

## X-Ray Structures of the N- and C-Terminal Domains of a Coronavirus Nucleocapsid Protein: Implications for Nucleocapsid Formation

Hariharan Jayaram,<sup>1†</sup> Hui Fan,<sup>2</sup> Brian R. Bowman,<sup>1‡</sup> Amy Ooi,<sup>2</sup> Jyothi Jayaram,<sup>3</sup> Ellen W. Collisson,<sup>3</sup> Julien Lescar,<sup>2\*</sup> and B. V. Venkataram Prasad<sup>1\*</sup>

*Verna and Marrs McLean Department of Biochemistry and Molecular Biology, Baylor College of Medicine, Houston, Texas 77030<sup>1</sup>; School of Biological Sciences, Nanyang Technological University, Singapore 637551<sup>2</sup>; and Department of Veterinary Pathobiology, Texas A&M University, College Station, Texas 77843<sup>3</sup>*

Received 23 January 2006/Accepted 10 April 2006

**Coronaviruses cause a variety of respiratory and enteric diseases in animals and humans including severe acute respiratory syndrome. In these enveloped viruses, the filamentous nucleocapsid is formed by the association of nucleocapsid (N) protein with single-stranded viral RNA. The N protein is a highly immunogenic phosphoprotein also implicated in viral genome replication and in modulating cell signaling pathways. We describe the structure of the two proteolytically resistant domains of the N protein from infectious bronchitis virus (IBV), a prototype coronavirus. These domains are located at its N- and C-terminal ends (NTD and CTD, respectively). The NTD of the IBV Gray strain at 1.3-Å resolution exhibits a U-shaped structure, with two arms rich in basic residues, providing a module for specific interaction with RNA. The CTD forms a tightly intertwined dimer with an intermolecular four-stranded central  $\beta$ -sheet platform flanked by  $\alpha$  helices, indicating that the basic building block for coronavirus nucleocapsid formation is a dimeric assembly of N protein. The variety of quaternary arrangements of the NTD and CTD revealed by the analysis of the different crystal forms delineates possible interfaces that could be used for the formation of a flexible filamentous ribonucleocapsid. The striking similarity between the dimeric structure of CTD and the nucleocapsid-forming domain of a distantly related arterivirus indicates a conserved mechanism of nucleocapsid formation for these two viral families.**

*Coronaviridae* is a family of viruses which are the causative agents of human upper respiratory infections including common colds, as well as severe illnesses such as severe acute respiratory syndrome (SARS). Avian infectious bronchitis virus (IBV) is a major source of mortality in chickens worldwide and has a significant impact on the poultry industry. Other coronaviruses affect domestic animals and are of veterinary significance. Coronaviruses are enveloped viruses with a diameter ranging from 80 to 160 nm. Their viral genome consists of positive-sense single-stranded (ss) RNA of approximately 30 kb (36). The genomic RNA encodes a 3' coterminal set of four or more subgenomic mRNAs with a common leader sequence at their 5' ends. These subgenomic RNA segments encode various structural and nonstructural viral proteins that are required to produce progeny virions. The viral particle consists of a nucleocapsid or core structure surrounded by a lipid envelope in which the membrane glycoprotein (M) and another small transmembrane protein (E) are embedded. A series of protrusions composed of glycoproteins (S) anchored in the

lipid envelope extend radially, forming up to 20-nm-long spikes which give the roughly spherical viral particles a crown (corona) appearance.

During the virus life cycle, multiple copies of the nucleocapsid phosphoprotein (N) interact intimately with genomic and subgenomic RNA molecules (1, 28) and together with M, the most abundant envelope protein, participate in genome condensation and packaging. The N and M proteins interact via their C termini, leading to specific genome encapsidation in the budding viral particle (19). Electron microscopic studies of detergent-permeabilized transmissible gastroenteritis virus capsids revealed that the internal ribonucleocapsid is a flexible filamentous structure with a diameter of approximately ~10 to 15 nm and up to several hundred nanometers in length (33, 34). The highly basic N protein has a molecular mass ranging between 45 and 60 kDa in the various groups of coronaviruses and, along with its coding RNA, is synthesized in large amounts during infection (20, 37). The N protein is able to bind ssRNA nonspecifically but displays an increased affinity for viral genomic RNA (9). Packaging signals have been identified at the 5' and 3' termini of the genome for several coronaviruses, but not unambiguously for the IBV genome. Biochemical studies of murine hepatitis virus (MHV), IBV, and SARS coronavirus (SARS-CoV) have mapped the RNA binding function to a segment of 55 residues located at the N-terminal half of the N protein and the dimerization function to its C-terminal half (14, 29, 45). In addition to its structural role, the N protein is also implicated in other processes during infection including mRNA transcription, replication, and host cell modulation (16, 20, 25, 35, 39, 41, 44). The N protein is

\* Corresponding author. Mailing address for B. V. V. Prasad: Department of Biochemistry and Molecular Biology, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030-3498. Phone: (713) 798-5686. Fax: (713) 798-1625. E-mail: vprasad@bcm.tmc.edu. Mailing address for J. Lescar: School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore 637551. Phone: 65-6316-2859. Fax: 65-6791-3856. E-mail: julien@ntu.edu.sg.

† Present address: Howard Hughes Medical Institute and the Department of Biochemistry, Brandeis University, Waltham, MA 02454.

‡ Present address: Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138.

also an important diagnostic marker for coronavirus disease and a major immunogen that can prime protective immune responses (23).

The recombinant N protein of coronavirus expressed in *Escherichia coli* is highly susceptible to proteolysis, making structural analysis of the full-length protein difficult. To date, there is only limited information on the structure of the ribonucleocapsid protein, including a nuclear magnetic resonance (NMR) analysis of the N-terminal domain of the SARS-CoV N protein (18) and our crystallographic study of the N-terminal domain of the IBV N protein (Beaudette strain) at 1.85-Å resolution (14). As yet, there is no X-ray crystallographic structure analysis of the C-terminal dimerization domain of a coronavirus N protein. In this paper, we present a structural analysis, at high resolution, of the two proteolytically stable domains of the IBV N proteins located at the N- and C-terminal ends, called NTD and CTD, respectively. The NTD structure of IBV (Gray strain), determined to a 1.3-Å resolution, is similar to the previously published Beaudette strain (12) but makes strikingly different quaternary associations. We describe the first crystal structure of the CTD dimerization domain of the coronavirus N protein at a resolution of 2.2 Å. Our X-ray crystallographic analysis of the NTD and CTD provides insight into the way these modules might interact with RNA and with the M protein. The various crystal forms also delineate a number of alternative protein surfaces that are likely to be used for the formation of a flexible filamentous ribonucleocapsid.

#### MATERIALS AND METHODS

**Purification of full-length nucleocapsid protein and limited proteolysis.** The full-length N protein was expressed as described previously (47). The protein was further purified by heparin affinity chromatography, concentrated to 1 to 2 mg/ml and checked for homogeneity using dynamic light scattering (Dynapro) and negative-stain electron microscopy. Limited proteolytic cleavage of the full-length N protein (1 to 2 mg/ml) was carried out with 2% (weight trypsin/weight protein) sequencing-grade trypsin (Roche) to identify stable domains. The identity of the amino termini of the proteolytic product(s) was determined by N-terminal amino acid sequencing of the band following gel electrophoresis and blotting onto a polyvinylidene fluoride membrane (PVDF-Immobilon-P<sup>SO</sup>; Millipore). For construct optimization the identities of the carboxy-terminal amino acids were estimated based on secondary structure prediction and mass spectrometric characterization of the proteolyzed protein.

