Crystal Structure of the Severe Acute Respiratory Syndrome (SARS) Coronavirus Nucleocapsid Protein Dimerization Domain Reveals Evolutionary Linkage between Corona- and Arteriviridae*

Received for publication, March 6, 2006, and in revised form, April 11, 2006 Published, JBC Papers in Press, April 20, 2006, DOI 10.1074/jbc.M602107200

I-Mei Yu†1, Michael L. Oldham†1, Jingqiang Zhang‡, and Jue Chen‡1

From the 1Department of Biological Sciences and the Cancer Center, Purdue University, West Lafayette, Indiana 47907 and the 2State Key Laboratory for Biocontrol, Zhongshan University, Guangzhou 510275, China

The causative agent of severe acute respiratory syndrome (SARS) is the SARS-associated coronavirus, SARS-CoV. The nucleocapsid (N) protein plays an essential role in SARS-CoV genome packaging and virion assembly. We have previously shown that SARS-CoV N protein forms a dimer in solution through its C-terminal domain. In this study, the crystal structure of the dimerization domain, consisting of residues 270–370, is determined to 1.75 Å resolution. The structure shows a dimer with extensive interactions between the two subunits, suggesting that the dimeric form of the N protein is the functional unit in vivo. Although lacking significant sequence similarity, the dimerization domain of SARS-CoV N protein has a fold similar to that of the nucleocapsid protein of the porcine reproductive and respiratory syndrome virus. This finding provides structural evidence of the evolutionary link between Coronaviridae and Arteriviridae, suggesting that the N proteins of both viruses have a common origin.

Coronaviruses are enveloped, single-stranded, positive-sense RNA viruses that infect a variety of mammals and birds. Although previously identified human coronaviruses cause only mild respiratory infections, in the 2003 outbreak of severe acute respiratory syndrome (SARS),3 a disease caused by a new type of coronavirus (SARS-CoV), there were more than 8,000 cases resulting in 774 deaths (~10% mortality). Phylogenetic analysis suggests that SARS-CoV diverged early from group 2 coronaviruses and has evolved independently for a long period of time (1).

The coronavirus genome, containing ~30,000 bases, is the largest among positive-sense RNA viruses (2, 3). It encodes non-structural proteins including the RNA polymerase and helicase, as well as the spike (S), envelope (E), membrane (M), and nucleocapsid (N) structural proteins. The coronavirus virion is about 120 nm in diameter and consists of a lipid envelope containing three or four anchored glycoproteins and a helical ribonucleoprotein core (4). The surface projections forming the crown-like structure observed via electron microscopy are made up of the S protein, which is responsible for receptor recognition and membrane fusion (5–9). The integral membrane proteins M and E are essential for virus budding. When co-expressed in animal cells, the M and E proteins are sufficient to form virus-like particles (10). The N protein interacts with the viral genome to form the ribonucleoprotein core and has been shown to be involved in viral RNA synthesis, transcriptional regulation of genomic RNA, translation of viral proteins, and budding (2, 11, 12).

Coronaviruses are related to arteriviruses by their similar genome organizations and viral replication mechanisms (3, 13). Recently, Coronaviridae and Arteriviridae were united to form the new order Nidovirales. The name of the order comes from a property common to both viruses, a nested set of subgenomic mRNAs for structural protein expression (Latin nidus, meaning nest). The replicate genes of arteriviruses and coronaviruses are thought to have a common origin since they both have conserved domains that are present at the same relative positions (2, 14, 15). The structural proteins, however, are thought to be unrelated due to differences in protein size and lack of sequence similarity (15).

Coronavirus N proteins are composed of three distinct domains. High resolution structures of the N-terminal domain (~130 residues) were determined by NMR (16) and crystallography (17). This region folds similarly to the U1A RNA-binding protein and is suggested to bind RNA (16, 17). The central region of the N protein has also been shown by several laboratories to be an RNA-binding domain (~21). No structural information is available for the central domain, possibly due to its high positive charge and flexible nature. Recently, secondary structure elements of residues 248–365 of SARS-CoV N protein have been reported from NMR studies; however, the tertiary structure was not resolved (22). We have previously shown that the full-length SARS N protein, consisting of 422 amino acids, forms a dimer in solution through its C-terminal domain. In this study, we report the crystal structure of the dimerization domain of SARS-CoV N protein consisting of residues 270–370. Consistent with biochemical studies showing that this domain mediates dimer formation of the N protein (17, 23), the structure consists of a dimer with extensive interactions between subunits, suggesting that the N protein is not stable in the monomeric form and that the dimeric structure represents the functional unit of the N protein. Furthermore, the dimerization domain of SARS-CoV N protein shows a similar fold to that of the N protein of porcine reproductive and respiratory syndrome virus (PRRSV), a member of Arteriviridae. The structural similarity between these two N proteins, in the absence

* This work was supported by National Institutes of Health Grant P01 AI055672 (to R.J.K.) and by a grant from the Pew Scholars Program in the Biomedical Sciences (to J.C.C.). The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

‡ These authors contributed equally to this work.

