

# Structure of the C-terminal domain of nsp4 from feline coronavirus

Ioannis Manolaridis,<sup>a</sup> Justyna A. Wojdyla,<sup>a</sup> Santosh Panjikar,<sup>a</sup> Eric J. Snijder,<sup>b</sup> Alexander E. Gorbalenya,<sup>b</sup> Hanna Berglind,<sup>c</sup> Pär Nordlund,<sup>c</sup> Bruno Coutard<sup>d</sup> and Paul A. Tucker<sup>a\*</sup>

<sup>a</sup>EMBL Hamburg Outstation, c/o DESY, Notkestrasse 85, D-22603 Hamburg, Germany, <sup>b</sup>Molecular Virology Laboratory, Department of Medical Microbiology, Center of Infectious Diseases, Leiden University Medical Center, PO Box 9600, 2300 RC Leiden, The Netherlands, <sup>c</sup>Division of Biophysics, Department of Medical Biochemistry and Biophysics, Scheeles väg 2, Karolinska Institute, SE-171 77 Stockholm, Sweden, and <sup>d</sup>Laboratoire Architecture et Fonction des Macromolécules Biologiques, UMR 6098, AFMB-CNRS-ESIL, Case 925, 163 Avenue de Luminy, 13288 Marseille, France

Correspondence e-mail: tucker@embl-hamburg.de

Coronaviruses are a family of positive-stranded RNA viruses that includes important pathogens of humans and other animals. The large coronavirus genome (26–31 kb) encodes 15–16 nonstructural proteins (nsps) that are derived from two replicase polyproteins by autoproteolytic processing. The nsps assemble into the viral replication–transcription complex and nsp3, nsp4 and nsp6 are believed to anchor this enzyme complex to modified intracellular membranes. The largest part of the coronavirus nsp4 subunit is hydrophobic and is predicted to be embedded in the membranes. In this report, a conserved C-terminal domain (~100 amino-acid residues) has been delineated that is predicted to face the cytoplasm and has been isolated as a soluble domain using library-based construct screening. A prototypical crystal structure at 2.8 Å resolution was obtained using nsp4 from feline coronavirus. Unmodified and SeMet-substituted proteins were crystallized under similar conditions, resulting in tetragonal crystals that belonged to space group  $P4_3$ . The phase problem was initially solved by single isomorphous replacement with anomalous scattering (SIRAS), followed by molecular replacement using a SIRAS-derived composite model. The structure consists of a single domain with a predominantly  $\alpha$ -helical content displaying a unique fold that could be engaged in protein–protein interactions.

Received 15 April 2009

Accepted 14 May 2009

**PDB Reference:** C-terminal domain of nsp4, 3gzf, r3gzfsf.

## 1. Introduction

The *Coronaviridae* family, which is comprised of the genera *Coronavirus* and *Torovirus*, and the more distantly related *Arteriviridae* and *Roniviridae* families together form the order *Nidovirales* (Gorbalenya *et al.*, 2006). Coronaviruses are positive-stranded RNA viruses that are frequently associated with enteric or respiratory diseases in humans, livestock and companion animals (Dye & Siddell, 2005). At present, they are formally classified into three genetic groups (1–3), with the first two groups further divided into two subgroups (1a/b and 2a/b; Gorbalenya *et al.*, 2004; Lai & Holmes, 2001), but as our understanding of natural coronavirus diversity progresses novel subgroups continue to be recognized (Woo *et al.*, 2009). Viruses that belong to different subgroups have diverged profoundly. A fraction of their proteins are subgroup-specific and the amino-acid sequences of their most conserved proteins may differ by as much as 50%. The best-known member of this family, severe acute respiratory syndrome coronavirus (SARS-CoV), belongs to subgroup 2b, whereas feline coronavirus (FCoV), characterized in this study, belongs to subgroup 1a (Gorbalenya *et al.*, 2004).

Feline infectious peritonitis virus (FIPV) is a pathogenic FCoV variant that emerged by mutation of the relatively benign enteric FCoV (Poland *et al.*, 1996; Vennema *et al.*, 1998) and causes a fatal immune-mediated disease in cats (Pedersen, 1995). Since the variations are minor and despite the fact that we are working with a construct derived from an FIPV strain, we will henceforth use FCoV as an abbreviation for the virus. The FCoV genome (strain FIPV WSU-79/1146) consists of 29 125 nucleotides and contains six open reading frames (ORFs; Dye & Siddell, 2005). The first two ORFs, namely ORF1a and ORF1b, comprising the 5'-most gene (*i.e.* gene 1), encode two large replicase polyproteins, pp1a and pp1ab. Proteolytic cleavage of these polyproteins by virus-encoded proteinases, *i.e.* the 3C-like main proteinase (M<sup>pro</sup> in nsp5) and the papain-like accessory proteinases (PL<sup>pro</sup> 1 and 2 in nsp3), are predicted to give rise to a total of 16 mature nonstructural proteins (nsps; Ziebuhr *et al.*, 2000; Dye & Siddell, 2005). In addition to the proteases mentioned above, associated enzymatic activities have been identified for several coronavirus nsps, including deubiquitinating and adenosine diphosphate-ribose-1'-phosphatase (ADRP) functions (nsp3), RNA-dependent RNA polymerases with low (nsp8) and high (nsp12) processivity, helicase (nsp13), RNA exonuclease and N7-methyltransferase (nsp14), RNA endoribonuclease (nsp15) and 2'-O-methyltransferase (nsp16) (Anand *et al.*, 2003; Bhardwaj *et al.*, 2004; Chen *et al.*, 2009; Cheng *et al.*, 2005; Gorbalenya *et al.*, 1989; Harcourt *et al.*, 2004; Imbert *et al.*, 2008; Ivanov, Hertzog *et al.*, 2004; Ivanov, Thiel *et al.*, 2004; Ratia *et al.*, 2006; Seybert *et al.*, 2000; Snijder *et al.*, 2003). Tertiary structures, solved using X-ray or/and NMR analyses, have been reported for a substantial number of nsps from at least one coronavirus, typically SARS-CoV (reviewed in Bartlam *et al.*, 2005; Mesters *et al.*, 2006). These structures represent a variety of separate domains, entire proteins and even multiprotein complexes. Despite these remarkable advances, many domains, including those residing in nsp4, remain poorly characterized.