**Cloning, expression, purification, and crystallization of the tryptic fragments of N protein.** The NTD and CTD proteins from two strains were employed in this study, namely, IBV Gray (CTD1, CTD2, and NTD1) and IBV Beaudette (CTD3) (Fig. 3). The proteins were cloned and expressed as glutathione S-transferase fusion proteins using the pet41 Ek-LIC vector (Novagen) or, for the Beaudette strain, as detailed previously (14). The expressed protein was purified using a glutathione S-Sepharose (Pharmacia) affinity column, followed by on-bead cleavage with enterokinase (EK-Max, Invitrogen). The cleavage reaction was performed by suspending 1 ml of beads in 40 ml of cleavage buffer (250 mM NaCl, 50 mM Tris-HCl [pH 8.0]) with 10 units of protease. Following proteolysis, the diluted supernatant was further purified by gel filtration chromatography on a Superdex 75 16/60 column (Pharmacia). The purified N- and C-terminal domains were concentrated to 5 to 8 mg/ml for crystallization.

**Data collection and phasing.** Diffraction data were collected at various synchrotron beam lines as indicated in Table 1. For each crystal, images were collected using an oscillation angle of 1° and integrated and scaled with HKL2000 (31). For the NTD, the diffraction data to 1.3 Å were phased using molecular replacement (MR) procedures in PHASER (38), using the previously published NTD structure (PDB ID, 2BTL) at 2.8-Å resolution (14). Following MR, further model building and refinement were performed in a similar manner to that for the CTD as described below. The CTD crystallized in three crystal forms (Table 1). Its structure was determined using selenomethionine (Se-Met)-substituted protein with crystal form CTD1 (Table 1) using multiwavelength

TABLE 1. Crystallization and data collection

Data set	CTD1	CTD2	CTD3	NTD1
Crystallization condition	PEG 4000, 28.75% to 29.5%, pH 4.8 citrate, 0.1 M MgCl <sub>2</sub>	30% PEG 4000, 100 mM Tris-HCl, pH 8.6, 800 mM LiCl	4.3M NaCl, 0.1 M Tris-Cl, pH 8.5	25% PEG 4000, 100 mM MES sodium salt, pH 6.2, 200 mM MgCl <sub>2</sub>
X-ray source	SBC-CAT 19ID advanced photon source (Argonne)	BIOCARS-14BM-C advanced photon source (Argonne)	ID14-4, ESRF (Grenoble)	BIOCARS-14BM-C advanced photon source (Argonne)
Wavelength	0.97937 Å (360°, 1° oscillation)	0.9000 Å (180°, 1° oscillation)	0.97626 Å (70°, 1° oscillation)	0.9000 Å (180°, 1° oscillation)
Cell parameters (Å)	P2(1)2(1)2(1), <i>a</i> = 38.39, <i>b</i> = 65.94, <i>c</i> = 92.31, $\alpha = \beta = \gamma = 90$	P2(1)2(1)2, <i>a</i> = 108.99, <i>b</i> = 128.53, <i>c</i> = 71.44, $\alpha = \beta = \gamma = 90$	P4(3), <i>a</i> = 61.59, <i>b</i> = 61.59, <i>c</i> = 91.88, $\alpha = \beta = \gamma = 90$	C2, <i>a</i> = 100.06, <i>b</i> = 46.21, <i>c</i> = 74.18, $\alpha = 90$ , $\beta = 121.06$ , $\gamma = 90$
Resolution (Å)	50–2.0	50–2.2	20–2.6	50–1.3
Total no. of reflections	15,139	97,377	31,078	204,381
Number of molecules in ASU (solvent fraction)	2 (33%)	8 (41.8%)	2 (58.4%)	2 (66%)
Completeness (%)	96 (68.9)	99.7 (98.8)	90.6 (90.5)	87.9 (60.5)
Redundancy <sup>a</sup>	11 (8.2)	3.4 (3.1)	1.65 (1.66)	3.3 (2.6)
R <sub>merge</sub> <sup>b</sup>	0.072 (0.282)	0.081 (0.588)	0.079 (0.417)	0.056 (0.227)
I/ $\sigma$ (I)	28.80 (0.73)	16.23 (2.83)	13.59 (3.1)	49.3 (13.3)

<sup>a</sup> The numbers in parentheses refer to the last (highest) resolution shell.

<sup>b</sup>  $R_{\text{merge}} = \sum_i \sum_h |I_{ih} - \langle I_h \rangle| / \sum_i \sum_h I_{ih}$ , where  $I_{ih}$  is the *i*th observation of the reflection *h*, while  $\langle I_h \rangle$  is its mean intensity.

TABLE 2. Refinement statistics

Domain (PDB code)	PDB CTD1 (2GE7)	PDB CTD2 (2GE8)	CTD3 (2CA1)	PDB NTD1 (2GEC)
Resolution range (Å)	50–2.0	48–2.2	20–2.6	50–1.3
Number of reflections	15,110	48,564	8,941	60,751
Rfactor <sup>a</sup>	0.238	0.236	0.204	0.210
Rfree <sup>b</sup>	0.269	0.291	0.256	0.250
Mean bond length deviation	0.005	0.006	0.009	0.008
Mean bond angle deviation	1.305	1.318	1.176	1.235
Ramachandran statistics				
Residues in most-favored regions (%)	94.2	91.6	89.8	88.6
Residues in additional allowed regions (%)	5.8	7.6	10.2	10.4
Residues in generously allowed regions (%)		0.6		0.5
Residues in disallowed (poor density) regions (%)		0.2		0.5

<sup>a</sup> Rfactor =  $\sum \|F_{\text{obs}} - F_{\text{calc}}\| / \sum F_{\text{obs}}$ .

<sup>b</sup> Rfree was calculated with 10% of reflections excluded from the whole refinement procedure.

anomalous dispersion data sets collected at two different wavelengths (Se peak, 0.9734 Å; Se inflection, 0.9748 Å). Positions of the four Se atoms were located using the SnB program (43) and refined using SHARP (overall figure of merit of 0.65) (2). An electron density map was calculated following density modification using CCP4 (8). An initial model was built using ARP/WARP (21) followed by manual model building using Coot (13). A few cycles of simulated annealing were performed with CNS (3) followed by model refinement using REFMAC5 (32). The structures of CTDs in the two other crystal forms (CTD2 and CTD3) were solved using MR procedures as implemented in PHASER (23). Model bias in both NTD and CTD structures was reduced by using the prime-and-switch technique implemented in SOLVE/RESOLVE (42). During the course of model building and refinement, the stereochemistry of the structures was checked by PROCHECK (22). The final statistics are provided in Table 2. Surface electrostatic potentials were calculated using DELPHI (30). All figures were generated using PyMOL (10) and ESPript (15).

**Protein structure accession numbers.** The coordinates for the molecules were deposited into the PDB with accession numbers 2GE7 (CTD1), 2GE8 (CTD2), 2CA1 (CTD3), and 2GEC (NTD1).

## RESULTS

**Identification of two stable independent domains of IBV N by limited proteolysis.** Since the full-length recombinant protein aggregated and was degraded under a variety of experimental conditions, we sought to identify stable domains that were resistant to mild proteolysis. We used limiting amounts of trypsin and V8 protease. The digestion pattern with the V8 protease was not very distinct, yielding several diffuse bands.