†1 To whom correspondence should be addressed. Tel.: 765-496-3113; Fax: 765-496-1189; E-mail: chenjue@purdue.edu.

‡ These abbreviations used are: SARS, severe acute respiratory syndrome; SARS-CoV, SARS-associated coronavirus; PRRSV, porcine reproductive and respiratory syndrome virus; SeMet, selenomethionine; BDV, borna disease virus; Bis-Tris, 2-[bis(2-hydroxyethyl)amino]-2-(hydroxymethyl)propane-1,3-diol.

§ This work was supported by National Institutes of Health Grant P01 AI055672 (to R.J.K.) and by a grant from the Pew Scholars Program in the Biomedical Sciences (to J.C.C.).
of significant sequence similarity, adds further evidence that Coronaviridae and Arteriviridae have evolved from a common ancestor.

EXPERIMENTAL PROCEDURES

Cloning, Expression, and Purification—DNAs encoding various fragments of SARS-CoV (strain GD01) nucleocapsid protein, SARS N, were cloned into a pMCSG7 plasmid with an N-terminal His6 tag and a tobacco etch virus protease cleavage site. For selenomethionine (SeMet) derivatized protein, the methionine auxotroph strain B834(DE3) cells (Invitrogen) were transformed with the expression plasmid. The cells were grown to log phase at 37 °C in minimal medium supplemented with SeMet (Sigma) and cooled to 16 °C, and protein expression was induced by the addition of 0.1 mM isopropyl-1-thio-D-galactopyranoside (28) and refinement in REFMAC5 in CCP4 (27).

SeMet protein crystal (see Table 1, refinement set) was used for refinement using CNS (26). The initial model was built automatically using the program SOLVE (25) and improved by density modification in O (28) and refinement in REFMAC5 in CCP4 (27).

Effect of Metal Ions on the dimerization domain—The full-length SARS-CoV N protein (SARS N) and 18 truncated constructs were expressed in BL21star(DE3) cells and purified by nickel affinity chromatography (Amersham Biosciences; Superdex 75) in lysis buffer plus 10 mM dithiothreitol to keep all SeMet in the reduced form. Incorporation of SeMet was confirmed by mass spectrometry and fluorescence scans around the selenium potassium absorption edge.

The native N protein fragments were expressed in BL21star(DE3) cells and purified similarly except that no dithiothreitol was used in the buffer.

Crystallization and Data Processing—Protein samples were dialyzed against 10 mM Tris at pH 8.5 and 70 mM NaCl and concentrated to 45 mg/ml. Crystals were grown by mixing protein solution with the reservoir solution containing 30–33% pentane-1,3-diol, 15/4 ethanediol/hydroxyethane, 50 mM (NH4)2SO4, and 50 mM Bis-Tris, pH 6.5, at 1:1 ratio in sitting drops by vapor diffusion at 4 °C. Crystals were looped out of the drop and flash-frozen in liquid nitrogen. Data were collected at 100 K with a Quantum-Q315 CCD (ADSC) detector on beamline 19-ID at the Advanced Photon Source (Structural Biology Center, Argonne National Laboratory). X-ray diffraction data for the initial structure determination were collected to 2.2 Å resolution from crystals containing SeMet protein (see Table 1, phasing set). Data from both lattices of the twin domains were processed separately with HKL2000 and merged together in Scalepack. Phases were obtained from the four selenium sites found by multiple wavelength anomalous diffraction using the program SOLVE (25) and improved by density modification using CNS (26). The initial model was built automatically using Arp/Warp in CCP4 (27).