Nsp4 is an approximately 500-amino-acid replicase subunit that is released by the combined activity of the nsp3 and nsp5 proteases. It is predicted to be one of the three membrane-spanning proteins (the others are nsp3 and nsp6) among coronavirus nsps and bioinformatic analyses consistently predict four transmembrane domains in nsp4 (Clementz *et al.*, 2008; Oostra *et al.*, 2007). An N-terminal transmembrane region (amino acids 1–30) is presumably followed by a large luminal domain (amino acids 30–280), three closely spaced additional transmembrane regions (amino acids 280–400) and finally a C-terminal domain of about 100 residues that is exposed at the cytoplasmic face of the membrane. Coronavirus infection induces the extensive reorganization of endoplasmic reticulum membranes into a reticulovesicular network (Knoops *et al.*, 2008) that includes many unusual double-membrane vesicles (Gosert *et al.*, 2002; Harcourt *et al.*, 2004; Shi *et al.*, 1999; Snijder *et al.*, 2006; Stertz *et al.*, 2007). It is currently believed that nsp4 functions in anchoring the viral replication–transcription complex (RTC) to these modified membranes and independent genetic studies have demon-

strated its importance for replication (Clementz *et al.*, 2008; Sparks *et al.*, 2007).

In this paper, we present the first X-ray structure of the C-terminal domain of the FCoV nsp4. Together with structural data, a family-wide comparative sequence analysis of the nsp4 C-terminal domain was performed in order to identify residues/regions that might be important for function rather than for structural integrity.

## 2. Experimental procedures

### 2.1. Library-based construct screening

The sequence encoding the FCoV nsp4 (residues 2337–2826 of the polyprotein pp1a from strain FIPV WSU-79/1146; Genebank/RefSeq accession No. NC\_007025.1) was RT-PCR amplified from viral RNA and cloned into the pMM8 vector. pMM8 is a modified pET-43 (Novagen) bacterial expression vector containing restriction sites suitable for exonuclease-based construct-library generation (Cornvik *et al.*, 2006) and a Gateway cassette for recombination cloning inserted downstream of the His-tag coding sequence. An N-terminally deleted construct library was generated using an exonuclease strategy and screened for a soluble construct using the colony-filtration blot (Cornvik *et al.*, 2005, 2006). A soluble and well expressing construct containing residues 2731–2826 (here called the nsp4ct domain) was chosen for scale-up expression and purification. This construct has 14 additional N-terminal residues, including a noncleavable His<sub>6</sub> tag.

### 2.2. Expression

The expression of soluble nsp4ct was performed in *Escherichia coli* strain BL21 (DE3) (Novagen). Cultures were grown at 310 K in LB medium containing 50 µg ml<sup>-1</sup> ampicillin until the OD<sub>600</sub> reached 0.8. Protein synthesis was induced by the addition of 1 mM isopropyl β-D-1-thiogalactopyranoside (IPTG) and the culture was grown to stationary phase overnight at 288 K. Cells were harvested by centrifugation at 4000g (30 min, 277 K) and frozen at 253 K. Selenomethionine-substituted nsp4ct was expressed in the non-methionine auxotrophic *E. coli* strain BL21 (DE3) (Novagen). Bacteria were grown in minimal medium at 310 K until the OD<sub>600</sub> reached 0.8. Feedback-inhibition amino-acid mix (Lys, Thr, Phe, Leu, Ile, Val and SeMet) was added and after 15 min cells were induced with 1 mM IPTG. The culture was left shaking overnight at 288 K and the cells were subsequently harvested by centrifugation at 4000g (30 min, 277 K). Cell pellets were frozen at 253 K.

### 2.3. Purification

Both the native and the SeMet-substituted nsp4ct proteins were purified following the same protocol. A pellet from 1 l cell culture was resuspended in 20 ml buffer A (10 mM CHES pH 9.1 and 300 mM NaCl, plus 2 mM β-mercaptoethanol in the case of SeMet-substituted nsp4ct). The cells were sonicated and the protein was purified from the soluble cellular fraction by Ni–NTA affinity chromatography and eluted with

buffer *A* containing 500 mM imidazole. The eluate was buffer-exchanged into buffer *A* using PD10 columns (GE Healthcare Life Sciences). nsp4ct was subsequently concentrated and applied onto a Superdex 75 (16/60) gel-filtration column (GE Healthcare Life Sciences) pre-equilibrated with buffer *A*. The protein was concentrated to 10 mg ml<sup>-1</sup> and its purity was examined by SDS-PAGE.

## 2.4. Crystallization

Initial crystallization trials were carried out using the sitting-drop vapour-diffusion method in 96-well plates (Greiner) at 292 K at the EMBL Hamburg High-throughput Crystallization Facility (Mueller-Dieckmann, 2006). Crystals were obtained under various conditions. Further optimization of these conditions was performed manually in 24-well plates (Qiagen) using the hanging-drop vapour-diffusion method at 292 K. Crystals were obtained at a protein concentration of 7 mg ml<sup>-1</sup> in 0.22 M ammonium sulfate and 25% (w/v) PEG 5000.

## 2.5. Data collection and processing

The crystals were cryoprotected in a solution consisting of 0.22 M ammonium sulfate, 25% (w/v) PEG 5000 and 15% (v/v) ethylene glycol prior to data collection. Three data sets were collected: two single-wavelength native data sets (data sets 1 and 2) and a single-wavelength anomalous diffraction (SAD) data set (at peak wavelength; data set 3). Data set 1 was collected from a single crystal at 100 K on the European Synchrotron Radiation Facility (ESRF) beamline ID23-2 using a MAR 225 CCD detector. The oscillation range was 1°, with a crystal-to-detector distance of 346.2 mm. 90 images were collected to a maximum resolution of 3.1 Å. Data set 2 was collected from a single crystal at 100 K on the EMBL beamline X12 at DESY using a MAR 225 detector. The crystal-to-detector distance was 300 mm, with an oscillation range of 0.25°. A total of 670 images were collected to a maximum resolution of 2.76 Å. Data set 3 was also collected from a single crystal on beamline X12 (EMBL Hamburg). The crystal-to-detector distance was 280 mm and the oscillation range was 1°. 200 images at the selenium absorption edge were collected to a maximum resolution of 3.3 Å.

