West Nile Virus Core Protein: Tetramer Structure and Ribbon Formation

Terje Dokland,1,2,* Martin Walsh,3 Jason M. Mackenzie,4 Alexander A. Khromykh,4 Kim-Huey Ee,1 and Sifang Wang1
1Institute of Molecular and Cell Biology Singapore
2Department of Microbiology University of Alabama at Birmingham Birmingham, Alabama 35294
3Medical Research Council France European Synchrotron Research Facility Grenoble France
4The Sir Albert Sakzewski Virus Research Centre Royal Children’s Hospital and Clinical Medical Virology Centre University of Queensland Brisbane Australia

Summary

We have determined the crystal structure of the core (C) protein from the Kunjin subtype of West Nile virus (WNV), closely related to the NY99 strain of WNV, currently a major health threat in the U.S. WNV is a member of the Flaviviridae family of enveloped RNA viruses that contains many important human pathogens. The C protein is associated with the RNA genome and forms the internal core which is surrounded by the envelope in the virion. The C protein structure contains four α helices and forms dimers that are organized into tetramers. The tetramers form extended filamentous ribbons resembling the stacked α helices seen in HEAT protein structures.

Introduction

West Nile virus (WNV) is an emerging, mosquito-borne human pathogen that first appeared in the U.S. in 1999 (Brinton, 2002) and currently constitutes a serious health threat, with 264 registered deaths and 9858 reported cases in 2003 (CDC, 2004). Sporadic outbreaks have also occurred in Europe, and the virus has been isolated from birds in the UK (Buckley et al., 2003). Typical symptoms of WNV infection include fever, rashes, and myalgia; however, occasionally more severe and potentially fatal symptoms develop, including acute encephalitis and fulminant hepatitis (Petersen et al., 2003). WNV are divided into two lineages: lineage 1 includes the New York strain (NY99) responsible for the recent outbreaks in the U.S., as well as the closely related Kunjin virus, which was recently classified as an Australian subtype of WNV (Lanciotti et al., 1999; Scherret et al., 2001). Unlike the highly pathogenic NY99 strain, humans infected with Kunjin virus do not develop the more severe symptoms (Hall et al., 2003). Lineage 2 consists of several African strains, which are apparently not involved in outbreaks of the disease (Lanciotti et al., 1999). There is currently no vaccine available, but inoculation with a cDNA for the whole genome of an attenuated form of Kunjin virus could confer resistance to infection by the NY99 strain of WNV in mice (Hall et al., 2003).

WNV is a member of the Flaviviridae family, which includes many important human pathogens, like dengue, yellow fever, tick-borne encephalitis (TBE), and hepatitis C viruses (Strauss and Strauss, 2002). The flaviviruses have a positive-sense, single-stranded RNA genome, containing one large open reading frame, which is translated as a single polyprotein (Lindenbach and Rice, 2002). The 3433 amino acid polyprotein of WNV is cleaved by host and viral proteases to the three structural proteins, C (core), prM (membrane), and E (envelope), and seven nonstructural (NS) proteins (Brinton, 2002). In the virion, the genome is contained within a spherical or isometric nucleocapsid or core made from the C protein. The 123 amino acid C precursor is targeted to the membrane by a carboxy-terminal hydrophobic sequence and is cleaved to its mature 105 residue form (C) protein from the Kunjin subtype of West Nile virus (WNV), closely related to the NY99 strain of WNV, currently a major health threat in the U.S. WNV is a member of the Flaviviridae family of enveloped RNA viruses that contains many important human pathogens. The C protein is associated with the RNA genome and forms the internal core which is surrounded by the envelope in the virion. The C protein structure contains four α helices and forms dimers that are organized into tetramers. The tetramers form extended filamentous ribbons resembling the stacked α helices seen in HEAT protein structures.

