Acta Crystallographica Section D
Biological Crystallography
ISSN 1399-0047

The structure of the C-terminal domain of the *Zaire ebolavirus* nucleoprotein

Paulina J. Dziubańska,[a,b]‡ Urszula Derewenda,[a] Jeffrey F. Ellena,[c] Daniel A. Engel[b] and Zygmunt S. Derewenda[a]*

[a]Department of Molecular Physiology and Biological Physics, University of Virginia School of Medicine, Charlottesville, VA 22908-0736, USA, [b]Department of Microbiology, Immunology and Cancer Biology, University of Virginia School of Medicine, Charlottesville, VA 22908-0736, USA, and [c]Department of Chemistry, University of Virginia, Charlottesville, VA 22904-4319, USA

‡ On leave from the Faculty of Chemistry, Wroclaw University of Technology, Wroclaw, Poland.

Correspondence e-mail: zsd4n@virginia.edu

Received 9 May 2014
Accepted 22 June 2014

PDB references: C-terminal domain of *Zaire ebolavirus* nucleoprotein, 4qb0; 4qaz

Ebolavirus (EBOV) causes severe hemorrhagic fever with a mortality rate of up to 90%. EBOV is a member of the order *Mononegavirales* and, like other viruses in this taxonomic group, contains a negative-sense single-stranded (ss) RNA. The EBOV ssRNA encodes seven distinct proteins. One of them, the nucleoprotein (NP), is the most abundant viral protein in the infected cell and within the viral nucleocapsid. Like other EBOV proteins, NP is multifunctional. It is tightly associated with the viral genome and is essential for viral transcription, RNA replication, genome packaging and nucleocapsid assembly prior to membrane encapsulation. NP is unusual among the *Mononegavirales* in that it contains two distinct regions, or putative domains, the C-terminal of which shows no homology to any known proteins and is purported to be a hub for protein–protein interactions within the nucleocapsid. The atomic structure of NP remains unknown. Here, the boundaries of the N- and C-terminal domains of NP from Zaire EBOV are defined, it is shown that they can be expressed as highly stable recombinant proteins in *Escherichia coli*, and the atomic structure of the C-terminal domain (residues 641–739) derived from analysis of two distinct crystal forms at 1.98 and 1.75 Å resolution is described. The structure reveals a novel tertiary fold that is distantly reminiscent of the β-grasp architecture.

1. Introduction

Viral hemorrhagic fevers (VHFs) constitute a group of severe illnesses in which the vascular system is damaged with accompanying internal bleeding, while regulatory functions of the body are critically impaired (Paessler & Walker, 2013). Several distinct families of viruses cause VHF, with varying disease severity. The most dangerous VHF, associated with a mortality as high as 90%, is caused by the filoviruses (*Filoviridae*), *e.g.* *Marburgvirus* (MARV) and *Ebolavirus* (EBOV) (Brauburger *et al.*, 2012; de Wit *et al.*, 2011). Although outbreaks of Ebola hemorrhagic fever, first identified in 1976, are sporadic and endemic to Africa, EBOV constitutes a grave global potential health threat. As this paper was being prepared for publication, the most challenging EBOV outbreak ever was reported by the World Health Organization in Guinea and Liberia, with possible cases in Sierra Leone, Mali and Ghana. This is the first outbreak of EBOV in West Africa. To date, more than 250 deaths have been confirmed. Owing to its pathogenicity, high mortality and human-to-human transmission, EBOV is considered to be a potential bioweapon and is classified as a Category A bioterrorism agent (Salvaggio & Baddley, 2004). Importantly, there are no approved vaccines or antiviral agents against

EBOV, while existing therapies for infected individuals have minimal effects.

Five strains of EBOV have been identified to date. These include four African strains, *i.e.* Tai Forest (also known as Ivory Coast), Sudan, Zaire and Bundibugyo, as well as the Reston strain from the Philippines (Kuhn *et al.*, 2013). There is a single known strain of the related MARV (Brauburger *et al.*, 2012), and the distantly related Lloviu virus (LLOV) is found in insectivorous bats in Spain (Negredo *et al.*, 2011). Reston EBOV and Lloviu virus do not appear to be pathogenic in humans. In contrast, the Sudan, Ivory Coast, Bundibugyo and Zaire EBOV strains, as well as MARV, have been associated with VHF outbreaks. Fatality rates for EBOV range from ∼40% for the Sudan and Bundibugyo strains to ∼90% for the Zaire strain (the rate for the Tai Forest strain is not known owing to its rarity).

There is intense interest in the molecular mechanisms of infectivity, replication, assembly and pathogenesis of EBOV, with the long-term objective of identifying suitable targets for drug discovery and the development of effective diagnostic tools. The ssRNA genome encodes seven proteins, most of which have multiple functions (Feldmann *et al.*, 1993). Two of them, the glycoprotein GP and the matrix protein VP40, are essential components of the viral envelope that surrounds the nucleocapsid (Beniac *et al.*, 2012). The nucleocapsid includes the viral negative-sense ssRNA complexed with five proteins, *i.e.* the nucleoprotein (NP), the structural proteins VP24, VP30 and VP35 and the viral polymerase (L). Cryo-EM and tomography allowed for the reconstruction of the EBOV nucleocapsid at 14–19 Å resolution (Beniac *et al.*, 2012; Bharat *et al.*, 2012; Booth *et al.*, 2013). Recent intensive efforts have resulted in structural (mainly crystallographic) characterization of five EBOV proteins: GP (Lee *et al.*, 2008), the matrix protein VP40 (Bornholdt *et al.*, 2013; Gomis-Rüth *et al.*, 2003; Dessen *et al.*, 2000), VP24 (Zhang *et al.*, 2012), VP30 (Hartlieb *et al.*, 2007) and VP35 (Leung, Shabman *et al.*, 2010; Leung, Prins *et al.*, 2010).

Proteins L and NP have so far eluded structural characterization. The latter plays a critical role in virus replication and maturation, and is the most abundant viral protein in infected cells and the viral nucleocapsid. NPs are found in all members of the order *Mononegavirales*, which groups together a number of important viruses that are highly pathogenic to humans, animals and plants, including *Filoviridae*, measles, mumps and rabies viruses, avian bornavirus and many others. The ssRNA in these viruses is packaged into a helical complex that includes multiple copies of NP. The architectures of the resulting NP–ssRNA complexes differ among the *Mononegavirales* families. Insights into the structure–function relationships underlying the physiological role of NPs from *Mononegavirales* have been made possible owing to crystallographic studies of the proteins from rabies virus and bornavirus (Albertini *et al.*, 2006; Rudolph *et al.*, 2003). Interestingly, the *Filoviridae* members appear to have unusual NPs characterized by a longer polypeptide chain than those of other *Mononegavirales*, with two distinct functional modules, with the N-terminal domain exhibiting the canonical ssRNA-packaging function (Noda *et al.*, 2010; Watanabe *et al.*, 2006). Recent data suggest that the C-terminal domain, with an amino-acid sequence that shows no homology to any other protein, may serve as a unique hub for protein–protein interactions in the nucleocapsid that are distinct from any other *Mononegavirales* (Beniac *et al.*, 2012). Moreover, recent data show that the C-terminal fragment of the EBOV NP is a major antigenic determinant, raising the possibility that it could be effective in virus detection and diagnostics (Sherwood & Hayhurst, 2013). Therefore, elucidation of the molecular architecture of the EBOV NP would be of considerable significance.