In all three cases the recorded images were processed with *XDS* (Kabsch, 1988) and the reflection intensities were processed with *COMBAT* and scaled with *SCALA* (Evans, 1993) from the *CCP4* program suite (Collaborative Computational Project, Number 4, 1994). Data-collection statistics are shown in Table 1.

## 2.6. Structure determination

The structure was solved using the SIRAS protocol of the *Auto-Rickshaw* automated crystal structure-determination platform (Panjikar *et al.*, 2005).  $F_A$  values were calculated using the program *SHELXC* (Sheldrick, 2008). Based on an initial analysis of the data, the maximum resolution for substructure determination and initial phase calculation was set to 3.8 Å. 20 selenium positions were found using the

program *SHELXD* (Sheldrick, 2008). The correct hand of the substructure was determined using the programs *ABS* (Hao, 2004) and *SHELXE* (Sheldrick, 2008). The occupancy of all substructure atoms was refined using the program *BP3* (Pannu *et al.*, 2003; Pannu & Read, 2004). The initial phases were improved using density modification, noncrystallographic symmetry (NCS) averaging and phase extension using the program *RESOLVE* (Terwilliger, 2000). A partial  $\alpha$ -helical model was produced using the program *HELICAP* (Morris *et al.*, 2004). The partial model contained 119 of the total of 440 residues expected for four molecules. The initial phases were improved by phase combination of experimental and model phases using the program *SIGMAA* (Read, 1986). The density modification and fourfold NCS averaging were repeated again as described above. The resultant phases were used to continue model building using the program *ARP/wARP* (Perrakis *et al.*, 1999), resulting in the placement of 242 residues. The partial models generated in the intermediate steps of *ARP/wARP* were then used to assemble an almost complete dimer using the graphics program *Coot* (Emsley & Cowtan, 2004). This dimer was then used to find the second dimer in the electron density using phased molecular-replacement techniques as implemented in *MOLREP* (Vagin & Teplyakov, 1997).  $2F_o - F_c$  and  $F_o - F_c$  electron-density maps calculated at this stage showed additional electron density indicating the presence of a fifth molecule in the asymmetric unit. The phased molecular replacement was repeated again to place the fifth molecule in the electron-density map. The resultant model was then used for restrained refinement in *REFMAC5* (Murshudov *et al.*, 1997), including use of the translation, libration and screw method (TLS; Schomaker & Trueblood, 1968) for describing group motions.

The structure was manually modified, followed by cycles of refinement, using the program *Coot*. The progress of the refinement was monitored by means of the free *R* factor (Brünger, 1992). Water molecules were included where clear peaks were present in both the  $2F_o - F_c$  and  $F_o - F_c$  maps and where appropriate hydrogen bonds could be made to surrounding residues or to other water molecules. The stereochemistry of the model was evaluated with the program *MOLPROBITY* (Davis *et al.*, 2007).

Interfaces between molecules were analyzed with the *PISA* server (Krissinel & Henrick, 2007). Interactions between molecules were initially evaluated using the *CCP4* program *CONTACT* with a maximum contact distance of 3.6 Å.

## 3. Results and discussion

### 3.1. Structure determination

The recombinant His<sub>6</sub>-tagged FCoV nsp4ct domain (residues 2731–2826 of pp1a and residues 395–490 of nsp4; here renumbered as 1–96) was expressed in *E. coli*. The protein was also expressed with the substitution of methionine by selenomethionine (SeMet). The incorporation of SeMet was verified by matrix-assisted laser desorption ionization (MALDI) mass spectrometry. Both native and SeMet-substituted proteins

**Table 1**

Data collection.

Values in parentheses are for the last resolution shell.

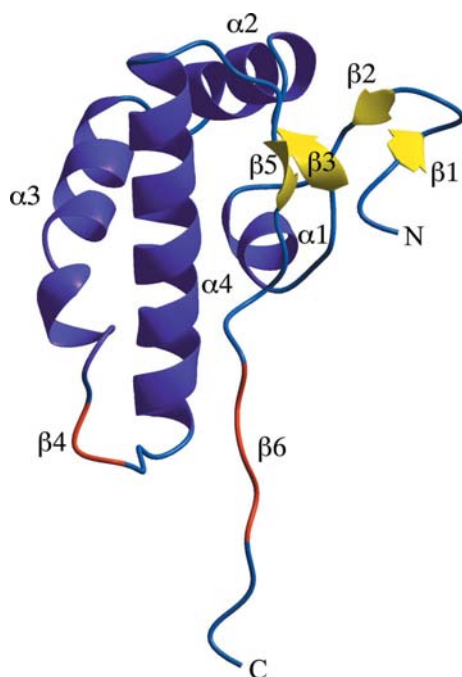
Data set	1 (native)	2 (native)	3 (SeMet)
Crystallization conditions	0.22 M (NH <sub>4</sub> ) <sub>2</sub> SO <sub>4</sub> , 25% (w/v) PEG 5000	0.22 M (NH <sub>4</sub> ) <sub>2</sub> SO <sub>4</sub> , 25% (w/v) PEG 5000	0.22 M (NH <sub>4</sub> ) <sub>2</sub> SO <sub>4</sub> , 25% (w/v) PEG 5000
X-ray source	ID23-2	X12	X12
Space group	<i>P</i> <sub>4</sub> <sub>3</sub>	<i>P</i> <sub>4</sub> <sub>3</sub>	<i>P</i> <sub>4</sub> <sub>3</sub>
Unit-cell parameters (Å)	<i>a</i> = <i>b</i> = 128.0, <i>c</i> = 43.7	<i>a</i> = <i>b</i> = 127.5, <i>c</i> = 42.7	<i>a</i> = <i>b</i> = 128.7, <i>c</i> = 42.3
Wavelength (Å)	0.872	0.978	0.9777
Resolution range (Å)	50.0–3.1	20.0–2.8	50.0–3.3
Mosaicity	0.14	0.16	0.2
Mean <i>I</i> /σ( <i>I</i> )	14.1 (2.9)	23.2 (3.6)	11.9 (4.0)
<i>R</i> <sub>fac</sub> (linear) (%)	8.5 (44.7)	5.3 (44.1)	9.7 (32.9)
Redundancy	3.6	6.8	3.8
<i>R</i> <sub>meas</sub> (%)	10.2 (56.2)	5.7 (48.0)	11.1 (37.8)
No. of observations	48030	109977	76895†
No. of unique reflections	13150	18178	20164†
Completeness (%)	99.3 (97.7)	99.1 (96.6)	99.2 (96.1)

† Friedel pairs were not merged.

