Natural Product Communications
May 2019: 1–11

DOI: 10.1177/1934578X19849202
journals.sagepub.com/home/npx

Structure of the SARS-Unique Domain C From the Bat Coronavirus HKU4

Andrew J. Staup[1], Ivon U. De Silva[1], Justin T. Catt[1], Xuan Tan[1], Robert G. Hammond[1], and Margaret A. Johnson[2]

Abstract

Coronaviruses (CoVs) that cause infections such as severe acute respiratory syndrome (SARS) and Middle East respiratory syndrome phylogenetically originate from bat CoVs. The coronaviral nonstructural protein 3 (nsp3) has been implicated in viral replication, polyprotein cleavage, and host immune interference. We report the structure of the C domain from the SARS-Unique Domain of bat CoV HKU4. The protein has a frataxin fold, consisting of 5 antiparallel β strands packed against 2 α helices. Bioinformatics analyses and nuclear magnetic resonance experiments were conducted to investigate the function of HKU4 C. The results showed that HKU4 C engages in protein-protein interactions with the nearby M domain of nsp3. The HKU4 C residues involved in protein-protein interactions are conserved in group 2c CoVs, indicating a conserved function.

Keywords

SARS-unique domain, coronavirus, non-structural protein, NMR, chemical shift perturbation, MERS, functional annotation

Received: November 10th, 2018; Accepted: January 18th, 2019.

Coronaviruses (CoVs) are known as single-stranded enveloped RNA viruses which possess positive-sense RNA genomes.[1] They are responsible for potentially lethal infections related to the human respiratory system, such as severe acute respiratory syndrome (SARS) and Middle East respiratory syndrome (MERS). Additionally, CoVs cause other types of infections, including acute respiratory distress and acute lung injury syndrome; upper and lower respiratory disease ranging from mild to severe; and gastrointestinal disease.[2]

While SARS and MERS CoVs are similar to other CoVs in genomic composition, they belong to two different phylogenetic lineages. Betacoronaviruses can be divided into 4 lineages known as A, B, C, and D. The SARS CoV belongs to the lineage B,[3] while the MERS CoV belongs to the C lineage. These two types of betacoronaviruses are predicted to originate zoonotically from bats, which are known to be a reservoir host for CoVs.[3]

Additionally, some bat species such as the lesser bamboo bat (*Tylonycteris pachypus*) can act as hosts for the bat CoV HKU4, which is a betacoronavirus in the C lineage.[4] Since this virus exhibits a high similarity to MERS with respect to the spike protein, ribonucleic acid (RNA) polymerase, and nucleocapsid protein, HKU4 is a relative of the MERS CoV.[5]

CoVs contain one of the largest RNA genomes, which encodes two important polyproteins known as the replicase polyproteins 1a and 1ab.[6] Viral proteases catalyze the cleavage of these polyproteins into several nonstructural proteins (nsp) which are involved in the formation of the replicase-transcriptase complex (RTC). This complex carries out RNA synthesis, RNA processing, and interference with the host cell innate immune system.[7,8] The largest nsp present in this complex is nsp3 which is the most variable region in the CoV genome.[9] nsp3 is a complex protein which may contain one or several macrodomains, several conserved uncharacterized domains, a papain-like cysteine protease, and transmembrane regions. Therefore, it can be identified as a multifunctional protein.[10]

The macrodomain is known to be a conserved region of nsp3 present in different viral species including CoVs.[8] This macrodomain plays a major role in the virulence of CoVs

[1] Department of Chemistry, 1720 2nd Avenue S. CHEM 201, University of Alabama at Birmingham, AL, USA
[2] Department of Chemistry, 1720 2nd Avenue S. CHEM 274, University of Alabama at Birmingham, AL, USA

Corresponding Author:
Margaret A. Johnson, Department of Chemistry, University of Alabama at Birmingham, 1720 2nd Avenue S. CHEM 274, Birmingham, AL 35294-0006, USA.
Email: maggiejohnson@uab.edu

since they can infect through protein-protein interactions related to host immune system.[11] There is an additional macrodomain called the SARS-unique domain (SUD) which contains two divergent domains and a conserved domain. This three-domain structure is capable of binding G-quadruplex nucleic acids[10,12] and of interacting with p53 to target it for degradation by stabilizing the RCHY E3 ligase.[13] We report here the first structural study of the HKU4 SARS-unique region.

We show that HKU4 C adopts a frataxin-like fold characteristic of the SUD-C domain. There is minimal conservation relative to group 2b viruses. Therefore, the function of this protein is likely to be specific to group 2c viruses including the *Tylonycteris* CoVs and MERS. We also characterized the folding of the neighboring SUD-M macrodomain and show that the M domain is a 150-residue independently folded, monomeric domain for which the fold does not change in the presence of the C domain.

Uniformly ^{15}N,^{13}C-labeled HKU4 C domain was expressed and purified from *Escherichia coli*. Domain boundaries were predicted employing secondary structure prediction with Jpred[14] and multiple sequence alignment with Fold and Function Assignment System,[15] resulting in a construct comprising the residues 1445 to 1522 of nsp3. A 6xHis fusion tag was employed to assist purification, leaving the non-native residues Ser-His-Met at the N-terminus after cleavage.

Assignment of 92% of the observable resonances of the backbone and side chains was achieved using triple resonance experiments.[16] All backbone ^{15}N, ^{1}H, and C′ atoms excluding the Ser-His-Met tag remaining from cloning were assigned. The assigned chemical shift list was deposited in the BioMagResBank[17] under the identifier 30531. Structure determination was accomplished utilizing 3D ^{15}N- and ^{13}C-resolved nuclear Overhauser effect spectroscopy (NOESY) experiments.

HKU4 C adopts a frataxin-like fold,[18] consisting of 5 antiparallel β strands packed against 2 α helices (Figure 1a). The helices are located at the N- and C- termini of the protein and are oriented parallel to each other. The hydrophobic core of