RESULTS

Crystallization and Structure Determination—The full-length SARS-CoV N protein (SARS N) and 18 truncated constructs were expressed in

### Table 1

<table>
<thead>
<tr>
<th>Data Collection</th>
<th>Refinement set</th>
<th>Se-Met (phasing set)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Peak</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Inflection</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Remote</td>
</tr>
<tr>
<td>Space group C2</td>
<td>124.2</td>
<td>124.3</td>
</tr>
<tr>
<td>α, Å</td>
<td>150.5</td>
<td>150.6</td>
</tr>
<tr>
<td>β, Å</td>
<td>124.5</td>
<td>124.6</td>
</tr>
<tr>
<td>β′, Å</td>
<td>108.9</td>
<td>108.8</td>
</tr>
<tr>
<td>Resolution range, Å</td>
<td>30-1.75</td>
<td>30-1.75</td>
</tr>
<tr>
<td>Wavelength, Å</td>
<td>0.97929</td>
<td>0.97929</td>
</tr>
<tr>
<td>Unique reflections</td>
<td>24,749</td>
<td>27,506</td>
</tr>
<tr>
<td>Completenessa</td>
<td>99.2 (99.6)</td>
<td>100.0 (100.0)</td>
</tr>
<tr>
<td>Redundancy</td>
<td>2.9 (2.9)</td>
<td>8.2 (7.8)</td>
</tr>
<tr>
<td>Rmergeb</td>
<td>7.4 (12.0)</td>
<td>10.6 (16.5)</td>
</tr>
<tr>
<td>r.m.s.d.</td>
<td></td>
<td>12.8 (15.5)</td>
</tr>
</tbody>
</table>

### Table Notes

- Space group C2
- α, β, γ, and β′ are lattice parameters.
- Resolution range is in Å.
- Wavelength is in Å.
- Unique reflections are the number of observed reflections.
- Completeness is in %.
- Redundancy is in %.
- Rmerge is the mean square deviation.

### Table Formulas

<table>
<thead>
<tr>
<th>Data Collection</th>
<th>Refinement set</th>
<th>Se-Met (phasing set)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Peak</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Inflection</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Remote</td>
</tr>
<tr>
<td>Space group C2</td>
<td>124.2</td>
<td>124.3</td>
</tr>
<tr>
<td>α, Å</td>
<td>150.5</td>
<td>150.6</td>
</tr>
<tr>
<td>β, Å</td>
<td>124.5</td>
<td>124.6</td>
</tr>
<tr>
<td>β′, Å</td>
<td>108.9</td>
<td>108.8</td>
</tr>
<tr>
<td>Resolution range, Å</td>
<td>30-1.75</td>
<td>30-1.75</td>
</tr>
<tr>
<td>Wavelength, Å</td>
<td>0.97929</td>
<td>0.97929</td>
</tr>
<tr>
<td>Unique reflections</td>
<td>24,749</td>
<td>27,506</td>
</tr>
<tr>
<td>Completenessa</td>
<td>99.2 (99.6)</td>
<td>100.0 (100.0)</td>
</tr>
<tr>
<td>Redundancy</td>
<td>2.9 (2.9)</td>
<td>8.2 (7.8)</td>
</tr>
<tr>
<td>Rmergeb</td>
<td>7.4 (12.0)</td>
<td>10.6 (16.5)</td>
</tr>
<tr>
<td>r.m.s.d.</td>
<td></td>
<td>12.8 (15.5)</td>
</tr>
</tbody>
</table>

#### Table Formulas

- Rmerge = Σ|F_o| − |F_c| / Σ|F_o|
- r.m.s.d. = root mean square deviation
bacteria and purified for crystallization. Although most constructs yielded either no crystals or merely clusters of small needles, a C-terminal construct containing residues 270–370 (cSARS-N) was readily crystallized to give well diffracting crystals. Although the crystals appear to be composed of a single lattice as viewed under a microscope fitted with a light polarizer, the diffraction pattern indicated that the crystals were unusually twinned, resulting in two overlapping monoclinic lattices of similar unit cell dimensions. The two lattices are related by a 180° rotation around the y axis followed by a 50° rotation around the z axis. As a result, nearly half of the observable reflections overlapped to some extent. Both native and SeMet derivatized proteins formed twinned crystals regardless of changes in temperature, precipitant, pH, and salt concentration. Because of the large number of partially overlapped reflections, only the central four pixels of each reflection were measured, these being proportional to the integrated intensity presuming the profile was fairly constant (details of the data processing procedure will be published elsewhere).

The structure of cSARS-N was determined by multiwavelength anomalous diffraction using a SeMet derivative to 1.75 Å resolution. The crystal belongs to space group C2, and the asymmetric unit contains a dimer of two subunits related by a 180° rotation. The final model, containing residues 270–366 of one subunit and residues 274–369 of the other, was refined to Rwork/Rfree of 18.7/23.5 (Table 1). A representative 2Fo – Fc electron density map is shown in Fig. 1. The model, analyzed with PROCHECK (29), has 97.5% of non-glycine residues in the most favored regions of the Ramachandran plot, with no disallowed residues.