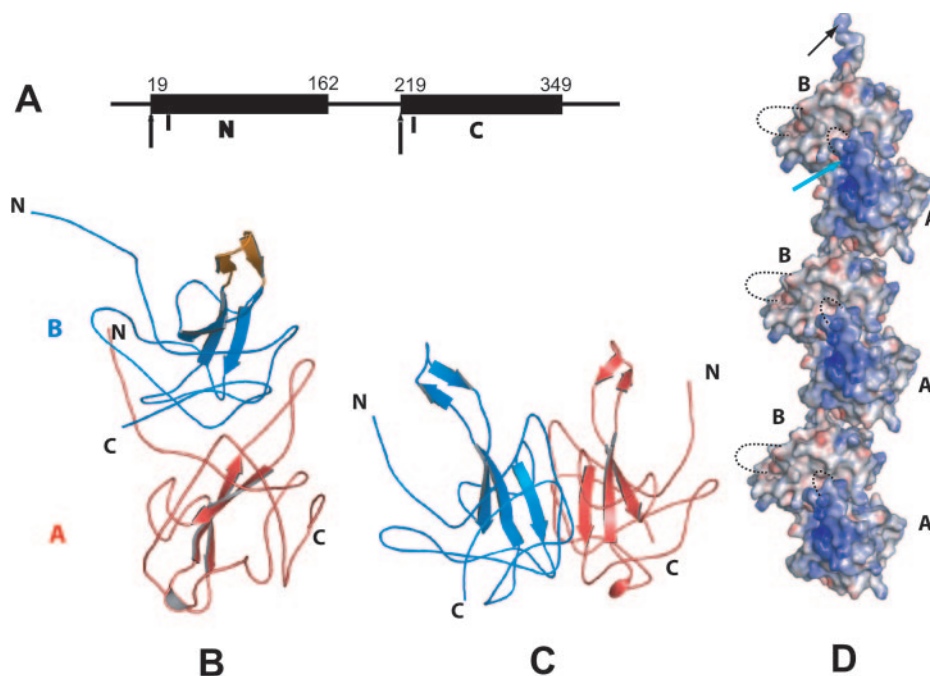


FIG. 1. Structural domains of the IBV N protein and structure of the NTD RNA binding domain. (A) Schematic diagram showing the major (arrow) and minor trypsinization sites (short vertical line) following limited proteolysis of the full-length IBV N protein. The locations of the N- and C-terminal domains (NTD spanning residues 19 to 162 and CTD spanning residues 219 to 349) are depicted as black rectangles. (B) Ribbon representation of the 1.3-Å structure of the NTD (Gray strain) asymmetric homodimer (molecules A and B as indicated). The region corresponding to the disordered internal arm is depicted in orange. (C) The NTD (Beaudette strain) determined by Fan et al. (14). (D) Electrostatic potential surface of the linear array of NTD dimers generated by the crystallographic translation. Molecules A and B that constitute the dimer are indicated. The N-terminal arm and the region corresponding to the internal arm, rich in basic residues, are indicated by black and cyan arrows, respectively. The disordered loop in the B molecule is indicated by a dotted line (see the text).

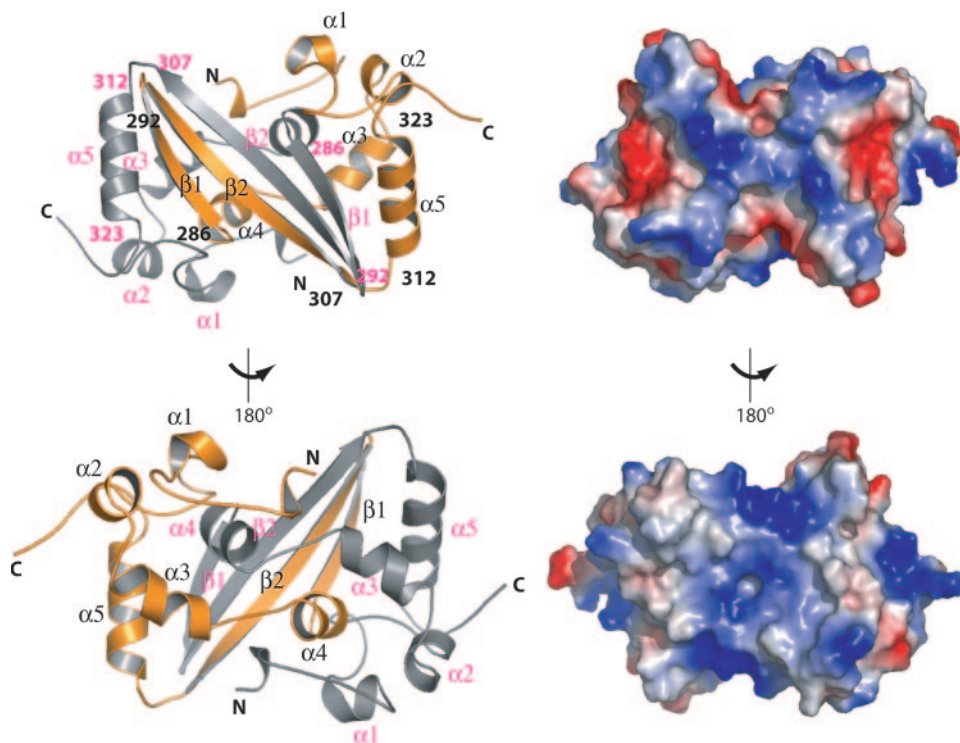


FIG. 2. Structure of the CTD dimerization domain. The left panel shows a ribbon representation of the “front” and “back” of the CTD dimer related by a rotation of  $180^\circ$  about the vertical axis, and the right panel shows the electrostatic potential surface of the dimer in the same orientations, in which the positively charged surface is represented in shades of blue and the negatively charged surface in shades of red. Left, the intertwined CTD dimer is formed by exchanging two  $\beta$  strands and one  $\alpha$  helix between the two monomers (“domain swapping”). The two monomers, in yellow and gray, respectively, are related by a noncrystallographic twofold axis of symmetry approximately perpendicular to the plane of the figure. The  $\beta$  strands from both monomers form an extended antiparallel  $\beta$ -sheet flanked by several  $\alpha$  helices. Secondary structural elements are labeled. Right, a large patch of positively charged residues (blue) that could be involved in RNA binding is visible on one of the faces of the CTD protein (bottom).

The full-length protein, however, could be cleaved into a “single” stable  $\sim 17$ -kDa band within 15 min of trypsinization. N-terminal sequencing allowed the identification of four tryptic fragments with two major cleavage sites at residues 19 and 219 and two secondary cleavage sites at residues 27 and 226 (Fig. 1A).

New constructs corresponding to these proteolytically resistant domains termed NTD (residues 19 to 162) and CTD (residues 219 to 349) were cloned, expressed, and purified. The NTD was monomeric at moderate protein concentrations, whereas the CTD was a dimer even at very low concentrations, as assayed by gel filtration chromatography (12). The NTD and CTD proteins tended to aggregate during the purification procedure and thus purified at very low concentrations and concentrated only prior to crystallization.