Cryo-electron microscopy (EM) reconstruction of WNV (Mukhopadhyay et al., 2003) showed an outer envelope layer consisting of 180 copies of E protein, very similar to that of dengue virus (Kuhn et al., 2002; Zhang et al., 2003). Consequently, the WNV E protein is assumed to have a similar structure to the E protein of dengue virus and TBE (Modis et al., 2003; Rey et al., 1995). The EM reconstructions of WNV and dengue virus did not show any structure for the inner core, however, presumably due to its being disordered or having different symmetry from the envelope.

There are only four amino acid differences (out of 105) in the mature C protein between Kunjin virus and the NY99 strain of WNV and seven differences between Kunjin virus and a Nigerian strain, a member of WNV lineage 2 (Figure 1). Dengue virus C protein shares about 57% identity with WNV. The WNV C protein contains a large number of positively charged residues, which are distributed throughout the protein, unlike the core proteins of togaviruses and arteriviruses, which have distinct, positively charged N-terminal regions (Choi et al., 1991; Doan and Dokland, 2003). The N-terminal and C-terminal charged regions have both been implicated in RNA binding (Khromykh and Westaway, 1996). The dengue virus C protein was recently solved by NMR methods (Ma et al., 2004). It forms a dimer in solution and contains four α helices. The N-terminal 20 or so residues were disordered, similar to the RNA binding regions of togaviruses and arteriviruses. Here, we report the crystal structure...
of the C protein from the Kunjin subtype of WNV. This is the first crystal structure of a flavivirus core protein. The C protein forms dimers, similar to those of dengue virus C, which are organized into tetramers with highly positively charged surfaces. The α helices in the tetramers stack up to form long filamentous ribbons in the crystal.

Results and Discussion

The first 103 amino acids of the C protein of Kunjin virus (strain MRM61C) were cloned with an N-terminal 6xHis tag and purified by affinity and ion exchange chromatography. Crystals of the full-length C protein were unstable and disordered and did not diffract X-rays to high resolution. Assuming that the crystal disorder was caused by flexible regions in the protein, we treated the full-length protein with trypsin, resulting in a stable fragment starting at Val23 and ending at Arg98 (Figure 1). This fragment yielded stable crystals which diffracted synchrotron radiation anisotropically to about 2.8 Å resolution. The crystals belong to space group I41 with a = 85.7 Å and c = 214.4 Å, with eight protein monomers in the asymmetric unit and V_M = 2.8. The structure was solved to 3.2 Å resolution by MAD on a Se-Met derivative, giving a final R_cryt = 0.25 and R_free = 0.31 (Table 1) and has been deposited in the Protein Data Bank with accession code 1SFK (Figures 2A and 2B).

The C protein structure consists of four α helices interspersed by short loops (Figures 1, 2B, and 3A). The first 103 amino acids of the C protein of Kunjin virus (strain MRM61C) were cloned with an N-terminal 6xHis tag and purified by affinity and ion exchange chromatography. Crystals of the full-length C protein were unstable and disordered and did not diffract X-rays to high resolution. Assuming that the crystal disorder was caused by flexible regions in the protein, we treated the full-length protein with trypsin, resulting in a stable fragment starting at Val23 and ending at Arg98 (Figure 1). This fragment yielded stable crystals which diffracted synchrotron radiation anisotropically to about 2.8 Å resolution. The crystals belong to space group I41 with a = 85.7 Å and c = 214.4 Å, with eight protein monomers in the asymmetric unit and V_M = 2.8. The structure was solved to 3.2 Å resolution by MAD on a Se-Met derivative, giving a final R_cryt = 0.25 and R_free = 0.31 (Table 1) and has been deposited in the Protein Data Bank with accession code 1SFK (Figures 2A and 2B).