In this paper, we report the identification of the boundaries of the two globular domains in Zaire EBOV NP, their overexpression in *Escherichia coli* and the structure determination of two crystal forms of the C-terminal domain spanning residues 641–739.

2. Materials and methods

2.1. *In silico* sequence analyses

Secondary-structure and tertiary-structure predictions, as well as disorder predictions, were carried out using the *Jpred* (Cole *et al.*, 2008; Cuff *et al.*, 1998), *Phyre*2 (Kelley & Sternberg, 2009), *EMBOSS* (Rice *et al.*, 2000), *GlobPlot* (Linding *et al.*, 2003) and *DisMeta* (Huang *et al.*, 2014) servers. Amino-acid sequence conservation was analyzed using *Geneious* (http://www.geneious.com) and *ConSurf* (Ashkenazy *et al.*, 2010).

2.2. Preparation of recombinant proteins

2.2.1. General remarks. cDNA constructs coding for the 1–412 and 641–739 fragments of the *Zaire ebolavirus* (EBOV) nucleoprotein (NP) were synthesized commercially (GENEWIZ) using optimized codon frequencies for *E. coli*. The constructs were cloned into the His_6-MBP-Parallel1 vector (Sheffield *et al.*, 1999). Consequently, the proteins were expressed as fusion proteins with MBP and could be purified using affinity chromatography. BL21-CodonPlus (DE3)-RIPL *E. coli* cells (Stratagene) were used for expression. Cells were grown in different types of media (see below) supplemented with 100 µg ml^{-1} ampicillin and 34 µg ml^{-1} chloramphenicol and were induced with 0.5 m*M* IPTG. All purification steps were carried out at 4°C. Protein concentrations were determined spectrophotometrically based on calculated molar absorption coefficients at 280 nm.

2.2.2. Expression and purification of recombinant N-terminal domain of EBOV NP (NP^{Nt}). NP^{Nt} protein was expressed in Terrific Broth. Induction was carried out at an OD_{600} of 2.0 and growth continued for 18 h at 16°C. Cells were harvested by centrifugation at 3500 rev min^{-1} for 30 min and frozen at −20°C. The pellet was resuspended in lysis buffer (50 m*M* Tris–HCl, 500 m*M* NaCl, 5 m*M* β-mercaptoethanol pH 8.0). Cells were disrupted by Dounce and high-pressure homogenizers and then by sonication and were centrifuged at 35 000 rev min^{-1} for 45 min. Clear supernatant was applied

Table 1
Crystallographic data.

Values in parentheses are for the last resolution shell.

	SeMet $P3_112$	$P2_12_12_1$
Data collection		
Wavelength (Å)	0.97907	1.0000
Beamline	ID	BM
Unit-cell parameters (Å)	$a = b = 56.5$, $c = 63.5$	$a = 36.6$, $b = 49.2$, $c = 53.8$
Resolution (Å)	1.98 (2.01–1.98)	1.75 (1.78–1.75)
Total No. of reflections	73492	63568
No. of unique reflections	8447	10026
Multiplicity	8.7	6.3
Completeness (%)	97.2 (76.3)	97.7 (82.2)
$R_{\rm merge}$† (%)	12.5 (26.2)	4.9 (29.2)
$\langle I/\sigma(I)\rangle$	12.5 (3.7)	31.6 (3.9)
Refinement statistics		
Model composition	901 non-H atoms, 100 solvent (oxygen)	900 non-H atoms, 95 solvent (oxygen)
Resolution limits (Å)	26.7–1.98	29.4–1.75
Reflections in working set	8427	9989
Reflections in test set	729	489
$R_{\rm cryst}$‡/$R_{\rm free}$ (%)	18.4/22.9	18.5/22.5
R.m.s. deviations		
Bond lengths (Å)	0.012	0.013
Bond angles (°)	1.3	1.3
Ramachandran plot, residues in (%)		
Favored regions	100	98.95
Generously allowed regions	0	1.05
Disallowed regions	0	0

† $R_{\rm merge} = \sum_{hkl}\sum_i |I_i(hkl) - \langle I(hkl)\rangle| / \sum_{hkl}\sum_i I_i(hkl)$, where $I_i(hkl)$ is the intensity of the ith observation and $\langle I(hkl)\rangle$ is the mean intensity of the reflections. ‡ $R_{\rm cryst} = \sum_{hkl}\big||F_{\rm obs}| - |F_{\rm calc}|\big| / \sum_{hkl}|F_{\rm obs}|$, the crystallographic R factor; $R_{\rm free} = \sum_{hkl}\big||F_{\rm obs}| - |F_{\rm calc}|\big| / \sum_{hkl}|F_{\rm obs}|$, where all reflections belong to a test set of randomly selected data.

onto an amylose resin column (Qiagen). After 1 h incubation, the flowthrough was collected and the resin was washed with 500 ml lysis buffer. The fusion protein was eluted with 50 m*M* Tris–HCl, 500 m*M* NaCl, 5 m*M* β-mercaptoethanol, 10 m*M* maltose pH 8.0. The fusion protein was digested with rTEV protease with concomitant dialysis against 4 l of a dialysis buffer (50 m*M* Tris–HCl, 500 m*M* NaCl, 5 m*M* β-mercaptoethanol pH 8.0) overnight. The solution was passed through an Ni–NTA agarose gravity column (Qiagen) and the flowthrough fraction containing NP^{Nt} was collected. Concentrated samples of NP^{Nt} were subjected to size-exclusion chromatography using a Superdex 200 column connected to a GE Healthcare ÄKTA FPLC system and were equilibrated with 50 m*M* Tris–HCl, 500 m*M* NaCl, 5 m*M* β-mercaptoethanol pH 8.0. Fractions containing NP^{Nt} were pooled and concentrated.

2.2.3. Expression and purification of recombinant C-terminal domain of EBOV NP (NP^{Ct}). Protein expression, cell disruption and centrifugation were carried out as described above, except that 300 m*M* NaCl was used in the lysis buffer. The supernatant was applied onto a column containing 3 ml Ni–NTA agarose resin (Qiagen) and incubated with the resin mixture for 1 h on a rocking platform. The fusion protein was eluted with a buffer consisting of 50 m*M* Tris–HCl, 300 m*M* NaCl, 5 m*M* β-mercaptoethanol, 250 m*M* imidazole pH 8.0. The eluted protein was digested with rTEV protease and dialyzed overnight against 4 l of solution containing 50 m*M* Tris–HCl, 300 m*M* NaCl, 5 m*M* β-mercaptoethanol pH 8.0. The sample was passed slowly through an Ni–NTA agarose column and the flowthrough containing NP^{Ct} was collected. Concentrated samples were subjected to size-exclusion chromatography on a Superdex 75 column connected to a GE Healthcare ÄKTA system and equilibrated with 50 m*M* Tris–HCl, 150 m*M* NaCl, 5 m*M* β-mercaptoethanol pH 8.0. Fractions containing the protein were pooled and concentrated. A fragment of cDNA corresponding to the N-terminally truncated NP^{Ct} (*i.e.* residues 660–739) was amplified by *Pfu* polymerase (Thermo Scientific) and cloned into *Nco*I and *Sal*I restriction sites in the 6×His_6-MBP-Parallel1 vector. The protein was expressed and purified exactly as the full-length NP^{Ct}.