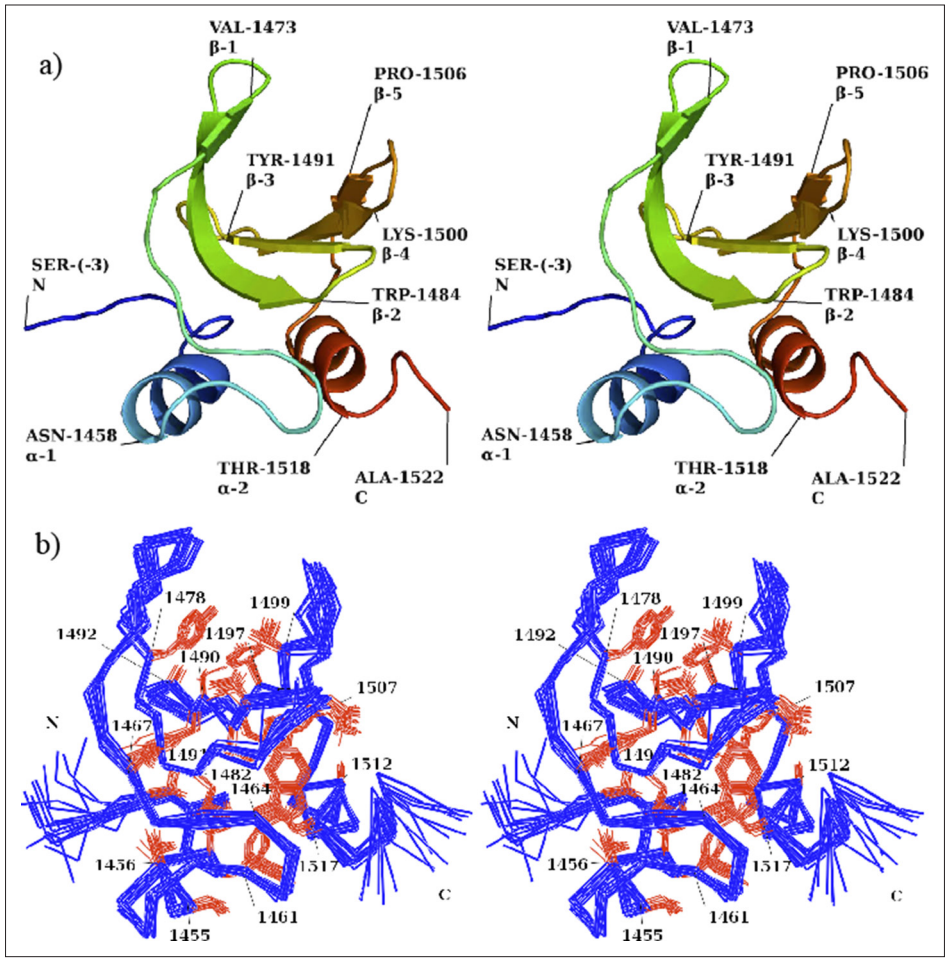

Figure 1. Solution nuclear magnetic resonance structure of HKU4 C. (a) The conformer with minimal RMSD to the mean coordinates of the ensemble of 20 conformers is displayed in stereo, and the secondary structures are indicated. (b) The ensemble of 20 conformers is shown in stereo. The N-terminus and C-terminus are labeled. RMSD, root-mean-square deviation.

Table 1. Statistics of the NMR Structure Determination of HKU4 C.

Quality	Value
Cyana-minimized target function, Å^2	2.20 ± 0.82
NOEs per residue	24.56
Long-range NOEs per residue (*i-j* ≥5)	8.94
RMSD drift (1st-7th cycle), Å	1.20
RMSD to the mean coordinates, Å	
Backbone	0.37 ± 0.08
Heavy atom	0.69 ± 0.11
Residual dihedral angle violations number ≥ 5°	0
Ramachandran statistics[22] (%)	
Most favored regions	85.1
Allowed regions	11.5
Disallowed regions	3.3
RMSD from ideal geometry	
Bond angles, degrees	0.2
Bond lengths, Å	0.001
Mean vdW contribution to the target function (A^2)	0.22 ± 0.05
vdW violations	0

NMR, nuclear magnetic resonance; vdW, van der Waals; RMSD, root-mean-square deviation; NOE, nuclear Overhauser effect;
Structure validation criteria for HKU4 C. The results indicate consistency in the ensemble of conformers and a well-defined fold.

the protein is comprised of residues from β strands 2 to 4 as well as α helices 1 and 2. The hydrophobic core residues of the ensemble (Figure 1b) are well defined across all conformers. The fold is classified in Structural Classification of Proteins[19] as similar to the N-terminal domain of CyaY.[20] The fold has also been identified in the CyaY family of regulatory proteins; frataxins, which are involved in the formation of iron-sulfur complexes in mitochondria; and C domains of the coronaviral SARS-unique region.

The structure consists of a β sheet composed of 5 β strands flanked by an N-terminal and a C-terminal α helix. The helix α1 consists of the residues 1448 to 1456. This is followed by several loop residues leading into the first β strand β1 (1471-1473) and is immediately followed by a β turn (1474-1476). A curved sheet formed by the remaining 4 β strands connected to one another by β turns is observed in the following residues: β2 1476 to 1483, β3 1487 to 1491, β4 1496 to 1499, β5 1504 to 1506. A final C-terminal helix α2 consists of residues 1510 to 1519.

Table 1 summarizes the input and statistics of the solution nuclear magnetic resonance (NMR) structure determination. The target function of the ensemble of 20 conformers of Figure 1b is 2.20 ± 0.82 Å^2. The backbone root-mean-square deviation (RMSD) is 0.37 ± 0.08 Å^2 measured from residues 1449 to 1519. Validation of the NMR structure was carried out on the basis of calculation input requirements, agreement with NMR observables, local residue geometry, and the fold of the protein.[21] The ensemble of 20 conformers representing the solution structure (RMSD 0.37 Å^2) is well defined by the experimental NOE and dihedral angle constraints, including 9 long-range NOEs per residue. Taken together, the results indicate a high-quality structure determination.

Alignment of the sequence and structure of HKU4 C with protein databases identified similarities to coronavirus C domains of SARS, mouse hepatitis virus (MHV), and HKU9 (Table 2). The highest similarity was obtained to the SARS C domain, with a distance matrix alignment (DALI) score of 9.6 being obtained.

DALI and TM-align were used to identify other proteins which are structurally similar to HKU4 C. Both DALI and TM-align predict the SARS C domain, the HKU9 C domain from the *Rousettus* BtCoV, and the MHV nsp3 C domain as the top three homologous proteins for HKU4 C. Both DALI and TM-align predict almost a similar sequence identity (25%) for HKU9 C, while the lowest sequence identity can be observed for SARS C according to both servers (Table 2).

A more detailed analysis of conservation patterns was carried out using the program ConSurf.[25,26] The residues

Table 2. Structural and Sequence Conservation of HKU4 Orthologues.

Server	PDB id	Protein name	RMSD	%ID	DALI/TM-align score
DALI	2kaf	SARS nsp3 C domain	2.0	10.0	7.0
	4ypt	MHV nsp3 C domain	2.5	22.0	6.5
	5utv	SARS unique fold in *Rousettus* BtCoVHKU9	2.5	25.0	9.6
TM-align	2kaf	SARS nsp3 C domain	1.79	17.2	0.63
	5utv	SARS unique fold in *Rousettus* BtCoVHKU9	2.22	24.7	0.70
	4ypt	MHV nsp3 C domain	2.38	17.8	0.64

DALI, distance matrix alignment; SARS, severe acute respiratory syndrome; PDB
, Protein Data Bank; MHV, mouse hepatitis virus; TM, template modeling; RMSD, root-mean-square deviation;
This table shows the top three homologous proteins predicted by DALI and TM-align. RMSD is a quantitative evaluation of similarity in protein structure. %ID is the percentage of identities among amino acid sequences of different proteins. DALI scores which are higher than 2.0 indicate high similarities to the query sequence.[23] TM-align score is a measurement of the similarity in 3D fold among proteins.[24]

