, …, J. N. Onuchic. 2004. Domain swapping is a consequence of minimal frustration. *Proc. Natl. Acad. Sci. USA.* 101:13786–13791.

44. Cho, S. S., Y. Levy, …, P. G. Wolynes. 2005. Overcoming residual frustration in domain-swapping: the roles of disulfide bonds in dimerization and aggregation. *Phys. Biol.* 2:S44–S55.

45. Ding, F., N. V. Dokholyan, …, E. I. Shakhnovich. 2002. Molecular dynamics simulation of the SH3 domain aggregation suggests a generic amyloidogenesis mechanism. *J. Mol. Biol.* 324:851–857.

46. Ono, K., M. Ito, …, S. Takada. 2015. Dimer domain swapping versus monomer folding in apo-myoglobin studied by molecular simulations. *Phys. Chem. Chem. Phys.* 17:5006–5013.

47. Whitford, P. C., J. K. Noel, …, J. N. Onuchic. 2009. An all-atom structure-based potential for proteins: bridging minimal models with all-atom empirical forcefields. *Proteins.* 75:430–441.

48. Yadahalli, S., and S. Gosavi. 2017. Packing energetics determine the folding routes of the RNase-H proteins. *Phys. Chem. Chem. Phys.* 19:9164–9173.

49. Zhu, N., D. Zhang, …, W. Tan; China Novel Coronavirus Investigating and Research Team. 2020. A novel coronavirus from patients with pneumonia in China, 2019. *N. Engl. J. Med.* 382:727–733.

50. Cho, S. S., Y. Levy, and P. G. Wolynes. 2009. Quantitative criteria for native energetic heterogeneity influences in the prediction of protein folding kinetics. *Proc. Natl. Acad. Sci. USA.* 106:434–439.

51. Sinner, C., B. Lutz, …, A. Schug. 2015. Revealing the global map of protein folding space by large-scale simulations. *J. Chem. Phys.* 143:243154.

52. Sugita, M., and T. Kikuchi. 2013. Incorporating into a Cα Go model the effects of geometrical restriction on Cα atoms caused by side chain orientations. *Proteins.* 81:1434–1445.

53. Hess, B., C. Kutzner, …, E. Lindahl. 2008. GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput.* 4:435–447.

54. Noel, J. K., P. C. Whitford, …, J. N. Onuchic. 2010. SMOG@ctbp: simplified deployment of structure-based models in GROMACS. *Nucleic Acids Res.* 38:W657–W661.

55. Yang, S., H. Levine, and J. N. Onuchic. 2005. Protein oligomerization through domain swapping: role of inter-molecular interactions and protein concentration. *J. Mol. Biol.* 352:202–211.

56. Mascarenhas, N. M., and S. Gosavi. 2016. Protein domain-swapping can be a consequence of functional residues. *J. Phys. Chem. B.* 120:6929–6938.

57. Nandwani, N., P. Surana, …, S. Gosavi. 2019. A five-residue motif for the design of domain swapping in proteins. *Nat. Commun.* 10:452.

58. Chavez, L. L., J. N. Onuchic, and C. Clementi. 2004. Quantifying the roughness on the free energy landscape: entropic bottlenecks and protein folding rates. *J. Am. Chem. Soc.* 126:8426–8432.

59. Cho, S. S., Y. Levy, and P. G. Wolynes. 2006. P versus Q: structural reaction coordinates capture protein folding on smooth landscapes. *Proc. Natl. Acad. Sci. USA.* 103:586–591.

60. Gosavi, S., L. L. Chavez, …, J. N. Onuchic. 2006. Topological frustration and the folding of interleukin-1 β. *J. Mol. Biol.* 357:986–996.

61. Zhuravlev, P. I., G. Reddy, …, D. Thirumalai. 2014. Propensity to form amyloid fibrils is encoded as excitations in the free energy landscape of monomeric proteins. *J. Mol. Biol.* 426:2653–2666.

62. Moschen, T., and M. Tollinger. 2014. A kinetic study of domain swapping of Protein L. *Phys. Chem. Chem. Phys.* 16:6383–6390.

63. Mondal, B., and G. Reddy. 2020. A transient intermediate populated in prion folding leads to domain swapping. *Biochemistry.* 59:114–124.

64. Tsutsui, Y., R. Dela Cruz, and P. L. Wintrode. 2012. Folding mechanism of the metastable serpin α1-antitrypsin. *Proc. Natl. Acad. Sci. USA.* 109:4467–4472.