2.2.4. Preparation of SeMet-labeled NP^{Ct}. SeMet-labeled protein was expressed in M9 minimal medium enriched with 40 µg ml^{-1} of each amino acid except methionine. The culture was grown at 37°C until the OD_{600} reached 0.8. The temperature was changed to 25°C and 50 mg each of leucine, isoleucine, valine and tryptophan and 100 mg each of threonine, lysine, phenylalanine, cysteine and selenomethionine were added per litre. After induction, growth was continued for 17 h. Labeled protein was purified exactly as for unlabeled NP^{Ct}.

2.2.5. Preparation of ^{15}N-labeled and ^{13}C,^{15}N-labeled NP^{Ct}. ^{15}N-labeled and ^{13}C,^{15}N-labeled protein samples were

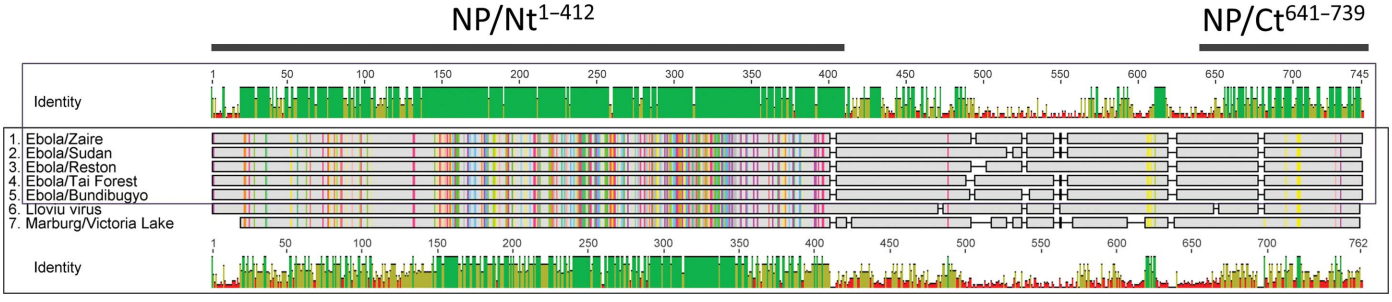


Figure 1
A graphical representation of the amino-acid conservation of the nucleoprotein NP among the members of the *Filoviridae*. The top graph shows the amino-acid identity level among the five subtypes of EBOV, while the bottom graph shows a comparison of all seven NPs, including MARV and LLOV. The colors denote the identity level: from 100% (green) to 0% (red). The center shows a schematic alignment of the sequences with possible sites of deletions (—|—) and insertions (—). Identical residues are shown as color bands running through all sequences. The figure was prepared using *Geneious* (http://www.geneious.com).

obtained by growing transformed cells in M9 minimal medium enhanced by the addition of labeled BioExpress1000 Cell Growth medium (Cambridge Isotope Laboratories, final concentration of 0.6%). Ammonium sulfate (^{15}N, 99%, Cambridge Isotope Laboratories, 1 g l^{-1}) and D-glucose (^{13}C, 99%, Cambridge Isotope Laboratories, 2 g l^{-1}) were used for the labeling. Protein expression was induced at an OD_{600} of 0.9–1.2 and the temperature was changed from 37 to 20°C. After 16 h the cells were harvested and the pellets were stored at −20°C.

Labeled proteins were purified in exactly the same manner as unlabeled NP-Ct, except that the buffer for size-exclusion chromatography consisted of 40 m*M* HEPES, 150 m*M* NaCl, 5 m*M* β-mercaptoethanol pH 7.5. For assignment experiments, a sample of 400 µ*M* ^{15}N-NP^{Ct} and 800 µ*M* ^{13}C,^{15}N-NP^{Ct} in 40 m*M* HEPES, 150 m*M* NaCl, 5 m*M* β-mercaptoethanol pH 7.5 buffer supplemented with 5% D_2O was prepared.

2.3. Crystallization of EBOV NP^{Ct}

NP^{Ct} concentrated to 7.4 mg ml^{-1} was used to set up screens using The JCSG+ Suite (Qiagen) and PEG/Ion HT (Hampton Research) with a Mosquito robot (TTP Labtech). For each crystallization condition, 1:1, 1:2 and 2:1 ratios of precipitant to protein solution were used. Crystals appeared in solutions consisting of 0.2 *M* magnesium formate, 20% PEG 3350 (the JCSG+ Suite) and 0.2 *M* calcium acetate hydrate, 20% PEG 3350 (PEG/Ion HT). Optimization of crystallization conditions with different concentrations of precipitants was carried out manually (hanging-drop method) or automatically (sitting-drop method). Single crystals suitable for X-ray experiments were then grown by the sitting-drop vapor-diffusion method using a 1:1 ratio of reservoir solution (50 m*M* magnesium formate, 19.3% PEG 3350) and protein solution with a protein concentration of ∼12.5 mg ml^{-1}. Crystals of SeMet-labeled NP^{Ct} were grown by the hanging-drop vapor-diffusion method using a 1:1 ratio of reservoir solution to protein solution with a protein concentration of ∼12 mg ml^{-1}. The final conditions found were 300 m*M* magnesium formate, 19.3% PEG 3350.

2.4. Data collection and structure determination

Crystals were cryoprotected under a range of conditions and then screened for diffraction quality. The crystals of unlabeled NP^{Ct} used for final data collection were transferred in a stepwise manner into 20%, 30% and finally 40% PEG 3350. Those of SeMet-labeled protein that gave the best diffraction were soaked in fresh well solution and then in 200 m*M* magnesium formate, 25% PEG 3350, 30% glycerol. All crystals were flash-cooled by immersion into liquid nitrogen. X-ray data were collected at ∼100 K on the SER-CAT beamlines (Southeast Regional Collaborative Access Team) at the Advanced Photon Source, Argonne National Laboratory, Chicago, USA. Data were indexed, integrated and scaled with *HKL*-2000 (Otwinowski & Minor, 1997). See Table 1 for details of data processing.

In the case of SeMet-labeled trigonal crystals, phase estimates were obtained by the SAD method using data collection at the absorption peak, λ = 0.97907 Å (see Table 1). The Se substructure was solved using *SHELXD* (Schneider &