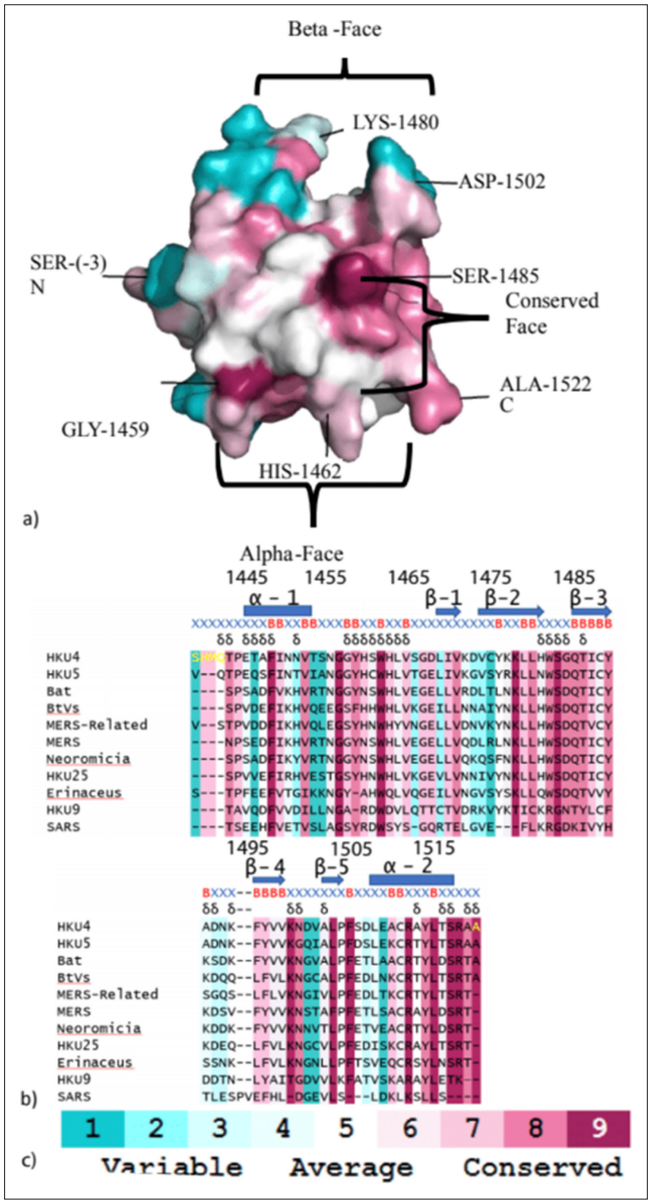

Figure 2. (a) Molecular surface of HKU4 C colored according to residue conservation among orthologues. Conservation levels are as follows: 9 represents the most highly conserved residues, while 1 represents the most variable residues. The structure is shown with the conserved face closest to the viewer, with the α face and β face indicated on the structure. The variable face of HKU4 C is on the opposite side of the protein relative to the viewer, and therefore cannot be seen from this perspective. (b) Multiple sequence alignment. Secondary structures of HKU4 C residues are indicated above them by rectangles and arrows, which represent α helices and β strands, respectively. Solvent-exposed residues of HKU4 C are represented by blue Xs, while residues buried in the core are represented by red Bs. Finally, HKU4 C residues marked by a "δ" experienced a chemical shift above the threshold in the HKU4 MC construct. (c) A graphical key of conservation levels determined by ConSurf is shown.[25] Residues in the sequence alignment that are yellow lack sufficient data to determine their conservation reliably.

F1451, G1459, W1464, K1480, S1485, K1500, L1505, F1507, R1514, L1517, S1519, and R1520 were most conserved with respect to orthologues. In addition, the following regions of the protein were most conserved overall: 1459 to 1467, 1480 to 1491, and 1513 to 1520 (Figure 2). A comparison with secondary structure and surface accessibility properties showed that the highly conserved regions of the protein correspond with the β strands 2 and 3 (residues 1480-1491) and α helix 2 (residues 1513-1520). Some of these residues, such as F1451, W1464, and F1507, are buried in the hydrophobic core of the protein, indicating that their interactions are likely to be necessary to maintain the fold of the protein. However, the conserved region 1459 to 1467, located between α helix 1 and β strand 2, contains several residues on the surface of the protein.

To facilitate further discussion, we define the surface regions of the protein as follows: the α face, which is composed of solvent-exposed residues from both α helices; the β face, which is located on the opposite side of HKU4 C relative to the α face and is composed of β strand and β turn residues; the conserved face, which contains several conserved residues in the C lineage; and the variable face, which contains several non-conserved residues and is located on the opposite side of HKU4 C relative to the conserved face.

HKU4 C was subjected to functional annotation using the prediction servers COACH, COFACTOR,[27] TMSite,[28] and RaptorX.[18] The results indicated that HKU4 C may bind ligands such as peptides and ions (Table 3). Based on the confidence scores (*C*-scores) predicted by both TMSite and COACH, binding of acetate ion to HKU4 C shows the highest probability. Both TM-Site and COACH predict the same binding site for acetate ion on HKU4 C, which is located on the variable face, while Raptor X predicts a binding site on the β face.

A second functional possibility is that of a peptide-binding site. Peptides are also possible ligands according to TM-Site and COFACTOR. The binding site for peptides predicted by TM-Site would be situated on the α face, near the C terminus (Table 3), while multiple binding sites are predicted by COFACTOR.

In addition to these ligands, calcium ions are also capable of binding with HKU4 C according to the predictions generated by TM-Site and Raptor X (Table 3). The probability of binding calcium ions to HKU4 C has been indicated by TM-Site with a *C*-score value of 0.21. Additionally, Raptor X is also able to predict that calcium ions can bind to the amino acid residues such as K1480 and G1486 which are located on the β face.

We next investigated how the HKU4 C protein might interact with other domains within the nsp3 protein. A prime candidate for such interactions is the neighboring domain of the SARS-unique region, the M domain. The C domain occurs together with the M domain in several CoVs[9,10,27] and has been found to modulate the binding specificity of the M domain.[10]

Table 3. Bioinformatics Results Obtained for Predicted Ligands and Their Possible Binding Sites on HKU4 C.

Ligand	Server	Predicted binding residues and regions	Structurally similar proteins	*C*-score
Peptide	TM-Site	A1515, S1519 (helical face)	Pentameric ligand gated ion channel from *Erwinia chrysanthemi* (3zkrD)	0.22
	COFACTOR	K1495, F1496, Y1497, V1498, V1499 (β face)	BmKX(irjiC)	0.1
		I1452, N1454, S1468, H1483, T1518 (around whole protein)	Knl1/Nsl1 complex (4nf9A)	0.1
Acetate	TM-Site	P1447, K1495, D1509, L1510 (variable face)	Dehydroascorbate reductase from *Pennisetum americanum* (5evoA)	0.21
	COACH	P1447, K1495, D1509, L1510 (variable face)	Dehydroascorbate reductase from *Pennisetum americanum* (5evoA)	0.7
	Raptor X	K1500, N1501, D1502 (β face)		
Calcium ion	Raptor X	K1480, G1486 (β face)		
	TM-Site	P1447, K1495, D1509, L1510 (variable face)	Dehydroascorbate reductase from *Pennisetum americanum* (5evoA)	0.21

Two additional constructs were cloned, purified, and subjected to preliminary NMR analysis: first, a construct spanning the residues 1319 to 1445 (M domain); and second, a didomain construct spanning the nsp3 residues 1319 to 1522, HKU4 MC. Both proteins were purified using similar protocols to HKU4 C and were monomeric in solution as shown by gel filtration chromatography (Supplemental Figure 1). Figure 3 shows analyses of the heteronuclear single quantum coherence (HSQC) spectra of these constructs. The assigned HSQC spectrum of HKU4 C is shown in Figure 3a, with 95 crosspeaks. The HSQC profile is shown in Figure 3b. The profile indicates a pure, well-folded protein sample that is suitable for NMR characterization due to the homogeneity of peak distributions and intensities.[29] A peak count and profile of the HKU4 M domain revealed a globular protein with high chemical shift dispersion and even peak intensities, indicating that the domain is independently folded in the absence of the C domain (Figure 3b,c). A total of 141 peaks were observed, which is within 10% of the expected 145.