The C protein structure consists of four α helices interspersed by short loops (Figures 1, 2B, and 3A). The protein forms a tight 2-fold symmetric dimer, in which the helices α1, α2, and α4 form three distinct “layers” and α3 forms a short, connecting helix flanking the dimer (Figure 3A). A DALI search for structural homologs (Holm and Sander, 1995) revealed a topological resemblance with the cyclin A-like domain, like that of archael transcription factor B (TFB; PDB code 1AIS; Z = 4.0) (Kos et al., 1997), and with the HEAT repeat domain of human

Table 1. Crystallographic Data

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Kunc29 Native</th>
<th>SF83 Se-Met</th>
</tr>
</thead>
<tbody>
<tr>
<td>X-ray source</td>
<td>ESRF BM14</td>
<td>ESRF BM14</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wavelength (Å)</td>
<td>0.8856</td>
<td>0.97956</td>
</tr>
<tr>
<td>Oscillation angle (°)</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Number of frames</td>
<td>139</td>
<td>200</td>
</tr>
<tr>
<td>Resolution range (Å)</td>
<td>30.0–3.2</td>
<td>40.0–3.7</td>
</tr>
<tr>
<td>Total number of reflections</td>
<td>72,872</td>
<td>70,802</td>
</tr>
<tr>
<td>Number of unique reflections</td>
<td>12,515</td>
<td>8,246</td>
</tr>
<tr>
<td>Completeness (%)</td>
<td>99.4 (95.3)</td>
<td>100</td>
</tr>
<tr>
<td>Redundancy</td>
<td>5.8 (4.8)</td>
<td>8.6 (8.6)</td>
</tr>
<tr>
<td>Rmerge^a</td>
<td>0.075 (0.32)</td>
<td>0.10 (0.45)</td>
</tr>
<tr>
<td>R cryst^a</td>
<td>0.251</td>
<td>0.087 (0.46)</td>
</tr>
<tr>
<td>R free^a</td>
<td>0.311</td>
<td>0.094 (0.51)</td>
</tr>
<tr>
<td>V (Å³)</td>
<td>21.2 (3.0)</td>
<td>26.7 (4.2)</td>
</tr>
</tbody>
</table>

^a Values in parentheses are for the high-resolution shell.

Results and Discussion

The first 103 amino acids of the C protein of Kunjin virus (strain MRM61C) were cloned with an N-terminal 6xHis tag and purified by affinity and ion exchange chromatography. Crystals of the full-length C protein were unstable and disordered and did not diffract X-rays to high resolution. Assuming that the crystal disorder was caused by flexible regions in the protein, we treated the full-length protein with trypsin, resulting in a stable fragment starting at Val23 and ending at Arg98 (Figure 1). This fragment yielded stable crystals which diffracted synchrotron radiation anisotropically to about 2.8 Å resolution. The crystals belong to space group I41 with a = 85.7 Å and c = 214.4 Å, with eight protein monomers in the asymmetric unit and V_M = 2.8. The structure was solved to 3.2 Å resolution by MAD on a Se-Met derivative, giving a final R_cryt = 0.25 and R_free = 0.31 (Table 1) and has been deposited in the Protein Data Bank with accession code 1SFK (Figures 2A and 2B).

The C protein structure consists of four α helices interspersed by short loops (Figures 1, 2B, and 3A). The protein forms a tight 2-fold symmetric dimer, in which the helices α1, α2, and α4 form three distinct “layers” and α3 forms a short, connecting helix flanking the dimer (Figure 3A). A DALI search for structural homologs (Holm and Sander, 1995) revealed a topological resemblance with the cyclin A-like domain, like that of archael transcription factor B (TFB; PDB code 1AIS; Z = 4.0) (Kos et al., 1997), and with the HEAT repeat domain of human
West Nile Virus Core Protein

1159

Figure 2. Stereoviews of the WNV C Protein Structure

(A) Detail of the weighted 2F₀ − F₁ electron density map, showing part of helix α4 from Asn72 to Ile94 (Esnouf, 1999).

(B) C backbone trace of CD dimer. C, blue; D, red.