Overall Structure—Sequence comparison shows that the dimerization domain of the nucleocapsid protein is conserved among the three groups of coronaviruses, suggesting a common structural and functional role of this domain (Fig. 2A). The monomer of cSARS-N contains five short α-helices, one 310 helix, and two β-strands. The overall shape of the monomer resembles the letter C, with one edge formed by a β-hairpin extending away from the rest of the molecule (Fig. 2B). In contrast to many small proteins that are remarkably compact, the monomeric cSARS-N domain folds into an extended conformation with a large cavity in its center. It is therefore likely that the N protein is not stable in the monomeric form and that higher oligomerization of the N protein is necessary to disrupt the dimer in solution (23). The strong interactions between the two subunits explain the previous observation that a significant amount of denaturant (4 M urea) was necessary to disrupt the dimer in solution (23).

A recent NMR study reported secondary structural assignments of a SARS N protein construct that is similar to cSARS-N (22). Despite the difficulty in resolving the three-dimensional structure, a dimeric interface consisting of a four-stranded anti-parallel β-sheet and two α-helices was proposed (22). Comparing the crystal structure with the proposed NMR structure, one of the most notable differences is the location of the longest helix, helix E (residues 346–357). In the crystal structure, helices E of both subunits are located at the edge of the dimer molecule, away from each other (Fig. 2C), whereas in the structure proposed by NMR, they are placed at the interface to stabilize the dimer (22).

Structural Similarity to the Nucleocapsid Protein of Arterivirus—A DALI search of the Protein Data Bank did not identify any other proteins with similar folds to that of cSARS-N. However, common structural features can be found between cSARS-N and the nucleocapsid protein of PRRSV (Fig. 2D). PRRSV belongs to the recently recognized family Arteriviridae, which is in the same order as Coronaviridae. The crystal structure of the C-terminal 65 residues of PRRSV N protein (30) is a homodimer. Each subunit consists of two central anti-parallel β-strands flanked by three α-helices. Although the sequences of PRRSV N and cSARS-N have no significant similarity (Fig. 3A), superposition of the monomeric N proteins of PRRSV and SARS-CoV shows that the two have similar folds in a core region of ~50 residues (Fig. 3B). The central β-strands are superimposable, and the helices following the β-strands are of similar length and position (Fig. 3B). In addition, as proposed in the recent NMR study (22), the dimer interface of the two N proteins are very similar, noted by the intertwining of the central β-strands to form a four-stranded sheet (Fig. 3C). This is evident by a root mean square deviation of 2.24 Å for the Cα positions among 38 residues at the dimer interface.
DISCUSSION

Specific packaging of the viral genome into the virion is a critical step in the life cycle of an infectious virus. The N protein plays an essential role in this process through self-association and interactions with viral RNA and other viral proteins. In an effort to understand the mechanism of how SARS N protein functions, we determined the crystal structure of its C-terminal dimerization domain. Extensive hydrogen-bonding and hydrophobic interactions are observed between the two subunits within the dimer found in the crystal, suggesting that the functional unit of the N protein is dimeric. Strong protein-protein interactions at the dimerization region may be essential to hold the putatively monomeric, highly charged RNA-binding domains in close proximity (16), thereby facilitating the formation of a large helical nucleocapsid core. Association of the N protein dimers is necessary for further assembly of the core. The full-length dimeric N protein has a propensity to form tetramers and higher molecular weight oligomers in vitro (23). A serine/arginine-rich motif (residues 184–196) was shown to be important for N protein oligomerization (31). Since constructs containing residues 211–422 or 285–422 of SARS N do not form oligomers larger than dimers (23), it is possible that the serine/arginine-rich motif located outside the dimerization domain is necessary to mediate further association of N protein dimers. Recently, the C-terminal 45 residues of the mouse hepatitis virus N protein were shown to be the major determinant for interaction with the M protein (32). Association of the N protein with the M protein may also play a role in the assembly of the nucleocapsid core into a progeny virion.