A detailed comparison of the spectra showed that the C domain overlays closely in the absence and in the presence of the covalently attached M domain (Figure 3e). This indicates that the fold of the protein is maintained in the didomain construct and there are no major conformational changes. However, significant chemical shift perturbations to the C spectrum were observed, indicating either an interaction between the two domains or minor conformational changes that occur in the didomain construct. These perturbations were mapped on the HKU4 C structure in Figure 4a. The perturbations are most intense on the conserved face of the protein. A similar analysis was carried out for the M domain, for which resonance assignments have not been obtained. Again, the HSQC spectra of this domain overlay closely in the absence and in the presence of the covalently attached C domain (Figure 3f). There are 31 residues in HKU4 C that experienced a chemical shift perturbation greater than or equal to 0.022 ppm (Figure 4c). There are 50 residues in HKU4 M that experience a chemical shift perturbation greater than or equal to 0.008 ppm. These observations are consistent with an interaction between the two domains in the didomain protein.

The flexibility of the linker region between the domains of HKU4 MC was further investigated using NMR relaxation experiments. $^{15}N\{^{1}H\}$-NOE plots of HKU4 C and the individual domains of HKU4 MC are shown in Figure 5. In the HKU4 C residues shown in Figure 5a, residues with NOE values less than 0.6 are the terminal residues Q1445 and A1521, indicating no flexible residues within HKU4 C. Although HKU4 C maintains its overall stability in the didomain protein, Q1445 loses flexibility as it is no longer a terminal residue, while R1520 becomes more flexible, as shown in Figure 5b. These experiments probed fast dynamics on the ps-ns timescale and do not preclude slower timescale conformational exchange.

The HSQC spectrum of HKU4 MC also shows a protein that is primarily globular (Figure 3d); 229 peaks were observed for this protein, which is close to the sum of the 95 peaks observed for HKU4 C and 141 observed for HKU4 M, minus the 3 peaks of the shared Q1445 residue.

There are 4 residues in the HKU4 M domain of the didomain protein with I_{rel} values within the 0.2 to 0.4 range, designating N-terminal or C-terminal residues that are flexible in the HKU4 M domain,[10] as shown in Figure 5c. There are 7 more residues within HKU4 M with values below 0.6, bringing the total number of flexible residues to 11. These data suggest a short linker between HKU4 M and HKU4 C domains no longer than 4 residues in length.

Additionally, there are 16 residues in the HKU4 MC didomain NOE plot that are flexible, as shown in Supplemental Figure 1. The additional 2 residues not accounted for in the HKU4 MC spectrum could be perturbed peaks that could not be assigned to either of their individual

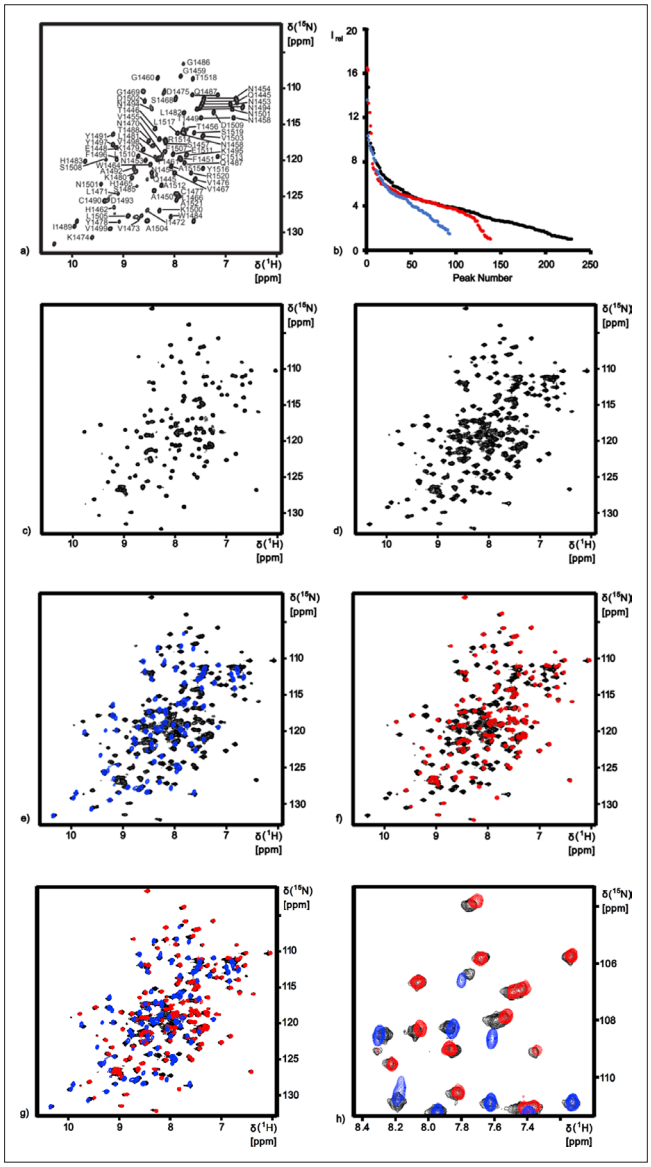

Figure 3. Comparison of [^{15}N, ^{1}H] HSQC spectra. (a) Labeled [^{15}N, ^{1}H] HSQC spectrum of HKU4 C domain. (b) [^{15}N, ^{1}H] HSQC profiles for HKU4 C (blue), HKU4 M (red), and HKU4 MC domains (black), which were calculated by dividing the peak intensities by the noise, concentrations, and the minimum peaks, respectively. (c) [^{15}N, ^{1}H] HSQC spectrum for HKU4 M. (d) [^{15}N, ^{1}H] HSQC spectrum for HKU4 MC. (e) Overlays of the HKU4 C [^{15}N, ^{1}H] HSQC spectrum (blue) with the HKU4 MC [^{15}N, ^{1}H] HSQC spectrum (black). The protein ratios were 2:3. The overlay was calibrated by shifting the HKU4 MC spectrum δ ^{1}H 0.003 ppm and δ ^{15}N 0.295 ppm and was then used to calculate the residue perturbations. (f) Overlays of the HKU4 M [^{15}N, ^{1}H] HSQC spectrum (red) with the HKU4 MC [^{15}N, ^{1}H] HSQC spectrum (black). The protein ratios were 1:1. The HKU4 M spectrum was calibrated by shifting the spectrum δ ^{15}N −0.060 ppm and was then used to calculate the peak perturbations. (g) Overlay of all 3 [^{15}N, ^{1}H] HSQC spectra, HKU4 C (blue), HKU4 M (red), and HKU4 MC (black) using the previous calibrations. The protein ratios were 2:2:3. (h) Expansion showing selected perturbed residues from the overlay in panel (g). The expansion is from δ ^{1}H 7.05 to 8.45 ppm and δ ^{15}N 103.5 to 111.5 ppm. HSQC, heteronuclear single quantum coherence.