(C) Superposition of subunit A (blue) with residues B1108 to B1206 of Pyrococcus woesei TFB (PDB code 1AIS; green), aligned using the helical segments identified by the DALI search (Holm and Sander, 1995).

(D) Superposition of the C protein CD dimer (purple) with the solution structure dimer of dengue C protein (1R6R) (green). The least mean square fit was calculated using residues 54–96.

elF4Gii (1HU3; Z = 3.9) (Marcotrigiano et al., 2001) (Figure 2C). Like the WNV C protein, both of these proteins are involved in nucleic acid interactions; however, unlike WNV C, they are monomeric. A comparison with the dengue virus C protein dimer (1R6R) (Ma et al., 2004) shows that while helices α2 to α4 superimpose relatively well, with an rmsd of 3.6 Å for the C atoms of residues 44–96, α1 is in a completely different orientation (Figure 2D).

In the crystal, two dimers form a tetramer with 222 symmetry, with a cluster of four NCS-related α4 helices forming a tunnel-like cavity (Figure 3B). Two such tetramers, rotated by 58° around an axis parallel to the crystal a axis, comprise the asymmetric unit of the crystal (Figure 3C). In the crystal, symmetry-related tetramers form long, helical ribbons that spiral around the 4-fold screw axes and extend the length of the crystal c axis (Figure 3C). Two such ribbons together form a hollow tubule at
Figure 3. Ribbon Diagrams of the WNV C Protein

The diagrams have been colored as follows: A, red; B, blue; C, green; D, yellow; E, cyan; F, magenta; G, orange; H, teal.

(A) Ribbon diagram of CD dimer (Esnouf, 1999). Secondary structure elements are indicated.

(B) Ribbon diagram of ABDC tetramer looking down the tunnel. N- and C-terminal residues are indicated.

(C) Arrangement of tetramers in the crystal, viewed down the b axis. The unit cell, the crystal a and c axes and the 4-fold screw axis are indicated. The ABCD and EFGH tetramers in the asymmetric unit are shown in color, while the symmetry-related ABCD tetramers that form one half of the tubule are shown in gray. The symmetry-related EFGH tetramers are omitted for clarity.

Each of the two 4-fold screw axes in the crystal. This leaves a large space between adjacent tubules, which is reflected in the relatively high solvent content (57%). Interestingly, the resulting repeated stacking of α helices in the tubules resembles the long, filamentous structures commonly formed by HEAT and ARM repeat proteins, multimeric structures typically involved in protein-protein interactions (Andrade et al., 2001).

The AB dimer encloses a hydrophobic pocket formed by Leu29, Leu36, Phe45, Leu49, Phe52, and Phe53 from the α1 and α2 helices (Figure 4A). The symmetry related C’ D’ dimer approaches the pocket, so that the α1 helices from A, C’, and D’ form a trimeric bundle. This would cause helix α1 from B and C’ to overlap; consequently, density for α1 in subunit B is missing, suggesting that this helix undergoes an order-disorder transition upon crystallization. Such order-disorder transitions are often of functional importance and may suggest a role in the conformational switching that is required during core assembly (Dokland, 2000). The density is very poor for the overlapping α helices in subunit C, perhaps due to two alternative orientations coexisting in the crystal. The GH dimer forms similar interactions with E’ F’; in this case α1 is missing from H where it would overlap with F’. The different orientation of α1 in the dengue C protein (Ma et al., 2004) also suggests that this helix is flexible, perhaps reflecting a functional role in assembly or RNA binding (Figure 2D). The hydrophobic pocket contains additional density that was not interpretable at 3.2 Å resolution, but could conceivably represent part of the
Figure 4. Interactions and Surface Potential of the WNV Core Protein Tetramer

(A) View of the interface between the AB dimer (red, blue) and the symmetry-related C′/D′ dimer (green, yellow) (Esnouf, 1999). The position that the missing helix α1 in the B subunit would have had where it overlaps with subunit C is shown in gray. The Tyr and Leu residues that comprise the hydrophobic pocket in AB are indicated, as is the density enclosed by the pocket in AB and C′/D′.