65. Zarrine-Afsar, A., S. Wallin, …, H. S. Chan. 2008. Theoretical and experimental demonstration of the importance of specific nonnative interactions in protein folding. *Proc. Natl. Acad. Sci. USA.* 105:9999–10004.

66. Azia, A., and Y. Levy. 2009. Nonnative electrostatic interactions can modulate protein folding: molecular dynamics with a grain of salt. *J. Mol. Biol.* 393:527–542.

67. Zhang, Z., and H. S. Chan. 2010. Competition between native topology and nonnative interactions in simple and complex folding kinetics of natural and designed proteins. *Proc. Natl. Acad. Sci. USA.* 107:2920–2925.

68. Kim, Y. C., and G. Hummer. 2008. Coarse-grained models for simulations of multiprotein complexes: application to ubiquitin binding. *J. Mol. Biol.* 375:1416–1433.

69. Cheung, M. S., and D. Thirumalai. 2006. Nanopore-protein interactions dramatically alter stability and yield of the native state in restricted spaces. *J. Mol. Biol.* 357:632–643.

70. Davtyan, A., N. P. Schafer, …, G. A. Papoian. 2012. AWSEM-MD: protein structure prediction using coarse-grained physical potentials and bioinformatically based local structure biasing. *J. Phys. Chem. B.* 116:8494–8503.

71. Dehouck, Y., C. Biot, …, M. Rooman. 2003. Sequence-structure signals of 3D domain swapping in proteins. *J. Mol. Biol.* 330:1215–1225.

72. Gallivan, J. P., and D. A. Dougherty. 1999. Cation-pi interactions in structural biology. *Proc. Natl. Acad. Sci. USA.* 96:9459–9464.

73. Ferreiro, D. U., E. A. Komives, and P. G. Wolynes. 2014. Frustration in biomolecules. *Q. Rev. Biophys.* 47:285–363.

74. Li, W., L. N. Kinch, …, N. V. Grishin. 2015. ChSeq: a database of chameleon sequences. *Protein Sci.* 24:1075–1086.

75. Drozdetskiy, A., C. Cole, …, G. J. Barton. 2015. JPred4: a protein secondary structure prediction server. *Nucleic Acids Res.* 43:W389–W394.

76. Geourjon, C., and G. Deléage. 1995. SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. *Comput. Appl. Biosci.* 11:681–684.

77. Stobart, C. C., N. R. Sexton, …, M. R. Denison. 2013. Chimeric exchange of coronavirus nsp5 proteases (3CLpro) identifies common and divergent regulatory determinants of protease activity. *J. Virol.* 87:12611–12618.

78. Barrila, J., S. B. Gabelli, …, E. Freire. 2010. Mutation of Asn28 disrupts the dimerization and enzymatic activity of SARS 3CL(pro). *Biochemistry.* 49:4308–4317.

79. Chen, S., T. Hu, …, X. Shen. 2008. Mutation of Gly-11 on the dimer interface results in the complete crystallographic dimer dissociation of severe acute respiratory syndrome coronavirus 3C-like protease: crystal structure with molecular dynamics simulations. *J. Biol. Chem.* 283:554–564.

80. Hu, T., Y. Zhang, …, X. Shen. 2009. Two adjacent mutations on the dimer interface of SARS coronavirus 3C-like protease cause different conformational changes in crystal structure. *Virology.* 388:324–334.

81. Kuo, C.-J., and P.-H. Liang. 2015. Characterization and inhibition of the main protease of severe acute respiratory syndrome coronavirus. *ChemBioEng Rev.* 2:118–132.

82. Altschul, S. F., W. Gish, …, D. J. Lipman. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403–410.

83. Robert, X., and P. Gouet. 2014. Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res.* 42:W320–W324.

84. Efron, B., and C. Stein. 1981. The jackknife estimate of variance. *Ann. Stat.* 9:586–596.

85. Miller, R. G. 1974. The jackknife–a review. *Biometrika.* 61:1--15.

86. Dawson, N. L., T. E. Lewis, …, I. Sillitoe. 2017. CATH: an expanded resource to predict protein function through structure and sequence. *Nucleic Acids Res.* 45:D289–D295.

87. Russell, R. B., and G. J. Barton. 1992. Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins.* 14:309–323.

88. Roberts, E., J. Eargle, …, Z. Luthey-Schulten. 2006. MultiSeq: unifying sequence and structure data for evolutionary analysis. *BMC Bioinformatics.* 7:382.

89. Zhang, L., D. Lin, …, R. Hilgenfeld. 2020. Crystal structure of SARS-CoV-2 main protease provides a basis for design of improved α-ketoamide inhibitors. *Science.* 368:409–412.