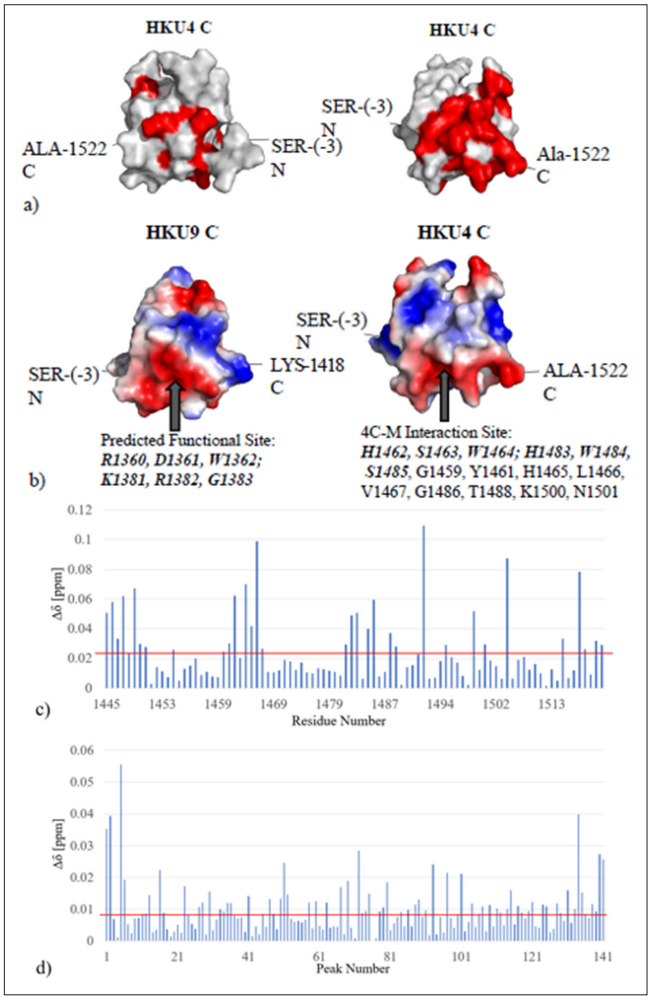

Figure 4. Interdomain interactions of HKU4 C. (a) Surface representation of HKU4 C with chemical shift perturbations mapped on the structure. The left panel shows the variable face of HKU4 C, while the right panel shows the conserved face of HKU4 C. Any perturbation above 0.022 ppm is indicated on the surface in red. (b) Electrostatic surface representations of the C domains of BtCoVHKU9 (left) and HKU4 (right). The location of a predicted functional site of HKU9 C is indicated in its panel. The residues on the conserved face of HKU4 C that interact with HKU4 M in the didomain construct are referenced in the bottom right panel. (c) Histogram of chemical shift perturbations of all HKU4 C residues in the 4MC didomain. Any residues with values above the red line (0.022 ppm) are considered perturbed. (d) Histogram of chemical shift perturbations of HKU4 M [^{15}N, ^{1}H] crosspeaks in the 4MC didomain. Any peaks with values above the red line (0.008 ppm) are considered perturbed.

domain. Further characterization of the didomain would need to be performed for elucidation.

The solution structure of the HKU4 C domain revealed a frataxin-like fold.[30] The fold family has similarity to other proteins present in CoV SUDs, implying the potential for a conserved function. This observation adds to other lines of evidence suggesting the SUD provides essential functions

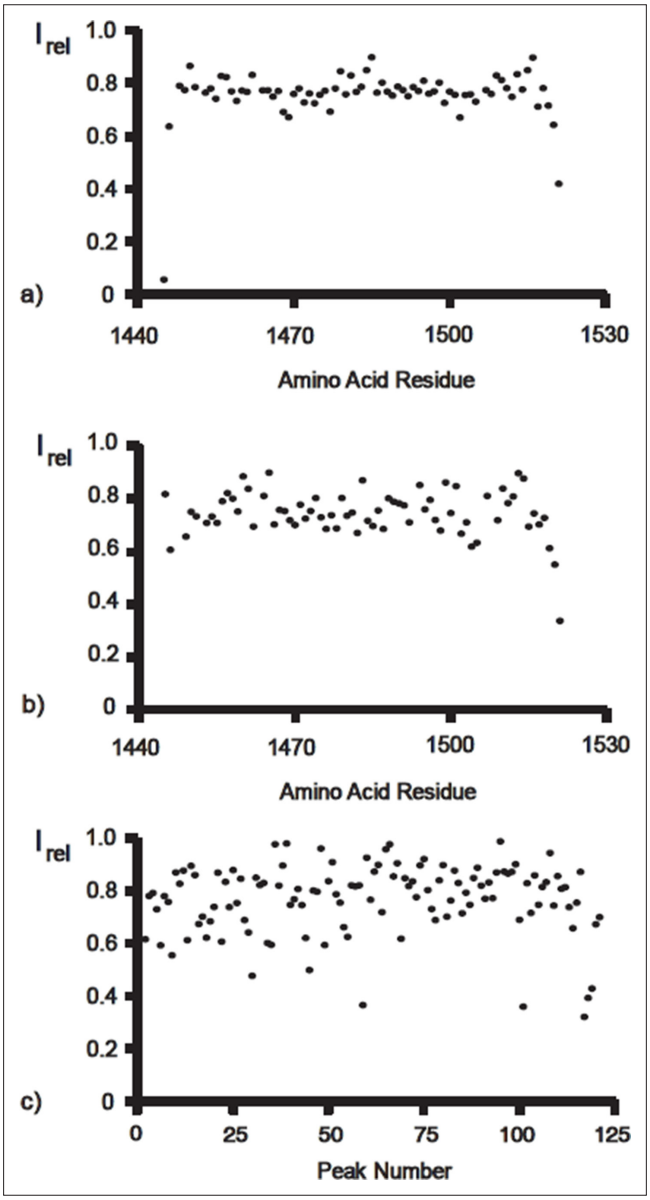

Figure 5. $^{15}N\{^{1}H\}$-NOE values for HKU4 protein constructs. (a) $^{15}N\{^{1}H\}$-NOE values, I_{rel}, plotted versus the sequence of the HKU4 C residues. (b) $^{15}N\{^{1}H\}$-NOE values, I_{rel}, plotted versus the crosspeaks corresponding to HKU4 C residues in the HKU4 MC didomain. (c) $^{15}N\{^{1}H\}$-NOE values, I_{rel}, plotted versus HKU4 M crosspeaks in the didomain, in order of increasing intensity. I_{rel} lower than 0.6 indicates flexibility in the residue. Q1445 and A1521 in the HKU4 C single domain fall below this threshold, as they are residues found in the N- and C-terminal respectively. In the didomain, R1520 gains greater flexibility, while Q1445 loses flexibility due to no longer being a terminal residue. There are 11 residues in the HKU4-M domain that are flexible. NOE, nuclear Overhauser effect.