**Table 2**

Refinement.

Space group	<i>P</i> <sub>4</sub> <sub>3</sub>
Resolution range (Å)	19.9–2.8
No. of reflections (working/free)	17247/927
No. of protein residues	<i>A</i> , 96; <i>B</i> , 93; <i>C</i> , 92; <i>D</i> , 91; <i>E</i> , 84
No. of waters	40
No. of sulfate molecules	2
<i>R</i> <sub>work</sub> / <i>R</i> <sub>free</sub> (%)	24.0/29.9
Average <i>B</i> (Å <sup>2</sup> )	68.7
R.m.s. deviation from ideal values	
Bond lengths (Å)	0.012
Bond angles (°)	1.6



**Figure 1**

Overall structure of nsp4ct domain (molecule *A* is shown).  $\alpha$ -Helices are shown in purple,  $\beta$ -strands are shown in yellow, loops and termini are shown in light blue and regions forming  $\beta$ -strands present only in the dimer interface are depicted in red.

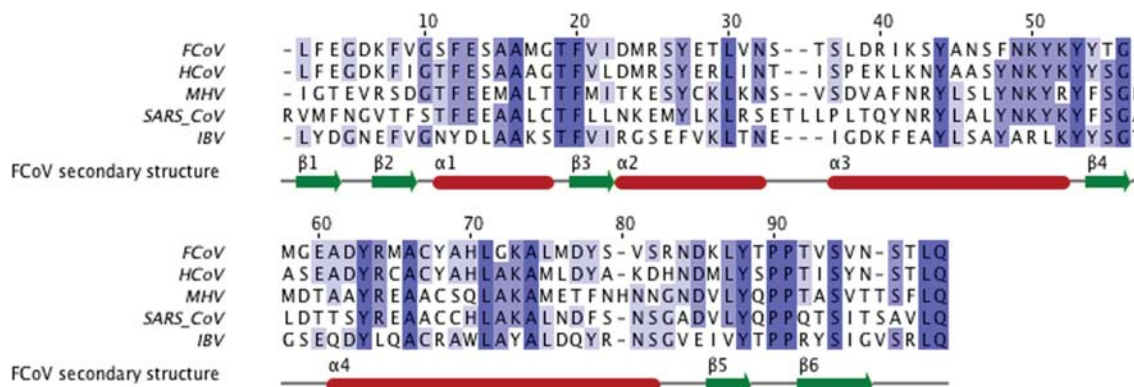
were crystallized from conditions containing ammonium sulfate and PEG 5000. The crystals belonged to space group *P*<sub>4</sub><sub>3</sub>, with unit-cell parameters *a* = *b* = 127.5, *c* = 42.8 Å (data set 2 in Table 1). There are five molecules (chains *A–E*) in the asymmetric unit, which corresponds to a 64% solvent content (Matthews, 1968). The structure was refined at 2.8 Å resolution to a final *R* value of 24.0% (*R*<sub>free</sub> = 29.9%). The final model contains 96 residues in molecule *A* (residues 0–95), 93 residues in molecule *B* (residues 0–49 and 53–95), 92 residues in molecule *C* (residues 0–91), 91 residues in molecule *D* (residues 1–91) and 84 residues in molecule *E* (residues 0–49 and 56–89). 88.0% of the residues are located in the preferred

regions of the Ramachandran diagram and 10.9% are in allowed regions. Residues Met55 (chains *A* and *B*), Glu57 (chains *C* and *D*) and Ala58 (chains *A–E*) are Ramachandran outliers. Glu57 and Ala58 are located in the N-terminus of helix  $\alpha$ 3. The geometry in this region may be influenced by hydrogen bonding between Glu57 O<sup>ε</sup> and Arg61 N<sup>η</sup>. The refined structure contains two sulfate ions and 40 solvent molecules. A detailed summary of the data-collection and structure-refinement statistics is given in Tables 1 and 2.

### 3.2. Overall structure

The FCoV nsp4ct structure contains six short  $\beta$ -strands  $\beta$ 1– $\beta$ 6 and four  $\alpha$ -helices  $\alpha$ 1– $\alpha$ 4 (Fig. 1). Strands  $\beta$ 1 and  $\beta$ 2 and strands  $\beta$ 3 and  $\beta$ 5 form small two-stranded antiparallel sheets. Strands  $\beta$ 4 and  $\beta$ 6 participate in the formation of the dimer interface. Strand  $\beta$ 4 is observed in molecules *C* and *D* and strand  $\beta$ 6 is observed in molecules *A–D*. The characteristic feature of the structure is the 21-residue-long helix  $\alpha$ 4. Analysis using EBI web tools (PDBsum/ProFunc, Catalytic site search; <http://www.ebi.ac.uk>), DALI ([http://ekhidna.biocenter.helsinki.fi/dali\\_server/](http://ekhidna.biocenter.helsinki.fi/dali_server/)) and GRATH (<http://protein.hbu.cn/cath/cathwww.biochem.ucl.ac.uk/cgi-bin/cath/Grath.html>) for both the monomer and dimer did not result in any significant indicators of function or similarities in structure. The structure therefore represents a novel protein fold. Nsp4ct has nine conserved and two nonconserved hydrophobic residues (Fig. 2) which form the hydrophobic core of the structure. These residues are grouped into a mainly aliphatic group (Phe19, Ile21, Leu29, Ile38, Leu68, Leu72) and two aromatic groups (Phe11, Tyr41, Tyr65 and Tyr26, Tyr75) (Fig. 3).