**NTD and CTD crystallized in multiple crystal forms.** The recently reported structure of NTD at  $1.85$ - $\text{\AA}$  resolution corresponds to the Beaudette strain of IBV. We used the IBV Gray strain which crystallized in a different crystal form and diffracts to  $1.3$ - $\text{\AA}$  resolution (Table 1, crystal form NTD1). The CTD yielded crystals with needle, rod, flat sheet, or hexagonal shapes under various conditions (Table 1, crystal forms CTD1, CTD2, and CTD3). Rod-shaped CTD1 crystals of Se-Met substituted protein diffracting to  $2.0$   $\text{\AA}$  were used for structure determination. The structures of CTD in the other crystal forms (CTD2 at  $2.2$ - $\text{\AA}$  and CTD3 at  $2.6$ - $\text{\AA}$  resolution) were

determined subsequently by molecular replacement. These various crystal forms exhibit different packing arrangements, a number of which could mimic the intermolecular interactions that trigger the formation of the coronavirus nucleocapsid.

**High-resolution structure of NTD.** With the exception of five additional residues discernible at its N terminus, the present structure of NTD of IBV Gray strain is quite similar to the structure of the NTD of IBV Beaudette strain (Fig. 1C) (14). Briefly, the NTD monomer features a relatively acidic globular core of twisted antiparallel  $\beta$ -sheet surrounded by several loop regions. Prominent among the loop regions are two long segments corresponding to the N-terminal 12 amino acids (residues 22 to 34) and an internal arm spanning residues 74 to 86. These loops protrude from the globular core resulting in a “U”-shaped monomer (Fig. 1B).

A dimer of NTD subunits is present in the crystallographic asymmetric unit: two interlocking NTD monomers are arranged in a head-to-tail fashion with the basic arms of one monomer interacting with the acidic base of the other monomer (Fig. 1B). The main structural variation between the two NTD monomers related by noncrystallographic symmetry is that in one monomer, one of the arms of the “U” (internal arm) is disordered. The buried surface area of  $\sim 2,150$   $\text{\AA}^2$  between the two NTD monomers indicates a rather strong interaction. This is in contrast with the previous structure of

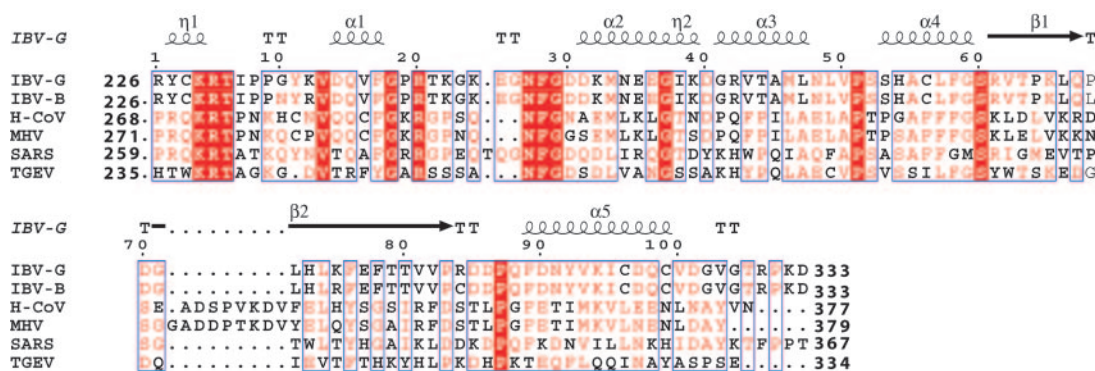


FIG. 3. Structure-based alignment of coronavirus nucleocapsid amino acid sequences corresponding to the CTD dimerization domain. Secondary structure elements are labeled above the sequence for the IBV CTD dimerization domain (this work). Sequences for IBV N proteins were obtained from Swiss-Prot (IBV-G [Gray strain], P32923; IBV-B [Beaudette strain], P69596). Sequences for human coronavirus (H-CoV; strain HKU1, YP\_173242); MHV (strain 1, AAA46439); SARS (SARS-CoV, NCAP\_CVHSA) and porcine transmissible gastroenteritis virus (TGEV; strain RM4, AAG30228) were obtained from GenBank. Helices ( $3_{10}$  and  $\alpha$  helices) are shown as squiggles and  $\beta$  strands as arrows. Boxes indicate residues that are fully or partially conserved. Fully conserved residues are shaded in red. Partially conserved residues are indicated by salmon-pink letters.

NTD (14), where the U-shaped monomers were kept parallel but which had a much smaller buried surface area of only  $\sim 590 \text{ \AA}^2$  (Fig. 1C).

**The structure of CTD is a tightly intertwined dimer.** In all three crystal forms obtained, the CTD exists as a tightly intertwined twofold symmetric dimer (Fig. 2) with two  $\beta$  strands and one  $\alpha$  helix from one monomer making extensive contacts with the other monomer and burying a total surface of approximately  $5,000 \text{ \AA}^2$  in their interaction. The distribution of secondary structure elements along the CTD amino acid sequence and the topology of the CTD dimer are shown in Fig. 3. The CTD dimer has a rectangular shape delimited by edges formed by the C-terminal  $\alpha$  helices  $\alpha 5$  (Fig. 2) with approximate dimensions of  $40 \text{ \AA} \times 40 \text{ \AA} \times 20 \text{ \AA}$ . It features a concave floor of  $\sim 400\text{-}\text{\AA}^2$  area consisting of an antiparallel  $\beta$  sheet ( $\beta 1A$ - $\beta 2A$ - $\beta 2B$ - $\beta 1B$ ) contributed by monomers A and B, respectively, surrounded by several  $\alpha$  helices and one short  $3_{10}$  helix (Fig. 2). Helices  $\alpha 3$  and  $\alpha 4$  are connected by a loop and, together with their dimeric partners, form a groove which arches inward over this floor constituting the other mostly basic face of the CTD dimer. Recent biochemical and mass spectrometric studies on the IBV N protein (Beaudette strain) have suggested the possibility of disulfide bridges in the CTD (5). However, no intramolecular disulfide bridge is seen in the CTD dimer of either strain of IBV reported here. The present structure of the CTD is consistent with previous observations that the CTD is a dimer in solution, with several biochemical studies which map the dimerization domain of the full-length protein to its C-terminal domain (40, 45).

**Multiple packing modes of CTD dimers.** The CTD dimer is involved in various intermolecular interactions in the three crystal forms reported here. The presence of one dimer (CTD1 and CTD3) and four dimers in the asymmetric unit of the CTD2 crystal form (Tables 1 and 2) permits an analysis of dimer-dimer interactions in various conditions of precipitant and pH. We focused on interactions with a buried surface area larger than  $1,000 \text{ \AA}^2$ . Such an analysis could help in the identification of molecular surfaces that are used to assemble the nucleocapsid.

CTD1 dimers (pH, 4.5) which are related by the crystallographic  $2_1$  screw axis display three kinds of intermolecular contacts. One interaction (burying  $\sim 1,100 \text{ \AA}^2$ ) brings two dimers in a tail-to-tail fashion (referred to as type S) (Fig. 4A). Interestingly, in the CTD2 form which crystallized at pH 8.5, a similar contact is observed between three of the four crystallographically independent dimers (Fig. 4B). However, unlike in CTD1, where the dimers form an infinitely long linear array, a small swivel between the three CTD2 dimers (dimers 1, 2, and 3) introduces a slight curvature. In both crystal forms, type S interactions are mediated by C-terminal residues located between positions 308 and 328, which include  $\alpha$  helix  $\alpha 5$  and a type II turn. A network of water-mediated polar interactions and a salt bridge between residues Arg 308 and Asp 314 are observed (Fig. 4A, bottom). Secondly, a lateral interaction between dimers 2 and 4 (Fig. 4B) mediated by their N-terminal residues (221 to 230) buries a comparable surface area of  $\sim 1,250 \text{ \AA}^2$  (type L). Dimer 4 extends the helical array formed by dimers 1, 2, and 3 in a lateral manner (Fig. 4B [type S']). Finally, in the CTD3 crystal form, dimers form a long helical polymer that spirals along the  $4_3$  screw axis, with a buried surface area of  $\sim 1,085 \text{ \AA}^2$  (type F). These interactions mediated by hydrogen bonds between Arg 230 of one monomer and carbonyl atoms from residues 263 to 266 in the other monomer bear some resemblance to the type L interactions seen in CTD1 and CTD2. These contacts result in an infinitely propagating tube of CTD3 molecules with a diameter of approximately  $60 \text{ \AA}$  (Fig. 4C and 5B). Interestingly, this arrangement of CTD subunits would place the RNA binding N-terminal domains towards the interior of the tube and the C-terminal domains pointing outside.