RNA viruses have high rates of sequence divergence and genome recombination, and thus it is often a challenge to study the evolutionary relationships among viruses. Structural studies have become an important method for revealing distantly related viruses. Among all lipid-enveloped RNA viruses, high-resolution structures of the proteins that package the viral genome into virions, often termed the nucleocapsid or capsid proteins, are currently available from five virus families (Fig. 4) (Retroviridae are not included in this discussion because their replication pathway is through a DNA intermediate (33)). Although functionally equivalent, the size and structure of nucleocapsid proteins are remarkably diverse among these five virus families. The nucleoprotein of borna disease virus (BDV), a single-stranded RNA virus in the family Bornaviridae, con-
Structure of SARS-CoV N Protein Dimerization Domain

FIGURE 4. Structural diversity of enveloped RNA virus nucleocapsid proteins. Representative known structural domains of nucleocapsid proteins of Bornaviridae (BDV), Togaviridae (dengue), Arteriviridae (PRRSV), and Coronaviridae (SARS-CoV) are shown. A, subunits of BDV N interact extensively to form a homotetramer, suggesting that the functional unit of BDV N is a tetramer (Fig. 4A). Both Semliki Forest virus (genus Alphavirus, family Togaviridae) and dengue virus (genus Flavivirus, family Flaviviridae) are small single-stranded RNA viruses with icosahedral symmetry. It has been suggested that they have a common ancestor because the structures of one of their surface glycoproteins are remarkably similar (35–37). Inside the lipid bilayer, however, the nucleocapsid cores of these two viruses are quite different. An alphavirus core consists of 240 copies of the capsid protein arranged in a T = 4 icosahedral lattice (38). Each capsid protein has ~270 residues, of which the N-terminal ~120 residues are flexible. The remaining 150 residues adopt a chymotrypsin-like fold and form homodimers in the crystal (39, 40) (Fig. 4B). In contrast, cryoelectron microscopy reconstructions of flaviviruses did not show a defined structure of the capsid core (41, 42), suggesting that the core either is disordered or has a different symmetry from the surface icosahedral lattice. In addition, the exact copy number of the capsid protein in the core is unclear. The C-terminal 100 residues of flavivirus capsid proteins fold into a four-helical structure with no similarity to that of an alphavirus capsid protein (43, 44) (Fig. 4C).

It is striking that, within the group of enveloped RNA viruses, no two nucleocapsid proteins are known to be structurally similar except those of Arteriviridae and Coronaviridae (Fig. 4). The structural similarity between the N proteins of SARS-CoV and PRRSV provides valuable information for understanding the evolutionary links between corona- and arteriviruses. Although these viruses were grouped in the same order, Nidovirales, their structural proteins were previously thought to be unrelated due to the marked difference in their sizes and lack of sequence similarity (15). In this study, a common fold of the dimerization domains of Nidovirales N proteins is observed, suggesting a possible common origin of these two proteins. The amino acid sequence and gene size diversity may have resulted from extensive mutation and RNA recombination during evolution. It is also important to note that the assembled cores of corona- and arteriviruses have completely different structures. PRRSV N protein interacts with the RNA genome to form a spherical, possibly icosahedral core (45), whereas coronaviruses are known to have helical nucleocapsids (46). It was previously suggested that the ancestral virus had an icosahedral nucleocapsid core, and the larger N protein of coronavirus, freed from its icosahedral package constraints, allowed the coronavirus genome to become larger during evolution (47).

Acknowledgments—We thank Dr. M. G. Rossmann for many helpful discussions and Drs. C. Gustafson, R. J. Ruhm, M. G. Rossmann, E. G. Strauss, and J. H. Strauss for critical reading of the manuscript. We also thank the beamline staff at SBC 19-ID at the Advanced Photon Source for assistance with data collection and the Purdue Cancer Center for x-ray and DNA sequencing facilities.

Addendum—While this work was in progress, the structure of the nucleocapsid protein of another coronavirus, the avian infectious bronchitis virus (IBV), was determined by Drs. H. Jayaram and B. V. Prasad at Baylor College of Medicine (unpublished data). The coordinates of the IBV N protein were kindly provided to us by them for phasing via molecular replacement. However, they were unnecessary since the experimental multilwavelength anomalous dispersion phasing was sufficient to build the initial model.

REFERENCES


Downloaded from http://www.jbc.org/ by guest on April 7, 2015
Protein Structure and Folding:
Crystal Structure of the Severe Acute Respiratory Syndrome (SARS) Coronavirus Nucleocapsid Protein Dimerization Domain Reveals Evolutionary Linkage between Corona- and Arteriviridae

I-Mei Yu, Michael L. Oldham, Jingqiang Zhang and Jue Chen
doi: 10.1074/jbc.M602107200 originally published online April 20, 2006

Access the most updated version of this article at doi: 10.1074/jbc.M602107200

Find articles, minireviews, Reflections and Classics on similar topics on the JBC Affinity Sites.

Alerts:
- When this article is cited
- When a correction for this article is posted

Click here to choose from all of JBC’s e-mail alerts

This article cites 48 references, 18 of which can be accessed free at http://www.jbc.org/content/281/25/17134.full.html#ref-list-1