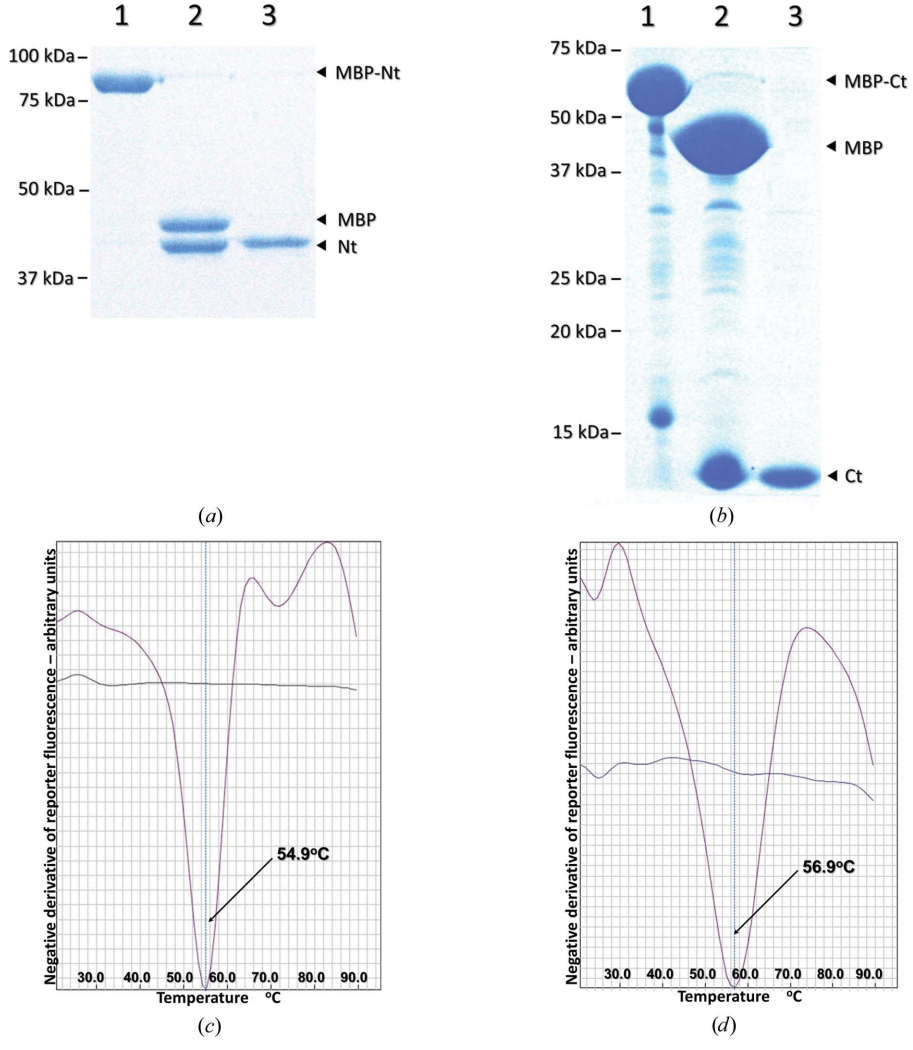


Figure 2
Overexpression, purification and thermostability of isolated N- and C-terminal globular domains of the Zaire EBOV NP. (*a*, *b*) SDS–PAGE gels showing single-step purification of the His-MBP fusion protein (lane 1), the sample following digestion with rTEV (lane 2) and the final sample after removal of the tag and additional purification as specified in §2 (lane 3). (*c*, *d*) Results of the thermal stability assays for the NP^{Nt} and NP^{Ct} domains, respectively. The minimum of the negative derivative of the reporter fluorescence indicates the midpoint of the melting (denaturation) temperature of the protein sample.

Sheldrick, 2002) and phases were calculated using *SHELXE* (Sheldrick, 2002). A large part of the model was automatically built with *ARP*/*wARP* (Langer *et al.*, 2008) and further improved manually with *Coot* (Emsley & Cowtan, 2004). $R_{\rm free}$ was monitored by setting aside ∼5% of the reflections as a test set. Restrained positional and isotropic atomic displacement parameter (ADP) refinement was performed with *PHENIX* (Adams *et al.*, 2010).

The structure of the orthorhombic form (unlabeled NP^{Ct}) was solved using the model from the trigonal form and the molecular-replacement method as implemented in *PHENIX* (Adams *et al.*, 2010). The atomic model of the orthorhombic structure was refined in a manner identical to the trigonal form (see Table 1). Structural figures were prepared using *PyMOL* (http://www.pymol.org/).

2.5. Heteronuclear NMR

A Varian VNMRS 600 MHz spectrometer equipped with a cryoprobe was used to obtain two-dimensional H–N and H–C HSQC and three-dimensional HNCO, HN(CA)CO, CBCA (CO)NH and HNCACB spectra at 25°C. *NMRPipe* (Delaglio *et al.*, 1995) was used to process the spectral data. *NMRView* (Johnson, 2004) and *Sparky* 3 (T. D. Goddard & D. G. Kneller, University of California, San Francisco, USA) were used for spectrum visualization and sequential assignment of backbone and C$^{\beta}$, ^{1}H (except H$^{\alpha}$), ^{13}C and ^{15}N resonances.

2.6. Thermal stability assay

The melting temperature ($T_{\rm m}$) of protein samples was determined by monitoring the fluorescence of SYPRO Orange dye (Life Technologies) in the presence of the protein as a function of temperature. All proteins used in the assays were dialyzed against 50 m*M* Tris–HCl, 250 m*M* NaCl, 5 m*M* β-mercaptoethanol pH 8.0. Assays were performed in 20 µl containing 20 ng protein and 10× the standard concentration of the dye; fluorescence was recorded as a function of temperature from 20 to 90°C using an Applied Biosystems StepOnePlus Real-Time PCR System (Life Technologies). This instrument uses wavelengths of 488 nm for excitation and 586 nm for emission.

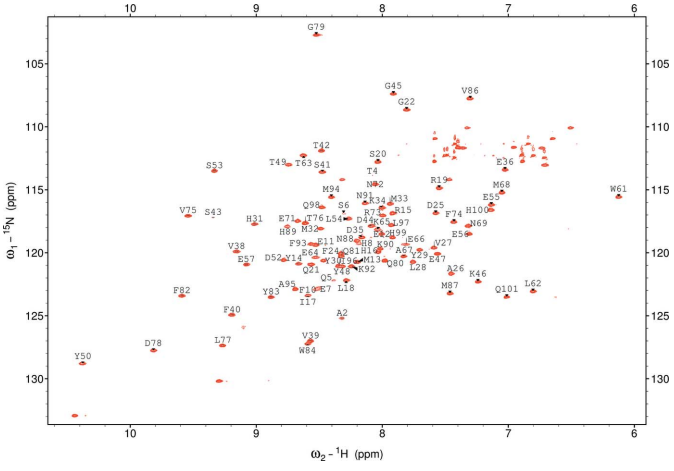


Figure 3
^{1}H–^{15}N HSQC spectrum of NP^{Ct} with backbone amide assignments. The two peaks with the highest ^{15}N p.p.m. are Trp indole NH. Unassigned peaks in the low-field ^{1}H–^{15}N region are owing to side chains.

3. Results and discussion

3.1. Identification of folded domains in Zaire EBOV NP

NPs of *Filoviridae* are significantly longer than those of other members of *Mononegavirales*, with Zaire EBOV NP containing 739 amino acids. There is evidence that this architecture is owing to the presence of two distinct modules, *i.e.* a hydrophobic N-terminal domain (∼450 amino acids), which is important for self-assembly, transcription and replication, and a hydrophilic C-terminal domain (∼150 amino acids) required for formation of the full nucleocapsid and for viral genome replication (Sanchez *et al.*, 1989; Watanabe *et al.*, 2006). It has also been established that the C-terminal fragment (residues 601–739) is involved in incorporation of the nucleocapsid into the virion and in interaction with VP40 (Noda *et al.*, 2007). Nevertheless, the precise boundaries of these functional units have not been identified, nor have they been shown to be stably folded domains.