The r.m.s.d. between C<sup>α</sup> atoms of molecules *A–E* is less than 0.1 Å (for the 63 common C<sup>α</sup> atoms). Differences between molecules are present in the N-terminus (residues 0–4), the C-terminus (from residue 88 onwards) and the region between residues 46 and 64, which includes the C-terminal part of helix  $\alpha$ 3, the flexible loop L <sub>$\alpha$ 3– $\alpha$ 4</sub> and the N-terminus of helix  $\alpha$ 4.



**Figure 2**

Alignment of amino-acid sequences from nsp4ct proteins, coupled with secondary-structure information from the FCoV nsp4ct three-dimensional structure. The alignment is based on amino-acid data for feline infectious peritonitis virus (FCoV, NC\_007025.1), human coronavirus NL63 (HCoV, ABE97129), murine hepatitis virus (MHV, NP\_001012459.1), severe acute respiratory syndrome coronavirus (SARS\_CoV, NP\_904322.1) and infectious bronchitis virus strain Beaudette (IBV, NP\_740625). The alignment was produced with *ClustalW2* (Larkin *et al.*, 2007) and edited with *JalView*. Residues are coloured according to conservation from fully conserved (dark blue) to nonconserved (colourless).

### 3.3. Dimer interface

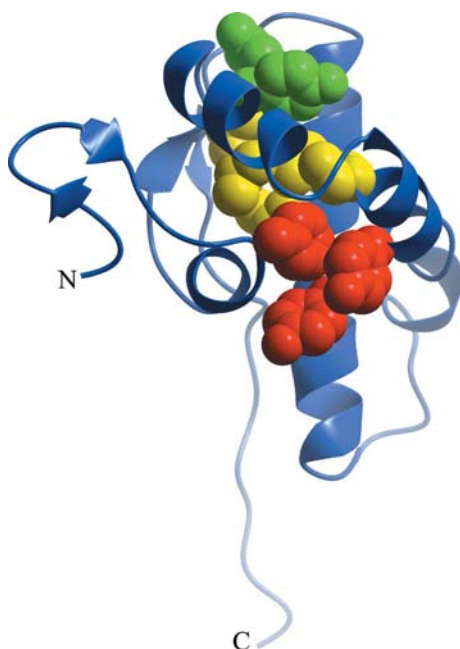
The FCoV nsp4ct crystal contains five molecules in the asymmetric unit. Molecules *A/C* and *B/D* form very similar dimers (Fig. 4a), each of approximate dimensions 60 × 20 × 20 Å. The average buried surface area per molecule is approximately 961 Å<sup>2</sup>. The buried surface of each dimer involves approximately 25 residues from α1, α3, α4, loop L<sub>α3-α4</sub> and the C-terminus. Interestingly, the interaction interface contains an intramolecular three-stranded anti-parallel β-sheet. The order of the strands in this sheet is β4<sub>C</sub>-

β6<sub>A</sub>-β6<sub>C</sub> in the case of dimer *A/C* and β4<sub>D</sub>-β6<sub>B</sub>-β6<sub>D</sub> in the case of dimer *B/D*. Strands β4<sub>C</sub> and β4<sub>D</sub> include residues 51–53, strands β6<sub>A</sub> and β6<sub>B</sub> include residues 89–92 and strands β6<sub>C</sub> and β6<sub>D</sub> include residues 88–90. The major interactions at this interface are the β-sheet hydrogen bonds Val89 N···Gly53 O, Val89 O···Gly53 N, Ser90 N···Ser90 O, Ser90 O···Ser90 N, Val91 N···Tyr51 O, Asn92 N···Tyr88 O and Asn92 O···Tyr88 N. Five hydrogen bonds located outside the β-sheet region are formed: Met55 O···Tyr60 O<sup>η</sup>, Tyr60 O<sup>η</sup>···Met55 O, Tyr60 O<sup>η</sup>···Met55 N, Thr94 O<sup>γ</sup>···Thr85 O and Thr94 O<sup>γ</sup>···Thr85 O<sup>γ</sup> (Fig. 4b). Strong van der Waals contacts between residues in the dimer buried area are also important in defining the interface.

The results obtained from analytical size-exclusion chromatography of nsp4ct demonstrated that the protein is monomeric in solution under the experimental conditions used (results not shown). Furthermore, the crystal structure contains a monomer as well as two dimers. The buried surface area supports the likelihood of dimerization and this may have physiological significance. It is conceivable that *in vivo* nsp4 dimerization during membrane modification or formation of the RTC may help in bringing the components together and could therefore aid their correct spatial orientation. This would agree with the previously proposed role of nsp4 as an anchor for the assembly of the viral RTC.

### 3.4. Nsp4ct sequence alignment

Fig. 2 shows the sequence alignment, produced with *ClustalW2* (Larkin *et al.*, 2007), of the C-terminal domain of nsp4 for the five coronaviral subgroups. These viruses are FCoV (group 1a), human coronavirus NL63 (HCoV-NL63; group 1b), murine hepatitis virus (MHV; group 2a), SARS-CoV (group 2b) and infectious bronchitis virus (IBV; group 3). The nsp4ct sequence identity between viruses belonging to the same group (but different subgroups) is higher than that for viruses belonging to different groups. The sequence identity between FCoV and HCoV-NL63 is 68% and that between MHV and SARS-CoV is 53%. IBV, on the other hand, is the