## DISCUSSION

A flexible filamentous nucleocapsid formed by the close association of N proteins with viral genomic RNA is a common feature in many enveloped ssRNA viruses including coronaviruses. Structural information on the N protein and a molecular

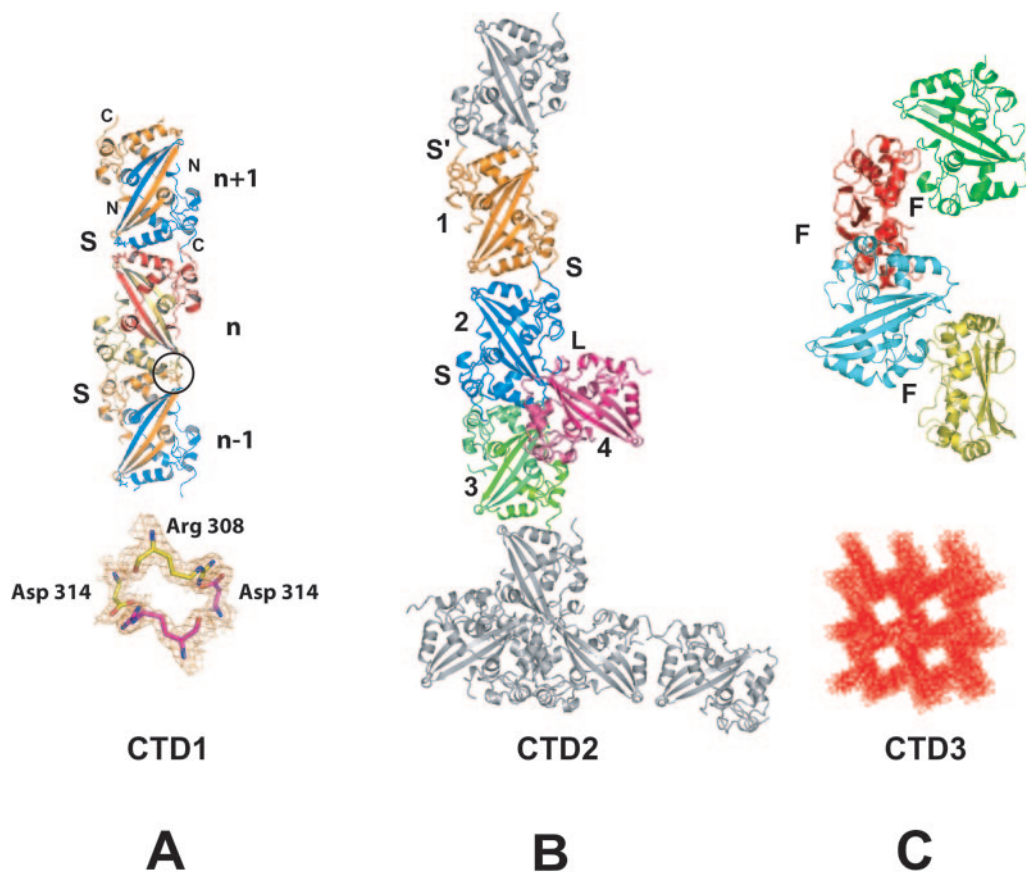


FIG. 4. Crystal-packing interactions between CTD dimers. (A) Crystal-packing interactions in CTD1 crystals grown at pH 4.5, with one dimer in the asymmetric unit (ASU). Three consecutive dimers from the neighboring ASU (numbered  $n$ ,  $n + 1$ , and  $n - 1$ ) related by one of the three orthogonal  $2_1$  screw axes are shown (type-S interaction). Each monomer has been given a different color. The N- and C-terminal ends for the  $n + 1$  are indicated. The salt bridge interaction between dimers  $n$  and  $n - 1$  seen in the type-S interface is circled. A closeup view of the salt bridge interaction with an electron density map is shown in the inset below. (B) Interactions between the four CTD dimers present in the CTD2 ASU (pH 8.5). Each dimer is shown in a different color and numbered from 1 to 4. The two classes of dimer-dimer interactions are indicated by S (between molecules 1 and 2 and molecules 2 and 3) and L (between molecules 2 and 4). The bridging type-S' interactions are shown with molecule 4 from two adjacent ASUs and molecules 1, 2, and 3 from the neighboring ASU (all shown in gray). (C) Dimer-dimer interactions observed in the CTD3 crystals having one CTD dimer in the ASU (type F). Each dimer related by the crystallographic  $4_3$  screw axis is shown in a different color. This type of packing gives rise to hollow tubes, as is evident from a projection along the fiber axis (along the c axis) below shown in red.

understanding of how this protein facilitates the formation of the nucleocapsid are limited. From the biochemical characterization of the N protein of IBV, a prototypical coronavirus, presented here, it is apparent that this protein has two major protease-resistant domains. Our X-ray crystallographic analysis of these two domains, NTD and CTD, provides some insights into how the two-domain organization of the N protein may coordinate nucleocapsid assembly.

**Interaction of N with RNA.** Biochemical studies (14, 29) have located the RNA binding site in the N-terminal domain with the minimal region being mapped to residues 177 to 231 in MHV (corresponding to residues 136 to 190 in IBV). In addition to the NTD, an involvement of the CTD in RNA binding has been shown by Fan et al. (14). Based on the structure of the NTD (IBV Beaudette strain), we proposed that the basic arms of the U-shaped monomer participate in RNA binding. This hypothesis is consistent with an NMR-heteronuclear single quantum coherence analysis of NTD-RNA interactions that was carried out for the SARS coro-

navirus N protein (18). A novel finding from our crystal structure analysis is the presence of strong interlocking dimers of the NTD (IBV Gray strain). This is in contrast with the weak dimeric interaction observed in the NTD of the IBV Beaudette strain reported earlier (14). These interlocking dimers associate to form a linear fiber with the basic tethers exposed along the surface. Such fibers could provide for closely packed interactions of NTD with the viral genomic RNA. Analysis of N protein-RNA interactions in MHV at different stages of the virus life cycle revealed that these interactions progress from an RNase-sensitive complex involving subgenomic RNA to an RNase-resistant complex involving genomic RNA (28). The strong and weak NTD dimer interactions seen in the two structures could correspond to these different states of N protein-RNA associations. The electrostatic potential surface of the CTD dimer shows one of its faces significantly more basic than the other, with a groove lined by  $\alpha$ -helices  $\alpha 3$  and  $\alpha 4$  that could interact with RNA (Fig. 2 and 5A and B).