for CoV virulence[31] and replication.[13,32] Despite this, sequence similarity to other viral proteins is low, with the maximum similarity being 25% to the HKU9 C protein (Table 2). Thus, while the C domain clearly belongs to a conserved fold as indicated by DALI scores of 9.6, 7.0, and 6.5 and TM-Align scores of 0.70, 0.63, 0.64 predicted for HKU9 C, SARS C, and MHV C respectively, it may not necessarily retain conserved functions among CoV phylogenetic groups.

According to Table 2, HKU4 C is similar in tertiary structure to other betacoronaviruses such as HKU9 C, SARS C, and MHV C. Even though SARS C is the lowest in sequence similarity toward HKU4 C, it has the highest structural similarity with HKU4 C. This is indicated by the low RMSD values according to both servers (Table 2). SARS and MHV CoVs belong to lineage B and A, respectively, while HKU4 and HKU9 CoVs belong to lineage C and D, respectively.[3,5,33] According to the structural data generated during our bioinformatics analysis (Table 2), HKU4 C has a similar fold as SARS C, MHV C, and HKU9 C. Therefore, it can be hypothesized that HKU4 C should have the same origin as the C domain present in other betacoronaviruses of the C lineage.

The results of functional annotation using prediction servers indicate that HKU4 C may have the ability to bind peptides and/or small ions such as acetate and calcium. Additionally, it can be hypothesized that the β face and the region opposite to the M domain, the variable face, have possible binding sites for small ligands such as acetate and calcium ions. Based on the *C*-score value predicted (0.7) for the binding reaction between acetate and HKU4 C by COACH, acetate should have the highest probability to bind HKU4 C. COACH is a meta server which is specific in predicting possible ligands and their corresponding binding sites for the given amino acid sequence of a particular protein. In order to generate these predictions it uses other predictions made by other online severs like TM-Site, S Site, COFACTOR, and FINDSITE among others.[26] TM-Site predicts the same *C*-score value for both acetate and calcium (Table 3). Therefore, both acetate and calcium have the same tendency for binding HKU4 C according to the results generated by TMSite.

Another possible function of HKU4 C is binding proteins or peptides. According to the results (Table 3), possible binding sites involved in protein-protein interactions can be found all over the protein, including the β face and α helices situated close to the C-terminus. TM-Site predicts binding of an amino acid sequence (Tyr-Ser-Arg-Ser-Pro-Thr-Ala) in peptides to the α face of HKU4 C and that prediction has a *C*-score value of 0.22 which is the highest *C*-score predicted for binding reactions between HKU4 C and peptides. Therefore, it can be considered as the most reliable prediction for protein-protein binding.

COFACTOR is an online server which can predict biological functions, homologous proteins, and amino acid residues of the possible binding site for a given protein structure. COFACTOR makes all these predictions based on other online servers and databases including TM-Align, Bio Lip Protein Database, Protein Data Bank (PDB), and UniProt.[25,34] Additionally, COFACTOR predicts binding of two other

oligopeptides and both binding reactions have a *C*-score value of 0.1. Therefore, those two protein-protein binding reactions should occur with a less probability than the protein-protein binding reaction predicted by TM-Site. One of the oligopeptides predicted by COFACTOR contains only 5 alanine residues which may bind with the β face of HKU4 C. The other oligopeptide which contains an amino acid sequence of Gln-Arg-Lys-Try-Pro-Leu-Arg-Pro may have the ability to bind with amino acid residues which are situated all over the protein. For this specific peptide, it is hard to predict a specific binding site within HKU4 C.

It is important to note that the Nsl complex was also identified as similar protein for HKU9 C, suggesting a possible surface and functional similarity. It may also be relevant that protein-protein binding has also been identified as a biochemical function of the divergent SUD-M domain of the SARS-CoV. This protein binds to host cell p53 and targets this protein for ubiquitination and degradation.[13] Hence, protein or peptide binding by the neighboring C domain of bat CoVs could have potential biological significance.

A predicted functional site previously identified in the HKU9 C domain[33] was found to not be conserved in HKU4 C. The predicted functional site of HKU9 C is comprised of residues 1360 to 1362 (RDW) and 1381 to 1383 (KRG), with all of the residues except W1362 being solvent exposed. The sequence alignment of Figure 2b shows that the analogous residues of HKU4 C are 1462 to 1464 (HSW) and 1483 to 1485 (HWS), where the conserved W1464 is inaccessible to solvent in both proteins. The aligned residues occupy similar regions of their respective proteins, though the lack of conservation of residues from the HKU9 C-predicted functional site suggests that its function is not retained in HKU4 C. When compared to group 2c CoVs such as the MERS CoV, residues 1464 (W) and 1483 to 1485 (HWS) of HKU4 C are highly conserved (Figure 2b).

The HKU4 C residues that are aligned with the HKU9 C predicted functional site residues experienced significant chemical shift perturbations in the HKU4 MC didomain, indicating that they form protein-protein interactions with HKU4 M (Figures 2 and 4). These residues are well conserved in group 2c CoVs including MERS, implicating protein-protein interactions with the M domain as a conserved function. Additionally, the residues of the HKU9 C-predicted functional site are not well conserved in HKU4 C. The different composition of the HKU4 C conserved face (HSW, WHS in HKU4 as opposed to RDW, KRG in HKU9) and its involvement in protein-protein interactions with HKU4 M both allude to a divergent function of the HKU4 C domain.

Based on the results of chemical shift perturbation of HKU4 C residues in the HKU4 MC didomain, we propose that HKU4 C engages in protein-protein interactions with HKU4 M on its conserved face. The conservation of residues in HKU4 C's conserved face in group 2c CoVs suggests that their purpose is the maintenance of an interface with their respective M domains. The establishment of this interaction surface as well as the lack of conservation of HKU9 C's predicted functional site indicates the possibility of a divergent function for group 2c C domains.

Experimental

Protein Expression and Purification

The residues 1445 to 1522 of nsp3 were amplified from a codon-optimized synthetic gene (Genscript, Piscataway, NJ, USA)[35] and cloned into the vector pET-15b-TE (Northeast Structural Genomics Consortium and DNASu) encoding a N-terminal 6xHis tag. The construct was confirmed by DNA sequencing. The protein was expressed in *E. coli* strain BL21 (DE3) pLysS employing 1 g/L NH_4Cl and 4 g/L ^{13}C-glucose for isotope labeling. Following purification by Ni^{2+} affinity chromatography, fusion tag cleavage with 4.0 mg tobacco etch virus protease was performed overnight at room temperature; in a second Ni^{2+} affinity step and gel filtration, the protein was concentrated into 20 mM sodium phosphate, 150 mM NaCl, and 5 mM d6-DTT using Sartorius Vivaspin 20 ultrafiltration concentrators (Littleton, MA, USA). The protein was monomeric as assessed by size-exclusion chromatography on a HiLoad 26/600 Sephadex 75 pg column (Buckinghamshire, United Kingdom).