Analysis of amino-acid conservation among *Filoviridae* NPs yielded results consistent with the suggested two-domain architecture (Fig. 1). Among all seven species, the N-terminal region spanning ∼400 amino acids is most conserved and the C-terminal ∼100 residues also show some conservation. Conservation in the C-terminal region is more prominent when the MARV and LLOV sequences are excluded, as expected. Next, we carried out detailed *in silico* secondary-structure and tertiary-structure predictions, as well as disorder predictions (Supplementary Fig. S1[1]), and we determined that the fragments 1–412 and 641–739 are likely to be globular modules. Polypeptides corresponding to those putative domains were expressed in *E. coli* in fusion with a His-MBP tag. Both were expressed in high yield and were easily purified (Figs. 2*a* and 2*b*). Circular-dichroism (CD) spectra (not shown) indicated a significant content of secondary structure in each domain. Further, a thermal stability assay (TSA) showed that the midpoints of unfolding transition occur at 54.9 and 56.9°C for the NP^{Nt} and NP^{Ct} domains, respectively (Figs. 2*c* and 2*d*). Further work on the NP^{Nt} domain is in progress; the remainder of this paper focuses on the C-terminal domain (NP^{Ct}).

3.2. Determination of the structure of Zaire EBOV NP^{Ct}

To assess whether recombinant NP^{Ct} is fully folded or whether it contains any significant stretch of unstructured polypeptide chain, we prepared ^{15}N-labeled samples and recorded an HSQC spectrum (Fig. 3). The spectrum was well dispersed and consistent with the protein being fully folded in solution. We were also able to determine backbone (except H$^{\alpha}$) and C$^{\beta}$ assignments for 93 out of 95 nonproline residues.

[1] Supporting information has been deposited in the IUCr electronic archive (Reference: BE5269).

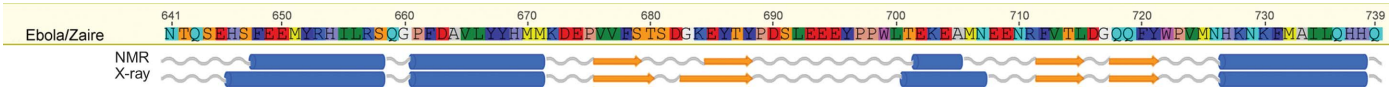


Figure 4
The secondary-structure elements of the C-terminal domain of the EBOV NP, as determined by heteronuclear NMR in solution and crystallographic analysis. The amino acids in the sequence are colored according to the *RASMOL* (Sayle & Milner-White, 1995) convention; the α-helices are shown as blue cylinders and the β-strands as yellow arrows.

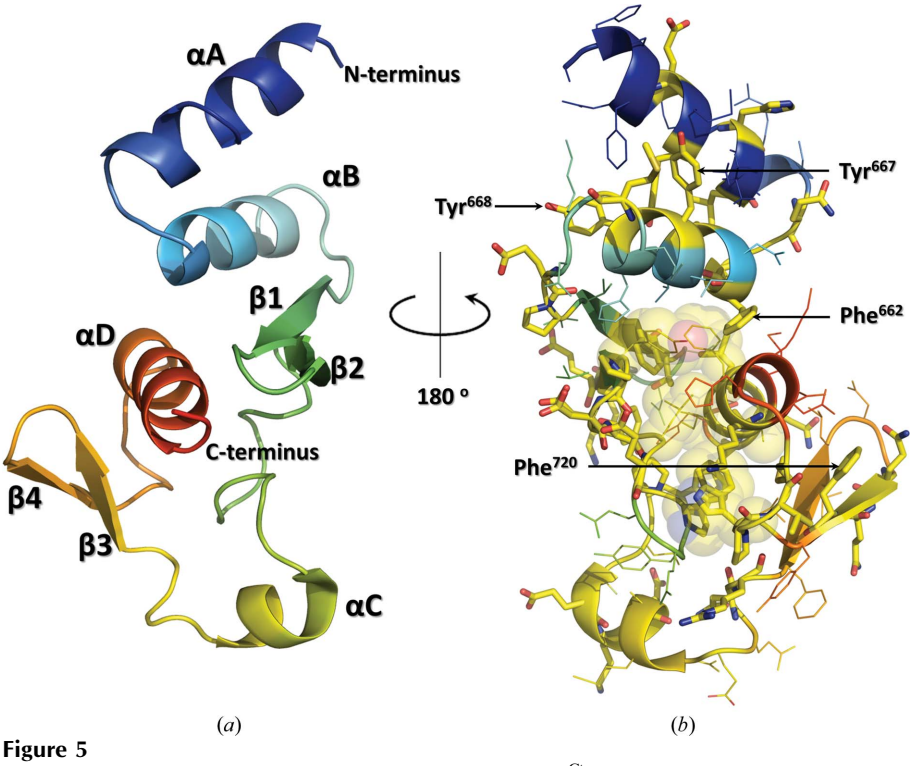


Figure 5
A diagrammatic representation of the structure of EBOV NP^{Ct}. (*a*) A ribbon diagram showing the tertiary structure with secondary-structure elements identified and labeled; the color scheme follows the rainbow convention with blue at the N-terminus and red at the C-terminus. (*b*) A view of the molecule after 180° rotation showing all the side chains as lines and all conserved side chains as sticks. The hydrophobic core made up of completely buried residues is visualized using van der Waals spheres. Four conserved residues are labeled.

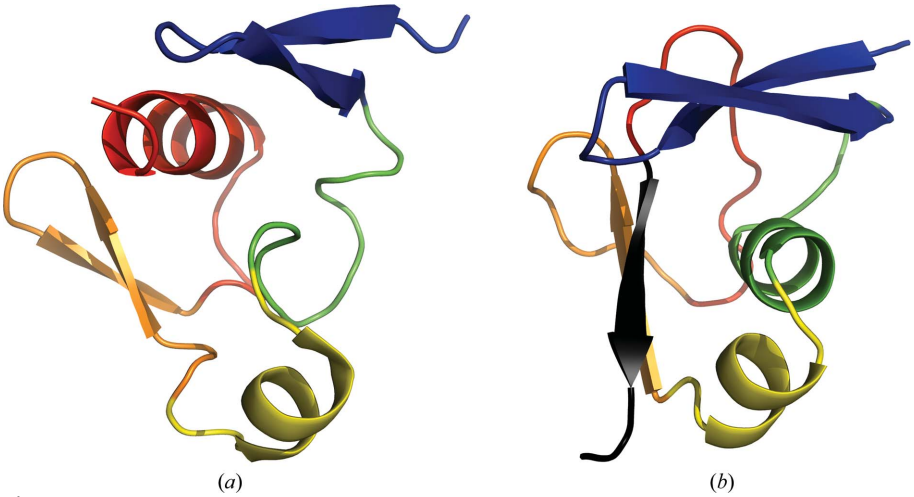


Figure 6
A comparison of the EBOV NP^{Ct} fold without the αA/αB hairpin (*a*) with the TGS domain of the CLOLEP-03100 protein from *Clostridium leptum* (*b*). The topologically corresponding fragments are colored identically in both molecules. Note that the critical difference in NP^{Ct} is the presence of the C-terminal helix in place of a β-strand in the TGS domain.

The assigned chemical shifts were used in *TALOS-N* (Shen & Bax, 2013) to define the secondary-structure elements. At this point, owing to the successful crystallization of NP^{Ct}, we discontinued further efforts to determine the full three-dimensional structure in solution by NMR. Nevertheless, the availability of assignments will allow future identification and characterization of interactions with binding partners.