(B) Electrostatic potential surface of C protein tetramer, viewed in the same orientation as Figure 3B.

missing density for the B and H subunits, or perhaps a polyethylene glycol molecule from the crystallization solution. There is a cation, probably Ca\(^{2+}\), sitting on the crystallographic 2-fold axis between 2-fold related Thr43 residues and an anion, possibly Cl\(^{−}\), held between 2-fold related Arg31 residues, about 3.5 Å from the cation, mediating the interactions between tubules in adjacent asymmetric units. The two ribbons within the tubule are held together by interactions between Arg69 residues in the A–G and B–E chains.

The WNV C protein dimer presumably represents the building block for core assembly, which also involves interactions with the viral RNA. Both the N-terminal (1–32) and C-terminal (83–103) sequences of the Kunjin virus C protein are involved in RNA binding (Khromykh and Westaway, 1996). The N-terminal region corresponds to the disordered region (residues 1–22) that was cleaved off by the trypsin treatment, and the C-terminal region corresponds to helix α4, which forms an 11 Å wide tunnel in the tetramer (Figure 3B). The tunnel contains a number of positively charged residues, consistent with a role in RNA binding (Figure 4B). The N- and C-terminal RNA binding regions are located at opposite ends of the dimer. While a conformational switch in the N-terminal region, perhaps reflected in the order-disorder transition in helix α1, could bring the two termini together on one side, the organization would appear to be very different from that of the togaviruses or the arteriviruses, which have a clear “inside” or RNA binding face (Choi et al., 1991; Doan and Dokland, 2003). It is possible that the RNA instead associates on the surface of the C tetramer, in a way similar to that proposed for the tetrameric Borna virus nucleoprotein (Rudolph et al., 2003). Indeed, the C tetramer has a strongly positively charged surface (Figure 4B). In this model, the tetramers would form the building blocks for capsid assembly. The order-disorder transition in helix α1 (Figure 4A) might thus reflect tetramer-tetramer interactions in the nucleocapsid. However, the lack of a structured core in EM reconstructions of dengue or WNV could mean that there is no ordered shell, and that the RNA is instead packed together with protein in a nonspecific manner (Kuhn et al., 2002; Mukhopadhyay et al., 2003; Zhang et al., 2003).

A hydrophobic region, corresponding to residues 45–65 in WNV, was proposed to localize the C protein to the membrane in dengue and TBE (Kofler et al., 2002; Markoff et al., 1997). In the C structure, this region forms the hydrophobic pocket that is partially protected by helix α1. Movement of α1 might render the hydrophobic region in a position able to interact with the membrane. Part of the hydrophobic region of TBE C (residues 28–43), which corresponds to parts of helices α1 and α2 in WNV, could be removed while still retaining some viability (Kofler et al., 2002). This suggests that the RNA binding properties of the C-terminal domain are more important than the specific fold of the protein, or that RNA binding is more unspecific in nature. Furthermore, viability was partially restored by mutations that increased
the hydrophobic nature of the remaining sequence, indicating that membrane localization is done through non-specific hydrophobic interactions (Kofler et al., 2003).

In vivo, C is localized to the ER membrane through prM, prior to cleavage of C-prM by the NS2B-NS3 protease (Lobigs and Lee, 2004; Mackenzie and Westaway, 2001). Incorporation of cores into virions occurs at the ER membrane and also involves nonstructural proteins NS2A and NS3. The EM structures of dengue and WNV (Kuhn et al., 2002; Mukhopadhyay et al., 2003; Zhang et al., 2003) did not reveal cytoplasmic extensions of the M and E proteins, so there may not be a direct interaction between C and prM/E. Virion budding is not dependent on C, and VLPs could be produced by expression of TBE prM and E (Ferlenghi et al., 2001), while expression of dengue C, prM, and E led to the formation of VLPs that lacked C (Sugrue et al., 1997).