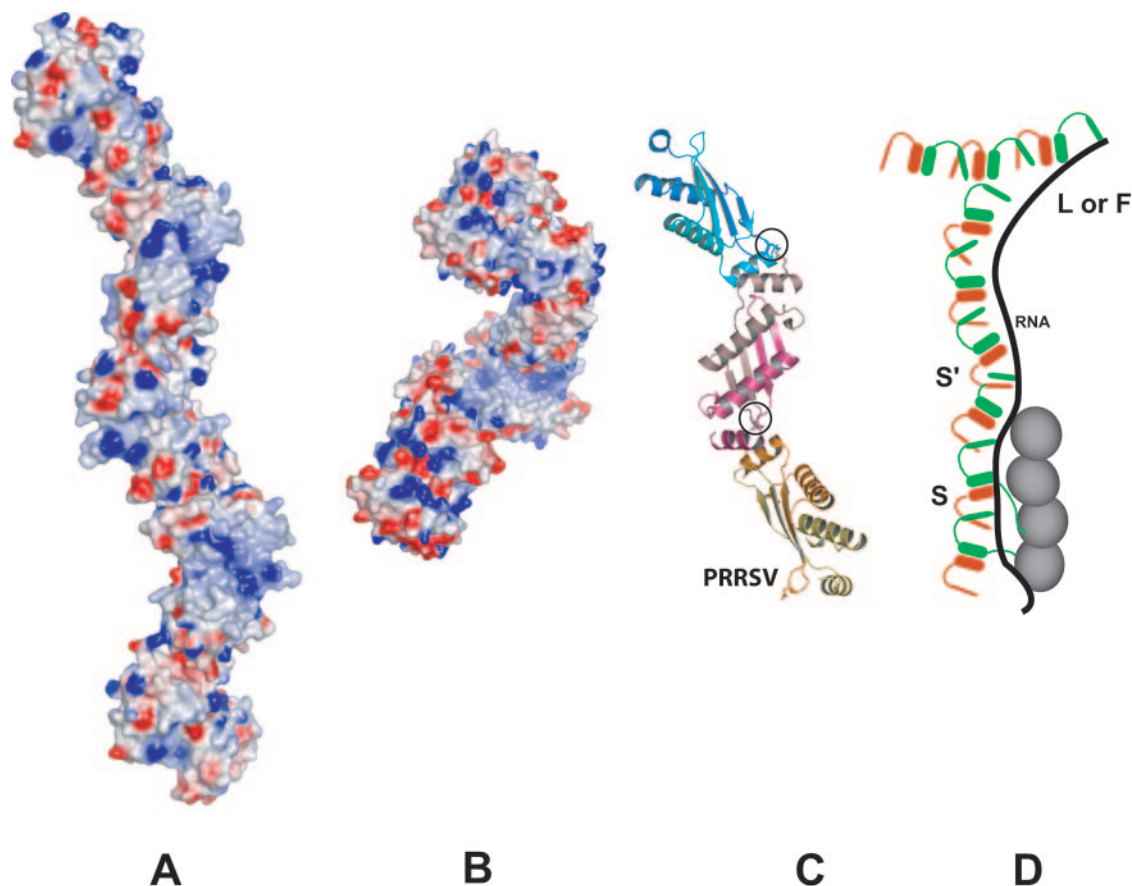


FIG. 5. A possible model for helical nucleocapsid formation. (A) Electrostatic potential surface of the CTD polymer formed through S' interactions in the CTD2 crystals. (B) View of the electrostatic potential surface of the assembly formed by the close association of five CTD dimers related by the  $4_3$  screw axis, from neighboring-unit cells, as observed in the CTD3 crystals. The grooves lined by basic  $\alpha$  helices are well exposed to the solvent and could thus participate in viral genomic RNA binding. (C) A possible model for the nucleocapsid formation based on the protein-protein interfaces observed in the crystallographic structures of the NTD and CTD domains of IBV. The NTD dimers (gray spheres) provide specific binding to the viral genomic RNA (black line). Secondary contacts with RNA with polymers of CTD are formed so that together the viral genome is condensed in a protective flexible tube. The CTD dimers (shown in red and green) could interact via type-S interface. Changes in the curvature and/or the direction of the nucleocapsid filament would derive from the incorporation of another type of isoenergetic interfaces (e.g., type-S', type-L, or type-F interactions). (D) This arrangement is reminiscent of the assemblies formed by the PRRSV capsid-forming domain (PDB ID, 1P65) (11).

**Interaction between the N and M proteins.** In addition to their interactions with RNA, N proteins also interact with the M proteins embedded in the viral membrane. Based on reverse genetic-complementation assays, the interaction region between these two proteins has been mapped to their C termini (19). The C terminus of the M protein is significantly basic, and recent mutational studies on the M protein have demonstrated that its interaction with the N protein is predominantly electrostatic in nature (26). The exposed acidic  $\beta$ -sheet floor, on the opposite side of the proposed RNA-binding region in the CTD dimer, may promote such an interaction. Thus, the CTD may serve a dual purpose of mediating the self-association of the N protein during nucleocapsid formation but also of providing a complementary surface for interaction with the endodomain of the M protein in the virus envelope.

**Possible model for nucleocapsid formation.** The formation of the coronavirus nucleocapsid involves self-association of the N protein and interaction with RNA resulting in a structure which is RNase resistant. Our crystal structure analysis of CTD

reveals a tight dimer mediated by an exchange of secondary structure elements (domain swapping), thus suggesting that the full-length N protein also functions as a dimer, with the CTD providing a structural scaffold while the NTD mainly serves as a module for RNA interaction. The relative orientation of the NTD with respect to the CTD in the N protein remains unknown, because in the full-length protein these two domains are connected by a 47-residue protease-sensitive loop, rich in Ser and Gly residues, which is presumably mobile. In the filamentous ribonucleocapsid, thanks to the flexibility provided by this linker, the RNA binding regions of these two domains could face each other, engulfing the RNA between them, thus conferring resistance to RNases.

In the various crystal forms studied, the NTD and CTD self-associate in multiple modes with buried surface areas greater than  $1,000 \text{ \AA}^2$ . A relevant question is which (if any) of these interactions are used in the formation of the nucleocapsid. Both the type-S and -F interactions between CTDs are conducive to the formation of fibril structures. Propagation of

any single type of interactions, however, would lead to a rigid helical nucleocapsid. Considering that the coronavirus nucleocapsid is not a rodlike structure, nucleocapsid assembly may thus involve a combination of the various interactions observed in our studies. The type-S interface could be used to a greater extent given that it occurs over a wider range of pH and seems to form regardless of the constraints imposed by crystal-packing forces (as in the CTD2 crystals). A combination of the type-S interactions with types L and F would modulate the curvature of the nucleocapsid in the virion (Fig. 5C).

**Similarity with other coronavirus N proteins.** Coronaviruses have been classified into four groups, with SARS-CoV being the founding member of an independent group. The N protein sequences are more similar within each group (~40% identity) than across groups (20 to 30% identity). The only X-ray structure of a coronavirus N protein available to date is that of the IBV NTD protein as described here and by Fan et al. (14). Despite very low sequence similarity between IBV and the SARS-CoV N proteins, their NTD and CTD structures adopt the same general polypeptide fold. This suggests that this fold is essentially preserved across the various coronavirus groups. This conclusion is also consistent with the NMR structures of the N- and C-terminal domains of the SARS N protein reported recently (4, 18).