Orthorhombic crystals of native NP^{Ct} that diffracted beyond 1.8 Å resolution were obtained using PEG 3350 as a precipitant (see §2). A SeMet-labeled sample yielded a trigonal crystal form that diffracted well to 2.0 Å resolution and allowed structure determination using SAD and subsequent refinement. The model includes residues 645–739, with the four N-terminal amino acids not being identifiable in the electron density. Further, the side chains of Glu645, His646, Glu649, Lys684, Glu695, Glu709, Lys728 and Gln739 are partly disordered so that some or all of their atoms are not visible in the electron density. Once the refinement of the trigonal form was completed, we used this structure to solve the orthorhombic crystal form by molecular replacement. This was successful, and the resulting structure was refined in a fashion similar to that described above. Most of the side-chain disorder was also observed in the orthorhombic form. The crystallographic details are given in Table 1.

3.3. EBOV NP^{Ct} represents a novel fold

The atomic models derived from the two crystal forms are very similar: the secondary and tertiary structures are virtually identical. Importantly, the secondary-structure elements agree well with the NMR data (Fig. 4). At its N-terminus, NP^{Ct} contains an antiparallel pair of α-helices (αA, residues 648–658; αB, 661–671) followed by two

β-strands arranged in an antiparallel hairpin (β1, residues 676–680; β2, 683–688). A stretch of irregular but structured polypeptide chain leads to a short α-helix (αC, 701–707) at the other extremity of the oblong molecule, followed by another β-hairpin (β3, 712–715; β4, 718–721) and a C-terminal α-helix, αD (residues 727–738) (Fig. 5*a*). The latter appears to be a critical element of the structure, as it runs through the center of the molecule, providing a scaffold around which most of the remainder of the molecule is folded. Residues that belong to the αD helix make contact with the αB helix, with both β-hairpins and with the coil structure connecting the two β-hairpins. Three hydrophobic residues within the αD helix, *i.e.* Phe731, Ala733 and Ile734, are completely buried. They form the small, densely packed hydrophobic core along with the equally occluded Tyr688, Leu692, Pro697 and Trp722 (Fig. 5*b*). Importantly, all of these residues are highly conserved among the five EBOV strains (the only differences are in the Sudan subtype, with Y688F and I734V substitutions).

Detailed *BLAST* searches confirmed that the amino-acid sequence of EBOV NP^{Ct} is unique and has no homologues in any other proteins except for other strains of EBOV. A search of the Protein Data Bank using the refined NP^{Ct} model with the *DALI* server (Holm *et al.*, 2008) did not reveal any similarities between NP^{Ct} and known tertiary folds. However, when the N-terminal pair of α-helices are removed, the program *FATCAT* (Ye & Godzik, 2004) detected weak similarity to a TGS domain (PDB entry 3hvz), which is one of the many representatives of the β-grasp superfamily (Burroughs *et al.*, 2007; Fig. 6). In general terms, the topologies of the two domains show similarities and several (though not all) secondary-structure elements are conserved, but these similarities cannot be taken as evidence of any evolutionary relationship.

Given the unusual antiparallel hairpin at the N-terminus of the domain, in which the αA helix makes no other contacts with the rest of the protein, we wondered whether the αA helix should be considered to be an integral part of the domain. To test this, we generated an N-terminally truncated protein, residues 660–739, purified it and measured its denaturation temperature (data not shown). Interestingly, the thermal stability of this variant was significantly lower, with a T_m of only 48.0°C, compared with 56.9°C for the NP^{Ct} protein. We conclude that the αA helix should be regarded as an integral part of the NP^{Ct} domain.

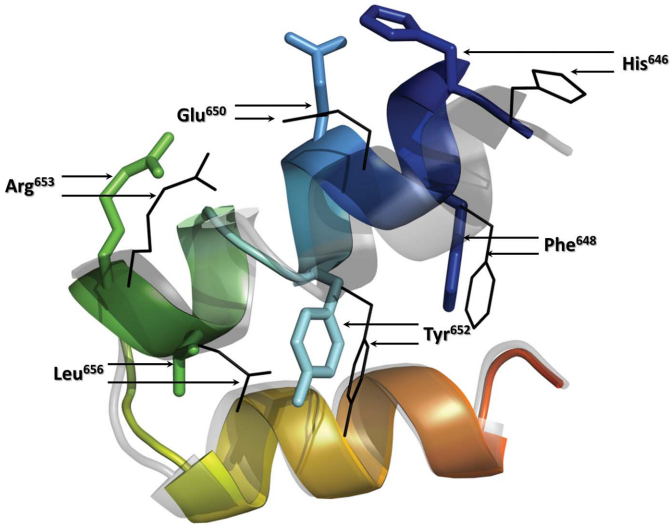
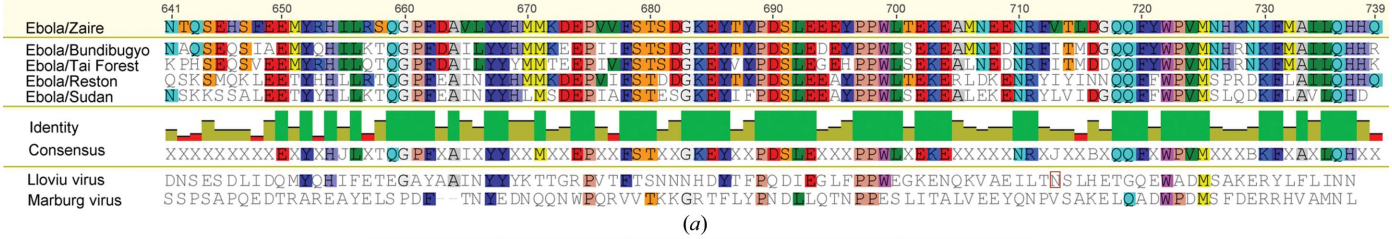


Figure 7
A comparison of the structures of the N-terminal helical fragment in the two crystal forms of NP^{Ct}. The colored ribbon with side chains depicted as sticks represents the structure in the orthorhombic form, whereas the lighter structure with selected side chains drawn with black lines is that in the trigonal form. Selected residues are labeled showing their positions in the two structures.

3.4. Comparison of the two crystal forms

As already mentioned, the atomic models derived from the two crystal forms are very similar. The only significant



(*a*)

Identity	Bundibugyo	Tai Forest	Reston	Sudan
Zaire	82%	79%	65%	60%
Sudan	61%	62%	74%	
Reston	65%	64%		
Tai Forest	86%			

(*b*)

Figure 8
Amino-acid conservation in the NP^{Ct} domain among *Filoviridae*. (*a*) Sequence alignment: the Zaire strain sequence (the subject of this study) is shown at the top; the sequences of the remaining four strains are below, followed by a graph of amino-acid identity level within the family (green denotes invariant residues) and the resulting consensus template sequence. Amino acids are colored according to the *RASMOL* convention. The sequences of the Marburg and Lloviu viruses are shown below the consensus sequence: only the residues that match the consensus are colored. (*b*) Pairwise amino-acid identity levels within the NP^{Ct} domain of the five EBOV strains.

difference is in the orientation of the αA helix (residues 648–658) relative to the main module (Fig. 7). In general terms, the N-terminal helix rotates as a rigid body by 8.8° and undergoes a translation of 1.7 Å. This conformational change, evidently owing to altered crystal packing, results in a substantial rearrangement of the interface between the αA and the αB helices. The αA helix makes no contact with the NP^{Ct} domain, other than with the αB helix, and consequently the different orientation of the αA helix does not affect any other parts of the structure. The remainder of the NP^{Ct} (residues 661–738) superposes in the two crystal forms with an r.m.s. (main chain) of 0.55 Å. The only other small, but significant, difference between the two structures is in the loop harboring residues 715–718; in one structure it is shifted by ∼2.0 Å compared with the other, and this also appears to be owing to a close crystal-packing contact.