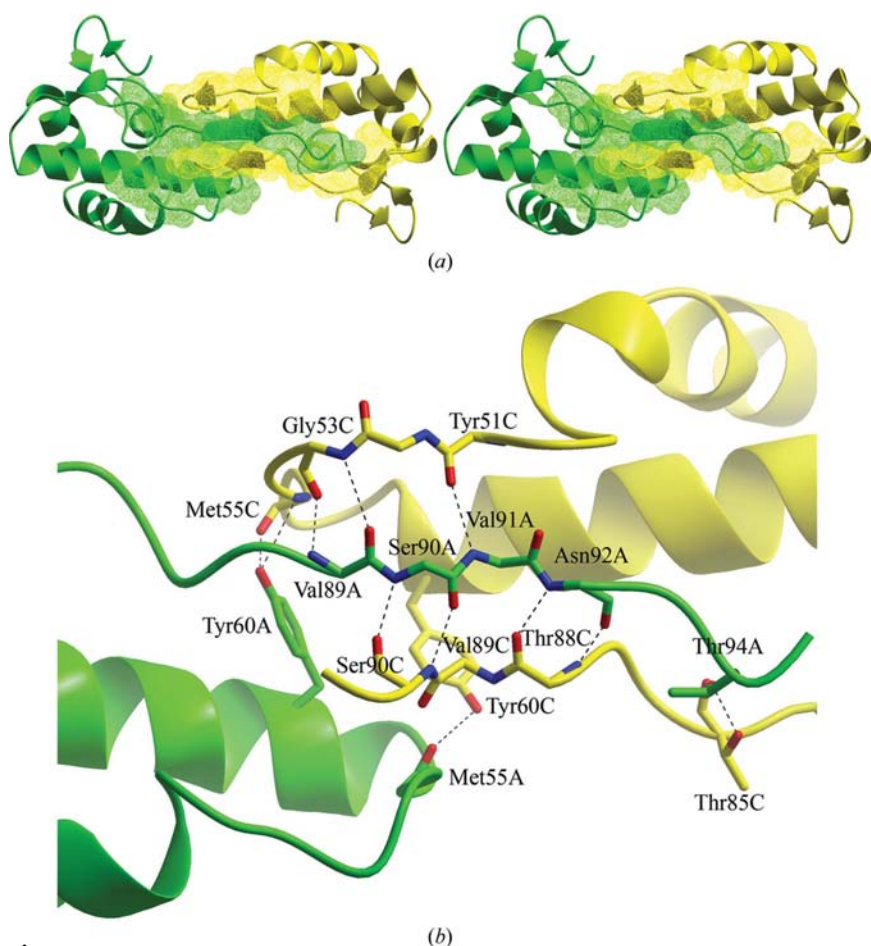
**Figure 3**

A ribbon view of FCoV nsp4ct is shown with the side chains of the hydrophobic residues important for protein folding depicted as van der Waals spheres. The residues are divided into three groups, namely the mainly aliphatic group (Phe19, Ile21, Leu29, Ile38, Leu68 and Leu72, yellow), aromatic group 1 (Phe11, Tyr41 and Tyr65, red) and aromatic group 2 (Tyr26 and Tyr75, green).

most distantly related virus and its sequence identity in all possible combinations with the other viruses is around 35%. This is consistent with previously published phylogenetic analyses of coronaviruses (Gorbalenya *et al.*, 2004). The sequence alignment shows a high level of conservation of the nsp4ct domain, with 17 of around 100 residues being identical between all five subgroups. Most of the aromatic amino acids of coronavirus nsp4ct are highly conserved. This includes residues Phe19, Tyr41, Tyr50, Tyr60 and Tyr84, which are fully conserved, and Phe11 (Tyr in IBV), Tyr26 (Phe in IBV), Phe45 (Tyr in four other coronaviruses) and Tyr51 and Tyr75 (both Phe in MHV and SARS-CoV). Interestingly, Phe45, Tyr50, Tyr51, Tyr60 and Tyr84 are part of the FCoV nsp4ct dimer interface and the fully conserved Tyr60 forms a side-chain ( $O^{\eta}$ ) hydrogen-bond interaction with the main-chain carbonyl of Met55 from the second monomer. The two fully conserved C-terminal residues (Leu95 and Gln96) are part of the

recognition site for the coronavirus M<sup>Pro</sup> (Hegyi & Ziebuhr, 2002).

There are four clusters of highly conserved residues. The first is between residues 9 and 19 and includes residues in helix  $\alpha 1$  and strand  $\beta 3$ . The second comprises residues 45–53 that belong to helix  $\alpha 3$  and part of loop  $L_{\alpha 3-\alpha 4}$ . Interestingly, the five independent chains of the FCoV nsp4ct structure differ most profoundly in this region, suggesting that it is highly flexible. In the cases of molecules *B* and *E* it was disordered and there was no electron density visible for residues 50–52 and 50–55, respectively. In molecules *A*, *C* and *D* this region could be placed into electron density and is involved in dimer formation. High sequence conservation of this cluster and its structural flexibility suggests that it may play an important role in the nsp4ct domain function. Residues 60–71 that belong to helix  $\alpha 4$  form the third highly conserved cluster and the fourth cluster consists of the C-terminal residues 81–96. This last cluster contains the highly conserved Tyr84, Pro86 and Pro87 which form the YxPP motif, which is the inverse of the consensus PPxY sequence recognized by the class I WW domains (Linn *et al.*, 1997). Di Leva *et al.* (2006) showed that the class I WW domain does not require a peptide with a consensus sequence and can also bind an inverted peptide sequence. The only condition is the presence of the polyproline II (PPII) conformation, which is observed in the case of FCoV nsp4ct. This suggests that region 84–87 is a reasonable candidate for protein–protein interactions. *PROSITE* (<http://www.expasy.org/prosite/>) analysis of all FCoV nsps did not identify any possible WW domains, suggesting that the YxPP motif interacting partner is a host protein. Furthermore, localization of the Pro–Pro motif may protect the extended unstructured C-terminus from proteolytic cleavage by host enzymes (Vanhoof *et al.*, 1995).