**Comparison with the N proteins of other positive-strand ssRNA viruses.** Crystal structures of nucleocapsid proteins from several other positive-strand ssRNA viruses, including porcine reproductive and respiratory syndrome virus (PRRSV) in the *Arteriviridae* (11), West Nile virus in the *Flaviviridae* (12), and Sindbis virus and Semliki Forest virus in the *Togaviridae* family (6, 7), have been reported. The C-terminal domains of nucleoproteins from Sindbis virus and Semliki Forest virus adopt a chymotrypsinlike  $\beta$ -barrel fold. Core proteins from West Nile and dengue virus are dimers composed entirely of  $\alpha$ -helical bundles. As indicated by a systematic structural homology search (17), the coronavirus N protein CTD closely resembles the nucleocapsid protein of PRRSV. The PRRSV capsid-forming domain (C-terminal 73 to 123 amino acid residues) has a similar dimeric structure and exhibits self-association mediated by a salt bridge as seen in the IBV CTD (11) (Fig. 5D). Although a "domain-swapping" mode of oligomerization has been observed in a variety of proteins which are known to self-aggregate (24), the nature of domain swapping observed in these two viral proteins appears to be unique.

Interestingly, unlike the CTD fold which is shared with the arterivirus PRRSV, the fold of the NTD is observed only in the coronavirus N proteins (14). The corresponding basic N-terminal RNA binding domain in the much shorter PRRSV N protein appears to be largely disordered (11). Thus, the observed structural similarity between the N proteins of IBV and PRRSV suggests that members of the *Coronaviridae* and *Arteriviridae* families (order *Nidovirales*) share a common mechanism of filamentous nucleocapsid formation with suitable alterations necessary to interact specifically with their respective genomes. Conversely, the structural differences with other nucleocapsid proteins from ssRNA enveloped viruses, such as flaviviruses and togaviruses, probably reflect variations in their replication strategies and assembly pathways. Indeed, flaviviruses and togaviruses exhibit icosahedrally symmetric exteriors and are not pleomorphic like coronaviruses (27, 46). One com-

mon feature in the nucleocapsid proteins across positive-strand ssRNA viruses, however, appears to be the partitioning of the nucleocapsid protein structure into two domains: one forming a protective scaffold around the RNA through self-association and the other providing specific interactions with the viral genome. Electron-microscopic studies of ribonucleocapsid assemblies at higher resolution are now needed in order to gain further insights into the molecular mechanism of nucleocapsid formation for coronaviruses. However, we hope that the atomic description of the N protein provided here will stimulate such studies and also the design of molecules that could disrupt viral assembly.

#### ACKNOWLEDGMENTS

This work was supported by grants from the NIH (AI36040) and the Robert Welch Foundation to B.V.V.P. and grants from the Singapore Biomedical Research Council (03/1/22/17/220) and the Academic Research Fund (RG119/05) to J.L.

We thank Jennifer Falon and Florante Quiocho for use of the in-house X-ray diffraction facility at BCM. H.J. wishes to thank Chris Miller and HHMI for support during the latter half of this project. We acknowledge use of the SBC-CAT 19ID and BIOCARS BM14 beam line and we acknowledge its staff for their help during data collection at the Advanced Photon Source supported by the U.S. Department of Energy, Basic Energy Sciences, Office of Science, under contract no. W-31-109-Eng-38.

#### REFERENCES

1. Baric, R. S., G. W. Nelson, J. O. Fleming, R. J. Deans, J. G. Keck, N. Casteel, and S. A. Stohman. 1988. Interactions between coronavirus nucleocapsid protein and viral RNAs: implications for viral transcription. *J. Virol.* **62**: 4280–4287.
2. Bricogne, G., C. Vonrhein, C. Flensburg, M. Schiltz, and W. Paoletti. 2003. Generation, representation and flow of phase information in structure determination: recent developments in and around SHARP 2.0. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **59**:2023–2030.
3. Brunger, A. T., P. D. Adams, G. M. Clore, W. L. DeLano, P. Gros, R. W. Grosse-Kunstleve, J. S. Jiang, J. Kuszewski, M. Nilges, N. S. Pannu, R. J. Read, L. M. Rice, T. Simonson, and G. L. Warren. 1998. Crystallography & NMR system: a new software suite for macromolecular structure determination. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **54**:905–921.
4. Chang, C. K., S. C. Sue, T. H. Yu, C. M. Hsieh, C. K. Tsai, Y. C. Chiang, S. J. Lee, H. H. Hsiao, W. J. Wu, C. F. Chang, and T. H. Huang. 2005. The dimer interface of the SARS coronavirus nucleocapsid protein adapts a porcine respiratory and reproductive syndrome virus-like structure. *FEBS Lett.* **579**: 5663–5668.
5. Chen, H., A. Gill, B. K. Dove, S. R. Emmett, C. F. Kemp, M. A. Ritchie, M. Dee, and J. A. Hiscox. 2005. Mass spectroscopic characterization of the coronavirus infectious bronchitis virus nucleoprotein and elucidation of the role of phosphorylation in RNA binding by using surface plasmon resonance. *J. Virol.* **79**:1164–1179.
6. Choi, H. K., G. Lu, S. Lee, G. Wengler, and M. G. Rossmann. 1997. Structure of Semliki Forest virus core protein. *Proteins* **27**:345–359.
7. Choi, H. K., L. Tong, W. Minor, P. Dumas, U. Boege, M. G. Rossmann, and G. Wengler. 1991. Structure of Sindbis virus core protein reveals a chymotrypsin-like serine proteinase and the organization of the virion. *Nature* **354**:37–43.
8. Collaborative Computational Project, No. 4. 1994. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **50**:760–763.
9. Cologna, R., and B. G. Hogue. 1998. Coronavirus nucleocapsid protein. RNA interactions. *Adv. Exp. Med. Biol.* **440**:355–359.
10. DeLano, W. L. 2002. The PyMOL Molecular Graphics System, version 0.98. [Online.] <http://www.pymol.org>.
11. Doan, D. N., and T. Dokland. 2003. Structure of the nucleocapsid protein of porcine reproductive and respiratory syndrome virus. *Structure* **11**:1445–1451.
12. Dokland, T., M. Walsh, J. M. Mackenzie, A. A. Khromykh, K. H. Ee, and S. Wang. 2004. West Nile virus core protein; tetramer structure and ribbon formation. *Structure* **12**:1157–1163.
13. Emsley, P., and K. Cowtan. 2004. Coot: model-building tools for molecular graphics. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **60**:2126–2132.
14. Fan, H., A. Ooi, Y. W. Tan, S. Wang, S. Fang, D. X. Liu, and J. Lescaer. 2005. The nucleocapsid protein of coronavirus infectious bronchitis virus: crystal