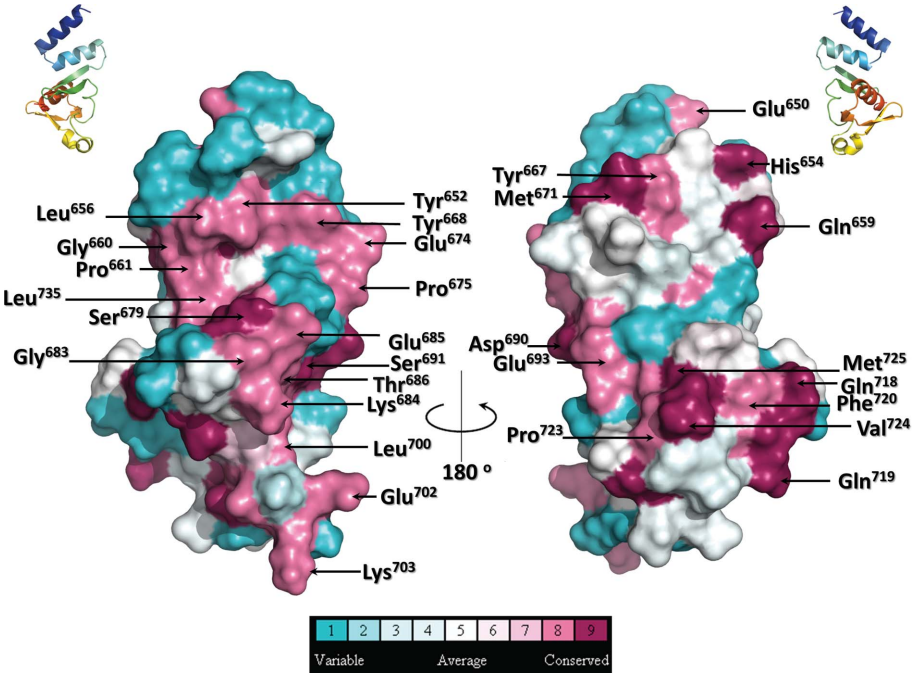


Figure 9
Graphical representation of the surface amino-acid conservation using the crystal structure of the Zaire EBOV NP^{Ct}. The color scale is based upon the level of conservation as determined by the *ConSurf* server. Categories 8 and 9 correspond to fully conserved residues. Small ribbon diagrams are shown at the top for the viewer's convenience.

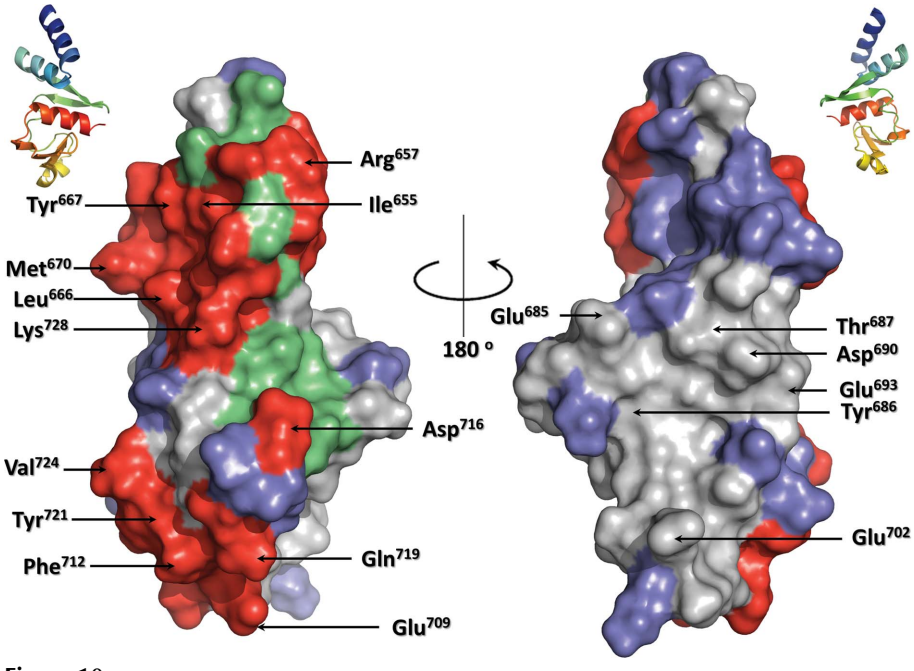


Figure 10
Surface patches involved in crystal contacts in the two EBOV NP^{Ct} structures. Green denotes patches involved in crystal contacts in the trigonal form and blue those in the orthorhombic form, while red denotes patches involved in both forms. The gray surfaces do not participate in any contacts. Ribbon diagrams are shown to assist the viewer with the orientation of the molecule. Selected residues involved in the patches are labeled.

3.5. Evolutionary conservation of NP^{Ct} among *Filoviridae*

As already pointed out, the NP^{Ct} domain is less stringently conserved among the five strains of EBOV than the NP^{Nt} domain. Pairwise comparisons reveal sequence-identity levels ranging from 60 to 86% (Fig. 8). A total of 47 residues are completely conserved among the five strains and constitute a consensus template. Within this group, a number are amino acids located in the small hydrophobic core, as expected. The LLOV and MARV sequences deviate significantly from the EBOV consensus. The LLOV sequence shows ∼25% identity to the Sudan and Reston strains and shares 16 residues with the EBOV consensus. MARV, on the other hand, shows such low similarity that *BLAST* does not pick up the relationship to EBOV when the MARV sequence is used. In our alignment, only 12 amino acids from the MARV sequence conform to the EBOV consensus. Overall, only seven amino acids are invariant among all the *Filoviridae* in NP^{Ct}.

3.6. Analysis of NP^{Ct} surfaces: implications for protein–protein interactions and antigenicity

Given the hypothesized function of the NP^{Ct} domain as a hub for protein–protein interactions, we carefully analyzed the molecular surfaces with reference to amino-acid conservation,

propensity to form interactions mediating crystal contacts and electrostatic potential. These analyses are also relevant with respect to the established high antigenicity of the EBOV NP^{Ct} domain.

Fig. 9 illustrates the amino-acid conservation among the surface-accessible residues. Of the total of 48 amino acids that were fully conserved among the five subtypes of EBOV, 36 have more than 15 Å^2 of exposed surface. The majority are scattered throughout the surface. The most conserved patch is located in a concave depression between the N-terminal αA and αB helices and the $\beta1/\beta2$ hairpin. This patch includes, among others, Tyr652, Leu656, Tyr668, Glu674, Ser679, Glu685 and Leu725. It is highly possible that this is the site of one or more of the protein–protein interactions involving the EBOV NP.