Based on differences in genomic organization as well as homologies in nonstructural proteins, the Flaviviridae are considered to belong to an evolutionarily distinct superfamily from Togaviridae and the Nidovirales, which includes coronaviruses and arteriviruses (Strauss and Strauss, 2002). Indeed, the structures of the core proteins from these three superfamilies are different. The Sindbis virus (Togaviridae) core protein has a chymotrypsin-like β-barrel fold (Choi et al., 1991). The recently determined structure of the core (N) protein of the arterivirus PRRSV displayed a novel motif based on an anti-parallel β sheet flanked by α helices (Doan and Dokland, 2003). As already suspected from analysis of charge distribution and secondary structure predictions, the flavivirus core structure presented here has a different structure, based entirely on α-helical bundles. Yet, there are some similarities between the WNV C and PRRSV N structures: Both proteins are dimeric; the dimers have an antiparallel α helix pair and are flanked by connecting α helices; both structures form long filamentous ribbon-like or tubular structures that extend in the direction of the long crystal c axis. In WNV, these filaments are formed by stacking of α-helical domains (Figure 3C), similar to those formed by HEAT repeat proteins (Andrade et al., 2001). These similarities are probably based on similar propensities or requirements for forming multimeric shells with a range of different interactions, and may reflect subunit and RNA interactions in the virus.

**Experimental Procedures**

The 103 first amino acids of the C protein of Kunjin virus (strain MR861C) were cloned with an N-terminal 6-His tag in the vector pET16b (Novagen), expressed in E. coli strain BL21(DE3) (Novagen) and purified by affinity and ion exchange chromatography. The protein was treated with low concentrations of trypsin, resulting in a stable 10.5 kDa fragment. N-terminal protein sequencing and mass spectrometry confirmed that the fragment started from Val23 and ended at Arg98. This 76 residue fragment was crystallized by vapor diffusion against 0.5%-2% PEG 3350 at pH 10.5 and yielded stable crystals of up to 0.2 mm length, which diffracted synchrotron radiation anisotropically to about 2.8 Å resolution. Se-Met derivative crystals were made by the same procedure by growing the cells in minimal media supplemented with amino acids Lys, Phe, Thr, Ile, Leu, Val, and Se-Met and 0.4% glucose (Doublie, 1997). Native and Se-Met derivative data were collected at the European Synchrotron Radiation Facility beamline BM14 using a marCCD detector (Mar USA, Inc.) and processed using HKL2000 (Owinowski and Minor, 1997) to a resolution of 2.8 Å for the native data and 3.6 Å for the Se-Met data (Table 1).

The crystals belonged to space group I4, with a = 85.7 Å and c = 214.4 Å, corresponding to a total of eight protein monomers in the asymmetric unit with V₉ = 2.8 Å³. Structure determination by Se-Met MAD methods was done using SOLVE and RESOLVE (Terwilliger, 2002, 2003). Starting from the initial RESOLVE trace, the noncrystallographic symmetry (NCS) that resulted in the generation of a total of four dimers in the asymmetric unit was identified and used for 8-fold NCS averaging and map improvement with DM (Cowtan and Main, 1998), as well as phase extension to 3.2 Å resolution using the native data. All eight subunits were built manually into the 2Fᵋ - F map using the program O (Jones et al., 1991), and the sequence could be unambiguously assigned in residues 24–96 in most of the chains (Figure 2A). Density for residues 24–39 was missing in two chains, and was poor for another two chains. Refinement with REFMAC5 (Murshudov et al., 1997) yielded a final R cryst = 0.25 and R free = 0.31 (Table 1).

Received: March 9, 2004
Revised: April 5, 2004
Accepted: April 6, 2004
Published: July 13, 2004

**References**


Accession Numbers
The structure has been deposited in the Protein Data Bank with accession code 1SFK.