**Figure 4**  
 (a) The dimer interface between molecules *A* and *C* is shown in cartoon representation as a stereo pair. Molecule *A* is shown in green and molecule *C* is shown in yellow. The surfaces of the monomers at the interface are shown in mesh representation. (b) Hydrogen bonds at the dimer interface Val89 N...Gly53 O, Val89 O...Gly53 N, Ser90 N...Ser90 O, Ser90 O...Ser90 N, Val91 N...Tyr51 O, Asn92 N...Tyr88 O, Asn92 O...Tyr88 N, Met55 O...Tyr60  $O^{\eta}$ , Tyr60  $O^{\eta}$ ...Met55 O, Tyr60  $O^{\eta}$ ...Met55 N, Thr94  $O^{\gamma}$ ...Thr85 O and Thr94  $O^{\gamma}$ ...Thr85  $O^{\gamma}$  are shown as dashed lines. Molecule *A* is shown in green, molecule *C* is shown in yellow, O atoms are shown in red and N atoms are shown in blue. For clarity, the arrows (cartoon representation) of the strands forming the dimer interface are not shown in this panel.

#### 4. Conclusions

The high conservation of the C-terminal domain of nsp4 suggests not only that it plays a ubiquitous role in the coronavirus life cycle, but also that nsp4 proteins from different subgroups are structurally similar and have similar modes of operation. In this context, it is a surprising finding that deletion of the nsp4ct of MHV (using a reverse genetics system) was reported to be tolerated by the virus (Clementz *et al.*, 2008; Sparks *et al.*, 2007), with the resulting mutant displaying a modestly attenuated phenotype. Thus, although a similar mutant has not been generated for FCoV or any other coronavirus, the conservation of the

nsp4ct domain outlined above would suggest that it is not absolutely required for coronavirus RNA synthesis and/or RTC formation per se. This opens the possibility that, like some other recently characterized coronavirus enzyme functions (Eriksson *et al.*, 2008; Roth-Cross *et al.*, 2009), the nsp4ct domain might play a role in specific virus–host interactions of the type that are not easily uncovered in cell culture-based systems for virus propagation. Further functional studies are required in order to better understand the detailed role of nsp4 and to identify its partners and therefore its significance in the viral life cycle.

We thank Dr Stuart Siddell (University of Bristol, England) for kindly providing feline coronavirus and Linda Boomaars-van der Zanden for excellent technical assistance. This work was supported by the European VIZIER project (Comparative Structural Genomics of Viral Enzymes Involved in Replication) funded by the Sixth Framework Programme of the European Commission under reference LSHG-CT-2004-511960.