- structure of its N-terminal domain and multimerization properties. *Structure* **13**:1859–1868.
15. **Gouet, P., E. Courcelle, D. I. Stuart, and F. Metz.** 1999. ESPript: analysis of multiple sequence alignments in PostScript. *Bioinformatics* **15**:305–308.
  16. **He, R., A. Leeson, A. Andonov, Y. Li, N. Bastien, J. Cao, C. Osioy, F. Dobie, T. Cutts, M. Ballantine, and X. Li.** 2003. Activation of AP-1 signal transduction pathway by SARS coronavirus nucleocapsid protein. *Biochem. Biophys. Res. Commun.* **311**:870–876.
  17. **Holm, L., and C. Sander.** 1998. Touring protein fold space with Dali/FSSP. *Nucleic Acids Res.* **26**:316–319.
  18. **Huang, Q., L. Yu, A. M. Petros, A. Gunasekera, Z. Liu, N. Xu, P. Hajduk, J. Mack, S. W. Fesik, and E. T. Olejniczak.** 2004. Structure of the N-terminal RNA-binding domain of the SARS CoV nucleocapsid protein. *Biochemistry* **43**:6059–6063.
  19. **Kuo, L., and P. S. Masters.** 2002. Genetic evidence for a structural interaction between the carboxy termini of the membrane and nucleocapsid proteins of mouse hepatitis virus. *J. Virol.* **76**:4987–4999.
  20. **Lai, M. M., and D. Cavanagh.** 1997. The molecular biology of coronaviruses. *Adv. Virus Res.* **48**:1–100.
  21. **Lamzin, V. S., A. Perrakis, and K. S. Wilson.** 2001. The ARP/WARP suite for automated construction and refinement of protein models, p. 720–722. *In* M. G. Rossmann and E. Arnold (ed.), *International tables for crystallography*, vol. F. Crystallography of biological macromolecules. Kluwer Academic Publishers, Dordrecht, The Netherlands.
  22. **Laskowski, R. A., M. W. MacArthur, D. S. Moss, and J. M. Thornton.** 1993. PROCHECK: a program to check the stereochemical quality of protein structures. *J. Appl. Crystallogr.* **26**:283–291.
  23. **Leung, D. T., F. C. Tam, C. H. Ma, P. K. Chan, J. L. Cheung, H. Niu, J. S. Tam, and P. L. Lim.** 2004. Antibody response of patients with severe acute respiratory syndrome (SARS) targets the viral nucleocapsid. *J. Infect. Dis.* **190**:379–386.
  24. **Liu, Y., and D. Eisenberg.** 2002. 3D domain swapping: as domains continue to swap. *Protein Sci.* **11**:1285–1299.
  25. **Luo, C., H. Luo, S. Zheng, C. Gui, L. Yue, C. Yu, T. Sun, P. He, J. Chen, J. Shen, X. Luo, Y. Li, H. Liu, D. Bai, J. Shen, Y. Yang, F. Li, J. Zuo, R. Hilgenfeld, G. Pei, K. Chen, X. Shen, and H. Jiang.** 2004. Nucleocapsid protein of SARS coronavirus tightly binds to human cyclophilin A. *Biochem. Biophys. Res. Commun.* **321**:557–565.
  26. **Luo, H., D. Wu, C. Shen, K. Chen, X. Shen, and H. Jiang.** 2006. Severe acute respiratory syndrome coronavirus membrane protein interacts with nucleocapsid protein mostly through their carboxyl termini by electrostatic attraction. *Int. J. Biochem. Cell Biol.* **38**:589–599.
  27. **Mukhopadhyay, S., R. J. Kuhn, and M. G. Rossmann.** 2005. A structural perspective of the flavivirus life cycle. *Nat. Rev. Microbiol.* **3**:13–22.
  28. **Narayanan, K., K. H. Kim, and S. Makino.** 2003. Characterization of N protein self-association in coronavirus ribonucleoprotein complexes. *Virus Res.* **98**:131–140.
  29. **Nelson, G. W., S. A. Stohman, and S. M. Tahara.** 2000. High affinity interaction between nucleocapsid protein and leader/intergenic sequence of mouse hepatitis virus RNA. *J. Gen. Virol.* **81**:181–188.
  30. **Nicholls, A., and B. Honig.** 1991. A rapid finite difference algorithm, using successive over-relaxation to solve the Poisson-Boltzmann equation. *J. Comput. Chem.* **12**:435–445.
  31. **Otwinowski, Z., and W. Minor.** 1997. Processing of X-ray diffraction data collected in oscillation mode, p. 307–326. *In* C. W. J. Carter and R. M. Sweet (ed.), *Methods in enzymology*, vol. 276. Academic Press, New York, N.Y.
  32. **Pannu, N. S., G. N. Murshudov, E. J. Dodson, and R. J. Read.** 1998. Incorporation of prior phase information strengthens maximum-likelihood structure refinement. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **54**:1285–1294.
  33. **Risco, C., I. M. Anton, L. Enjuanes, and J. L. Carrascosa.** 1996. The transmissible gastroenteritis coronavirus contains a spherical core shell consisting of M and N proteins. *J. Virol.* **70**:4773–4777.
  34. **Risco, C., M. Muntion, L. Enjuanes, and J. L. Carrascosa.** 1998. Two types of virus-related particles are found during transmissible gastroenteritis virus morphogenesis. *J. Virol.* **72**:4022–4031.
  35. **Schelle, B., N. Karl, B. Ludewig, S. G. Siddell, and V. Thiel.** 2005. Selective replication of coronavirus genomes that express nucleocapsid protein. *J. Virol.* **79**:6620–6630.
  36. **Siddell, S. G.** 1995. *The Coronaviridae*. Plenum Press, New York, N.Y.
  37. **Stohman, S. A., and M. M. Lai.** 1979. Phosphoproteins of murine hepatitis viruses. *J. Virol.* **32**:672–675.
  38. **Storoni, L. C., A. J. McCoy, and R. J. Read.** 2004. Likelihood-enhanced fast rotation functions. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **60**:432–438.
  39. **Surjit, M., B. Liu, S. Jameel, V. T. Chow, and S. K. Lal.** 2004. The SARS coronavirus nucleocapsid protein induces actin reorganization and apoptosis in COS-1 cells in the absence of growth factors. *Biochem. J.* **383**:13–18.
  40. **Surjit, M., B. Liu, P. Kumar, V. T. Chow, and S. K. Lal.** 2004. The nucleocapsid protein of the SARS coronavirus is capable of self-association through a C-terminal 209 amino acid interaction domain. *Biochem. Biophys. Res. Commun.* **317**:1030–1036.
  41. **Tahara, S. M., T. A. Dietlin, G. W. Nelson, S. A. Stohman, and D. J. Manno.** 1998. Mouse hepatitis virus nucleocapsid protein as a translational effector of viral mRNAs. *Adv. Exp. Med. Biol.* **440**:313–318.
  42. **Terwilliger, T. C., and J. Berendzen.** 1999. Automated MAD and MIR structure solution. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **55**:849–861.
  43. **Weeks, C. M., and R. Miller.** 1999. Optimizing shake-and-bake for proteins. *Acta Crystallogr. Sect. D Biol. Crystallogr.* **55**:492–500.
  44. **Wurm, T., H. Chen, T. Hodgson, P. Britton, G. Brooks, and J. A. Hiscox.** 2001. Localization to the nucleolus is a common feature of coronavirus nucleoproteins, and the protein may disrupt host cell division. *J. Virol.* **75**:9345–9356.
  45. **Yu, I. M., C. L. Gustafson, J. Diao, J. W. Burgner II, Z. Li, J. Zhang, and J. Chen.** 2005. Recombinant severe acute respiratory syndrome (SARS) coronavirus nucleocapsid protein forms a dimer through its C-terminal domain. *J. Biol. Chem.* **280**:23280–23286.
  46. **Zhang, Y., J. Corver, P. R. Chipman, W. Zhang, S. V. Pletnev, D. Sedlak, T. S. Baker, J. H. Strauss, R. J. Kuhn, and M. G. Rossmann.** 2003. Structures of immature flavivirus particles. *EMBO J.* **22**:2604–2613.
  47. **Zhou, M., A. K. Williams, S. I. Chung, L. Wang, and E. W. Collisson.** 1996. The infectious bronchitis virus nucleocapsid protein binds RNA sequences in the 3' terminus of the genome. *Virology* **217**:191–199.