The HKU4 M domain and HKU4 MC didomain were expressed using the same vector and *E. coli* strain. These domains spanned residues 1319 to 1445 and 1319 to 1522 of nsp3, respectively. Each protein was purified using analogous steps, with the exception that HKU4 MC used 7.5 mg tobacco etch virus protease for cleavage, and was then concentrated into 20 mM sodium phosphate, 300 mM NaCl, and 3 mM d_6-DTT buffer.

NMR Structure Determination

NMR samples were prepared at 2.0 mM (HKU4 C), 2.0 mM (HKU4 M), or 3.0 mM (HKU4 MC) protein concentration; 3% D_2O (v/v) and 0.02% NaN_3 (w/v) were added. Experiments were carried out on Bruker Avance III HD spectrometers at 600 (Faellanden, Switzerland) and 850 MHz ^{1}H frequencies (Karlsruhe, Germany) equipped with Bruker 5 mm TCI cryoprobes or alternatively on a Bruker Avance II spectrometer at 700 MHz ^{1}H frequency (Faellanden, Switzerland) equipped with a CP TCI H-C/N-D cryoprobe.

Backbone assignments employed triple resonance HNCA, HNCOCA, HNCACB, CBCA(CO)NH, and HNCO experiments analyzed with the program CARA.[36] Side chain assignments employed 3D ^{13}C-resolved NOESY (aliphatic and aromatic) and 3D ^{15}N-resolved NOESY experiments computationally analyzed using the program ASCAN.[37] These assignments were subsequently interactively verified and completed employing the 3D ^{15}N-TOCSY, HBHA(CO)NH, HCCH-TOCSY, aromatic ^{13}C-resolved NOESY,

aliphatic ^{13}C-resolved NOESY, and ^{15}N-resolved NOESY experiments.

^{1}H chemical shifts were calibrated using DSS sample and ^{15}N and ^{13}C shifts were referenced using the Ξ ratio.[38] The CANDID[35] algorithm of the J-UNIO suite[21] was used to pick NOE resonances in the NOESY spectra to calculate the structure of HKU4 C. The first cycle of the calculation resulted in a fold that was consistent throughout the remaining cycles of the calculation. The completion of more cycles of the calculation showed a decrease in target function and RMSD of the ensemble of 20 conformers. The ensemble was then energy-minimized through simulated annealing using CYANA.[39] Validation of the structure was performed using ProCheck 3.5.4,[40] Protein Data Bank validation suite,[41,42] Protein Structure Validation Software,[43] and MolMol v1.0.7.[44] The input and output of UNIO calculations were also evaluated in addition to the agreement of the structure with NMR observables.

The structure for HKU4 C was deposited to the PDB under the accession code 6MWM. The chemical shift assignments for HKU4 C atoms were deposited to the BioMagResBank under accession code 30531.

NMR Interaction Studies

Chemical shift perturbations were measured by comparing the ^{1}H and ^{15}N chemical shifts of the HKU4 C and HKU4 M constructs to those of the HKU 4-MC didomain. ^{1}H chemical shift perturbations were measured from the center of each peak with no scaling, while ^{15}N chemical shift perturbations were measured from the center of each peak and were later multiplied by a scaling coefficient, α, which is equal to the quotient of the range of observed ^{1}H chemical shifts divided by the range of observed ^{15}N chemical shifts.[45] A threshold for chemical shifts that were considered perturbed was established by calculating the standard deviation of all chemical shift perturbations in each protein shifts.[45] Chemical shifts were considered perturbed in the didomain construct if the magnitude of the chemical shift perturbation was greater than the threshold value.

The ^{15}N{^{1}H}-NOE experiments were performed using the Bruker 600 MHz, and the results were processed with the Bruker Dynamics Center.

Functional Annotation

DALI and TM-Align[33] were used to identify homologous proteins which are very close to the primary and tertiary structures of HKU4 C. First, the amino acid sequence was submitted to COFACTOR online server which uses TM-Align to identify homologous proteins for a given protein from the PDB. The structure of HKU4 C was submitted to DALI which compares the 3D structures of proteins in the PDB with the given protein to identify homologous proteins

During the second part of this bioinformatics search, it was aimed to identify possible ligands which can bind to HKU4 C and their possible binding pockets. In this step, online binding site prediction tools like TM-Site, COACH, COFACTOR, and RaptorX[12] were used.

Acknowledgments

We thank the members of Biomolecular NMR Class for technical assistance.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The UAB Comprehensive Cancer Center NMR facility is supported by grants 1P30 CA-13148 and 1S10 RR022994-01A1. This work was supported by the National Institutes of Health grant R35GM119456, University of Alabama at Birmingham Faculty Startup Funding, and the Department of Chemistry, University of Alabama at Birmingham.

Supplemental Material

Supplemental material for this article is available online.

References

1. Fehr AR, Athmer J, Channappanavar R, Phillips JM, Meyerholz DK, Perlman S. The NSP3 macrodomain promotes virulence in mice with coronavirus-induced encephalitis. *J Virol*. 2015;89(3):1523-1536.
2. Gralinski LE, Baric RS. Molecular pathology of emerging coronavirus infections. *J Pathol*. 2015;235(2):185-195.
3. Woo PCY, Wang M, Lau SKP, et al. Comparative analysis of twelve genomes of three novel group 2C and group 2D coronaviruses reveals unique group and subgroup features. *J Virol*. 2007;81(4):1574-1585.
4. Anthony SJ, Ojeda-Flores R, Rico-Chávez O, et al. Coronaviruses in bats from Mexico. *J Gen Virol*. 2013;94(Pt 5):1028-1038.
5. Hu B, Ge X, Wang L-F, Shi Z. Bat origin of human coronaviruses. *Virol J*. 2015;12(1):221.
6. Hurst KR, Koetzner CA, Masters PS. Characterization of a critical interaction between the coronavirus nucleocapsid protein and nonstructural protein 3 of the viral replicase-transcriptase complex. *J Virol*. 2013;87(16):9159-9172.
7. Mu J, Myers RA, Jiang H, et al. Plasmodium falciparum genome-wide scans for positive selection, recombination hot spots and resistance to antimalarial drugs. *Nat Genet*. 2010;42(3):268-271.
8. Fehr AR, Channappanavar R, Jankevicius G, et al. The conserved coronavirus macrodomain promotes virulence and suppresses the innate immune response during severe