## References

- Anand, K., Ziebuhr, J., Wadhvani, P., Mesters, J. R. & Hilgenfeld, R. (2003). *Science*, **300**, 1763–1767.
- Bartlam, M., Yang, H. & Rao, Z. (2005). *Curr. Opin. Struct. Biol.* **15**, 664–672.
- Bhardwaj, K., Guarino, L. & Kao, C. C. (2004). *J. Virol.* **78**, 12218–12224.
- Brünger, A. T. (1992). *Nature (London)*, **355**, 472–475.
- Chen, Y., Cai, H., Pan, J., Xiang, N., Tien, P., Ahola, T. & Guo, D. (2009). *Proc. Natl Acad. Sci. USA*, **106**, 3484–3489.
- Cheng, A., Zhang, W., Xie, Y., Jiang, W., Arnold, E., Sarafianos, S. G. & Ding, J. (2005). *Virology*, **335**, 165–176.
- Clementz, M. A., Kanjanahaluethai, A., O'Brien, T. E. & Baker, S. C. (2008). *Virology*, **375**, 118–129.
- Collaborative Computational Project, Number 4 (1994). *Acta Cryst. D* **50**, 760–763.
- Cornvik, T., Dahlroth, S. L., Magnusdottir, A., Flodin, S., Engvall, B., Liew, V., Ekberg, M. & Nordlund, P. (2006). *Proteins*, **65**, 266–273.
- Cornvik, T., Dahlroth, S. L., Magnusdottir, A., Herman, M. D., Knaust, R., Ekberg, M. & Nordlund, P. (2005). *Nature Methods*, **2**, 507–509.
- Davis, I. W., Leaver-Fay, A., Chen, V. B., Block, J. N., Kapral, G. J., Wang, X., Murray, L. W., Arendall, W. B. III, Snoeyink, J., Richardson, J. S. & Richardson, D. C. (2007). *Nucleic Acids Res.* **35**, 375–383.
- Di Leva, F., D'Adamo, P., Cubellis, M. V., D'Eustacchio, A., Errichiello, M., Saulino, C., Auletta, G., Giannini, P., Donaudy, F., Ciccodicola, A., Gasparini, P., Franze, A. & Marciano, E. (2006). *Audiol. Neurootol.* **11**, 157–164.
- Dye, C. & Siddell, S. G. (2005). *J. Gen. Virol.* **86**, 2249–2253.
- Emsley, P. & Cowtan, K. (2004). *Acta Cryst. D* **60**, 2126–2132.
- Eriksson, K. K., Cervantes-Barragan, L., Ludewig, B. & Thiel, V. (2008). *J. Virol.* **82**, 12325–12334.
- Evans, P. R. (1993). *Proceedings of the CCP4 Study Weekend. Data Collection and Processing*, edited by L. Sawyer, N. Isaacs & S. Bailey, pp. 114–122. Warrington: Daresbury Laboratory.
- Gorbalenya, A. E., Enjuanes, L., Ziebuhr, J. & Snijder, E. J. (2006). *Virus Res.* **117**, 17–37.
- Gorbalenya, A. E., Koonin, E. V., Donchenko, A. P. & Blinov, V. M. (1989). *Nucleic Acids Res.* **17**, 4847–4861.
- Gorbalenya, A. E., Snijder, E. J. & Spaan, W. J. (2004). *J. Virol.* **78**, 7863–7866.
- Gosert, R., Kanjanahaluethai, A., Egger, D., Bienz, K. & Baker, S. C. (2002). *J. Virol.* **76**, 3697–3708.
- Hao, Q. (2004). *J. Appl. Cryst.* **37**, 498–499.
- Harcourt, B. H., Jukneliene, D., Kanjanahaluethai, A., Bechill, J., Severson, K. M., Smith, C. M., Rota, P. A. & Baker, S. C. (2004). *J. Virol.* **78**, 13600–13612.
- Hegyi, A. & Ziebuhr, J. (2002). *J. Gen. Virol.* **83**, 595–599.
- Imbert, I., Snijder, E. J., Dimitrova, M., Guillemot, J. C., Lecine, P. & Canard, B. (2008). *Virus Res.* **133**, 136–148.
- Ivanov, K. A., Hertzog, T., Rozanov, M., Bayer, S., Thiel, V., Gorbalenya, A. E. & Ziebuhr, J. (2004). *Proc. Natl Acad. Sci. USA*, **101**, 12694–12699.
- Ivanov, K. A., Thiel, V., Dobbe, J. C., van der Meer, Y., Snijder, E. J. & Ziebuhr, J. (2004). *J. Virol.* **78**, 5619–5632.
- Kabsch, W. (1988). *J. Appl. Cryst.* **21**, 916–924.
- Knoops, K., Kikkert, M., Worm, S. H., Zevenhoven-Dobbe, J. C., van der Meer, Y., Koster, A. J., Mommaas, A. M. & Snijder, E. J. (2008). *PLoS Biol.* **6**, e226.
- Krissinel, E. & Henrick, K. (2007). *J. Mol. Biol.* **372**, 774–797.
- Lai, M. M. C. & Holmes, K. V. (2001). *Fields Virology*, 4th ed., edited by D. M. Knipe & P. M. Howley, pp. 1163–1185. Philadelphia: Lippincott, Williams & Wilkins.
- Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J. & Higgins, D. G. (2007). *Bioinformatics*, **23**, 2947–2948.
- Linn, H., Ermekova, K. S., Rentschler, S., Sparks, A. B., Kay, B. K. & Sudol, M. (1997). *Biol. Chem.* **378**, 531–537.
- Matthews, B. W. (1968). *J. Mol. Biol.* **33**, 491–497.
- Mesters, J. R., Tan, J. & Hilgenfeld, R. (2006). *Curr. Opin. Struct. Biol.* **16**, 776–786.
- Morris, R. J., Zwart, P. H., Cohen, S., Fernandez, F. J., Kakaris, M., Kirillova, O., Vonnrhein, C., Perrakis, A. & Lamzin, V. S. (2004). *J. Synchrotron Rad.* **11**, 56–59.
- Mueller-Dieckmann, J. (2006). *Acta Cryst. D* **62**, 1446–1452.
- Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst. D* **53**, 240–255.
- Oostra, M., te Lintelo, E. G., Deijs, M., Verheije, M. H., Rottier, P. J. & de Haan, C. A. (2007). *J. Virol.* **81**, 12323–12336.
- Panjikar, S., Parthasarathy, V., Lamzin, V. S., Weiss, M. S. & Tucker, P. A. (2005). *Acta Cryst. D* **61**, 449–457.
- Pannu, N. S., McCoy, A. J. & Read, R. J. (2003). *Acta Cryst. D* **59**, 1801–1808.
- Pannu, N. S. & Read, R. J. (2004). *Acta Cryst. D* **60**, 22–27.
- Pedersen, N. C. (1995). *Feline Pract.* **23**, 13.
- Perrakis, A., Morris, R. & Lamzin, V. S. (1999). *Nature Struct. Biol.* **6**, 458–463.
- Poland, A. M., Vennema, H., Foley, J. E. & Pedersen, N. C. (1996). *J. Clin. Microbiol.* **34**, 3180–3184.
- Ratia, K., Saikatendu, K. S., Santarsiero, B. D., Barretto, N., Baker, S. C., Stevens, R. C. & Mesecar, A. D. (2006). *Proc. Natl Acad. Sci. USA*, **103**, 5717–5722.
- Read, R. J. (1986). *Acta Cryst. A* **42**, 140–149.
- Roth-Cross, J. K., Stokes, H., Chang, G., Chua, M. M., Thiel, V., Weiss, S. R., Gorbalenya, A. E. & Siddell, S. G. (2009). *J. Virol.* **83**, 3743–3753.
- Schomaker, V. & Trueblood, K. N. (1968). *Acta Cryst. B* **24**, 63–76.
- Seybert, A., Hegyi, A., Siddell, S. G. & Ziebuhr, J. (2000). *RNA*, **6**, 1056–1068.
- Sheldrick, G. M. (2008). *Acta Cryst. A* **64**, 112–122.
- Shi, S. T., Schiller, J. J., Kanjanahaluethai, A., Baker, S. C., Oh, J. W. & Lai, M. M. (1999). *J. Virol.* **73**, 5957–5969.
- Snijder, E. J., Bredendiek, P. J., Dobbe, J. C., Thiel, V., Ziebuhr, J., Poon, L. L., Guan, Y., Rozanov, M., Spaan, W. J. & Gorbalenya, A. E. (2003). *J. Mol. Biol.* **331**, 991–1004.
- Snijder, E. J., van der Meer, Y., Zevenhoven-Dobbe, J., Onderwater,

- J. J., van der Meulen, J., Koerten, H. K. & Mommaas, A. M. (2006). *J. Virol.* **80**, 5927–5940.
- Sparks, J. S., Lu, X. & Denison, M. R. (2007). *J. Virol.* **81**, 12554–12563.
- Stertz, S., Reichelt, M., Spiegel, M., Kuri, T., Martinez-Sobrido, L., Garcia-Sastre, A., Weber, F. & Kochs, G. (2007). *Virology*, **361**, 304–315.
- Terwilliger, T. C. (2000). *Acta Cryst.* **D56**, 965–972.
- Vagin, A. & Teplyakov, A. (1997). *J. Appl. Cryst.* **30**, 1022–1025.
- Vanhoof, G., Goossens, F., De Meester, I., Hendriks, D. & Scharpe, S. (1995). *FASEB J.* **9**, 736–744.
- Vennema, H., Poland, A., Foley, J. & Pedersen, N. C. (1998). *Virology*, **243**, 150–157.
- Woo, P. C., Lau, S. K., Lam, C. S., Lai, K. K., Huang, Y., Lee, P., Luk, G. S., Dyrting, K. C., Chan, K. H. & Yuen, K. Y. (2009). *J. Virol.* **83**, 908–917.
- Ziebuhr, J., Snijder, E. J. & Gorbalenya, A. E. (2000). *J. Gen. Virol.* **81**, 853–879.