acute respiratory syndrome coronavirus infection. *MBio*. 2016;7(6):1-12.
9. Neuman BW, Joseph JS, Saikatendu KS, et al. Proteomics analysis unravels the functional repertoire of coronavirus nonstructural protein 3. *J Virol*. 2008;82(11):5279-5294.
10. Johnson MA, Chatterjee A, Neuman BW, Wüthrich K. SARS coronavirus unique domain: three-domain molecular architecture in solution and RNA binding. *J Mol Biol*. 2010;400(4):724-742.
11. Mielech AM, Deng X, Chen Y, et al. Murine coronavirus ubiquitin-like domain is important for papain-like protease stability and viral pathogenesis. *J Virol*. 2015;89(9):4907-4917.
12. Tan J, Vonrhein C, Smart OS, et al. The SARS-unique domain (SUD) of SARS coronavirus contains two macrodomains that bind G-quadruplexes. *PLoS Pathog*. 2009;5(5):e1000428.
13. Kusov Y, Tan J, Alvarez E, Enjuanes L, Hilgenfeld R. A G-quadruplex-binding macrodomain within the "SARS-unique domain" is essential for the activity of the SARS-coronavirus replication-transcription complex. *Virology*. 2015;484:313-322.
14. Drozdetskiy A, Cole C, Procter J, Barton GJ. JPred4: a protein secondary structure prediction server. *Nucleic Acids Res*. 2015;43(W1):W389-W394.
15. Jaroszewski L, Rychlewski L, Li Z, Li W, Godzik A. FFAS03: a server for profile--profile sequence alignments. *Nucleic Acids Res*. 2005;33(Web Server issue):W284-W288.
16. Griesinger C, Sorensen OW, Ernst RR. Three-dimensional Fourier spectroscopy. Application to high-resolution NMR. *J Magn Reson*. 1969;84(1):14-63.
17. Ulrich EL, Akutsu H, Doreleijers JF, et al. BioMagResBank. *Nucleic Acids Res*. 2008;36(Database issue):D402-D408.
18. Peng J, Xu J. A multiple-template approach to protein threading. *Proteins*. 2011;79(6):1930-1939.
19. Murzin AG, Brenner SE, Hubbard T, Chothia C. Scop: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*. 1995;247(4):536-540.
20. Adinolfi S, Iannuzzi C, Prischi F, et al. Bacterial frataxin CyaY is the gatekeeper of iron-sulfur cluster formation catalyzed by IscS. *Nat Struct Mol Biol*. 2009;16(4):390-396.
21. Serrano P, Pedrini B, Mohanty B, Geralt M, Herrmann T, Wüthrich K. The J-UNIO protocol for automated protein structure determination by NMR in solution. *J Biomol NMR*. 2012;53(4):341-354.
22. Chen VB, Arendall WB, Headd JJ, et al. *MolProbity* : all-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D Biological Crystallography*. 2010;66(1):12-21.
23. Holm L, Kääriäinen S, Rosenström P, Schenkel A. Searching protein structure databases with DaliLite v.3. *Bioinformatics*. 2008;24(23):2780-2781.
24. Zhang Y, Skolnick J. Scoring function for automated assessment of protein structure template quality. *Proteins*. 2004;57(4):702-710.
25. Ashkenazy H, Abadi S, Martz E, et al. ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules. *Nucleic Acids Res*. 2016;44(W1):W344-W350.
26. Landau M, Mayrose I, Rosenberg Y, et al. ConSurf 2005: the projection of evolutionary conservation scores of residues on protein structures. *Nucleic Acids Res*. 2005;33(Web Server issue):W299-W302.
27. Zhang C, Freddolino PL, Zhang Y. Cofactor: improved protein function prediction by combining structure, sequence and protein-protein interaction information. *Nucleic Acids Res*. 2017;45(W1):W291-W299.
28. Yang J, Roy A, Zhang Y. Protein-ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics*. 2013;29(20):2588-2595.
29. Bencze KZ, Kondapalli KC, Cook JD, et al. The structure and function of frataxin. *Crit Rev Biochem Mol Biol*. 2006;41(5):269-291.
30. Tan J, Kusov Y, Mutschall D, et al. The "SARS-unique domain" (SUD) of SARS coronavirus is an oligo(G)-binding protein. *Biochem Biophys Res Commun*. 2007;364(4):877-882.
31. Ma-Lauer Y, Carbajo-Lozoya J, Hein MY, et al. P53 down-regulates SARS coronavirus replication and is targeted by the SARS-unique domain and PLpro via E3 ubiquitin ligase RCHY1. *Proc Natl Acad Sci U S A*. 2016;113(35):E5192-E5201.
32. Hammond RG, Tan X, Johnson MA. SARS-unique fold in the *Rousettus* bat coronavirus HKU9. *Protein Sci*. 2017;26(9):1726-1737.
33. Mielech AM, Chen Y, Mesecar AD, Baker SC. Nidovirus papain-like proteases: multifunctional enzymes with protease, deubiquitinating and deISGylating activities. *Virus Res*. 2014;194:184-190.
34. Herrmann T, Güntert P, Wüthrich K. Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. *J Mol Biol*. 2002;319(1):209-227.
35. Keller R. Optimizing the process of nuclear magnetic resonance spectrum analysis and computer-aided resonance assignment. A dissertation submitted to the Swiss federal Institute of technology Zurich (ETH Zürich) for the degree of doctor of natural sciences. *Diss. ETH Nr*;*15947*.
36. Fiorito F, Herrmann T, Damberger FF, Wüthrich K. Automated amino acid side-chain NMR assignment of proteins using (13) C- and (15)N-resolved 3D [(1)H, (1)H]-NOESY. *J Biomol NMR*. 2008;42(1):23-33.
37. Markley JL, Bax A, Arata Y, et al. Recommendations for the presentation of NMR structures of proteins and nucleic acids. IUPAC-IUBMB-IUPAB Inter-Union task group on the standardization of data bases of protein and nucleic acid structures determined by NMR spectroscopy. *J Biomol NMR*. 1998;12(1):1-23.
38. Güntert P, Buchner L. Combined automated NOE assignment and structure calculation with CYANA. *Journal of Biomolecular NMR*. 2015;62(4):453-471.

39. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK: a program to check the stereochemical quality of protein structures. *Journal of Applied Crystallography*. 1993;26(2):283-291.
40. Berman H, Henrick K, Nakamura H. Announcing the worldwide protein data bank. *Nat Struct Mol Biol*. 2003;10(12):980.
41. Berman H, Henrick K, Nakamura H, Markley JL. The worldwide protein data Bank (wwPDB): ensuring a single, uniform archive of PDB data. *Nucleic Acids Res*. 2007;35(Database issue):D301-D303.
42. Bhattacharya A, Tejero R, Montelione GT. Evaluating protein structures determined by structural genomics consortia. *Proteins*. 2007;66(4):778-795.
43. Koradi R, Billeter M, Wuthrich K. MOLMOL: a program for display and analysis of macromolecular structures. *Journal of Molecular Graphics and Modelling*. 1996;14:29-32.
44. Williamson MP. Using chemical shift perturbation to characterise ligand binding. *Prog Nucl Magn Reson Spectrosc*. 2013;73:1-16.
45. Källberg M, Wang H, Wang S, et al. Template-based protein structure modeling using the RaptorX web server. *Nat Protoc*. 2012;7(8):1511-1522.