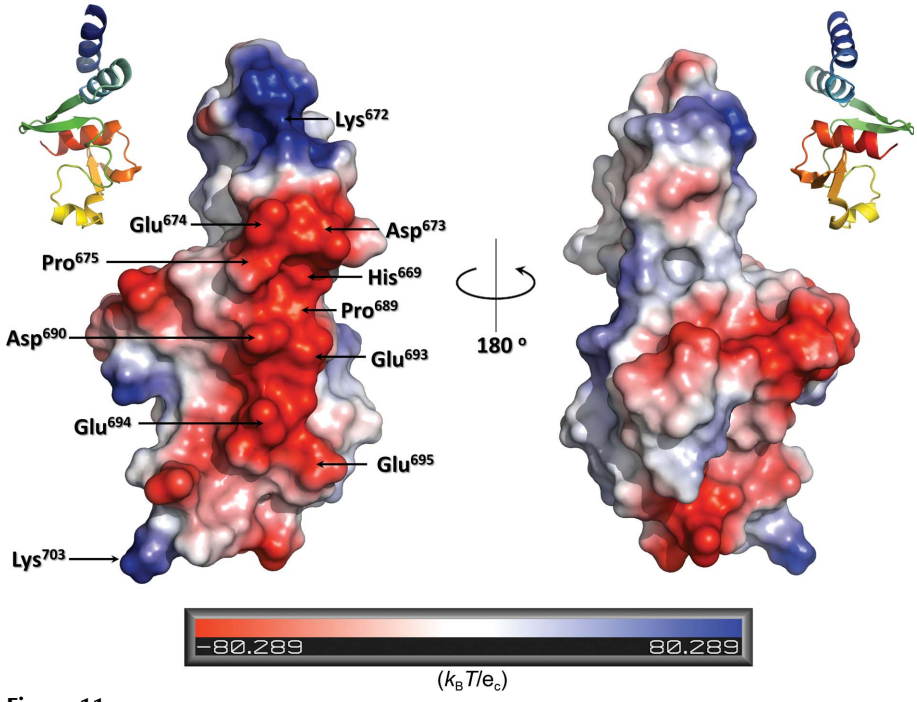


Figure 11
The electrostatic potential calculated in *PyMOL* mapped onto the solvent-accessible surface. Ribbon diagrams are shown to assist the viewer with the orientation of the molecule. Selected amino acids are labeled.

We also compared crystal contacts in the two forms of NP^{Ct} that we have studied. Intermolecular contacts in protein crystals are often mediated by surface patches that are physiologically relevant for protein–protein interactions. This is particularly true for those patches that mediate intermolecular contacts in different crystal forms of the same protein (Xu & Dunbrack, 2011). Fig. 10 shows that there are several distinct sets of intermolecular contacts in the two forms. Two patches are involved in crystal contacts in both forms. Predictably, both are hydrophobic in nature. One of them includes Phe712 and Tyr721, while the other is made up mainly of Ile655, Leu666 and Tyr667. Interestingly, the first four are completely conserved in the Zaire, Bundibugyo and Tai Forest strains, but not in the Sudan and Reston strains, which either show lower mortality upon infection (Sudan) or do not infect humans (Reston).

Finally, Fig. 11 shows the electrostatic surface potential of the Zaire EBOV NP^{Ct}. The pI of this domain is estimated at 4.9 and ranges among the other strains from 4.6 in Sudan to 5.5 in Reston. The low pI is owing to a preponderance of acidic residues, six of which create a contiguous patch on the protein surface: Asp673, Glu674, Asp690, Glu693, Glu694 and Glu695. The first four are either completely conserved or are conserved as Glu/Asp acids in all five strains of EBOV; Glu694 is replaced by a Gly in the Tai Forest substrain, whereas Glu695 is replaced by Ala in the Reston and Sudan strains.

4. Conclusions

Our work demonstrates for the first time that the nucleoproteins of the ebolaviruses contain two distinct, globular domains that can be produced in a recombinant form in *E. coli*. The N-terminal domain is within the fragment 1–412, although the precise boundaries will have to be confirmed. The second globular domain is located at the C-terminal end of the protein. We were able to show, both by NMR and X-ray crystallography, that a fragment encompassing residues 641–739 is completely folded. Crystal structures obtained for two distinct crystal forms of this NP^{Ct} domain revealed a novel fold, with a topology distantly related to some members of the β-grasp superfamily.

An intriguing aspect of EBOV NP^{Ct} is its relatively low amino-acid sequence conservation among the fives subtypes of EBOV and particularly when compared with proteins from LLOV and MARV. This is surprising given that the NP^{Ct} has been implicated in several protein–protein interactions involving other EBOV proteins, and therefore one might expect a significantly higher conservation of solvent-exposed residues. Future studies will show how the high sequence variation impacts on these interactions. The assignment of NMR chemical shifts will assist in those studies.

This research was supported by the University of Virginia School of Medicine (ZSD) and a contract from DTRA, HDTRA1-14-C-0006 (DAE). The use of the Advanced Photon Source (APS) was supported by the US Department of Energy, Office of Science and Office of Basic Energy Sciences under contract No. W-31-109-Eng-38. Supporting institutions of SER-CAT may be found at http://www.ser-cat.org/members.html. We want to thank Dr Darkhan Utepbergenov and Ms Natalya Olekhnovich for help with protein expression and purification, Dr John Bushweller and Adam Boulton for help with the NMR work, Dr Adam Godzik (Burnham-Safford Institute, La Jolla) for help with the

bioinformatics searches and Dr Michał Jakób (NCI, Frederick, USA/Wroclaw University of Technology, Poland) for discussions and advice.

References

Adams, P. D. *et al.* (2010). *Acta Cryst.* D**66**, 213–221.
Albertini, A. A., Wernimont, A. K., Muziol, T., Ravelli, R. B. G., Clapier, C. R., Schoehn, G., Weissenhorn, W. & Ruigrok, R. W. (2006). *Science*, **313**, 360–363.
Ashkenazy, H., Erez, E., Martz, E., Pupko, T. & Ben-Tal, N. (2010). *Nucleic Acids Res.* **38**, W529–W533.
Beniac, D. R., Melito, P. L., Devarennes, S. L., Hiebert, S. L., Rabb, M. J., Lamboo, L. L., Jones, S. M. & Booth, T. F. (2012). *PLoS One*, **7**, e29608.
Bharat, T. A., Noda, T., Riches, J. D., Kraehling, V., Kolesnikova, L., Becker, S., Kawaoka, Y. & Briggs, J. A. (2012). *Proc. Natl Acad. Sci. USA*, **109**, 4275–4280.
Booth, T. F., Rabb, M. J. & Beniac, D. R. (2013). *Trends Microbiol.* **21**, 583–593.
Bornholdt, Z. A., Noda, T., Abelson, D. M., Halfmann, P., Wood, M. R., Kawaoka, Y. & Saphire, E. O. (2013). *Cell*, **154**, 763–774.
Brauburger, K., Hume, A. J., Mühlberger, E. & Olejnik, J. (2012). *Viruses*, **4**, 1878–1927.
Burroughs, A. M., Balaji, S., Iyer, L. M. & Aravind, L. (2007). *Biol. Direct*, **2**, 18.
Cole, C., Barber, J. D. & Barton, G. J. (2008). *Nucleic Acids Res.* **36**, W197–W201.
Cuff, J. A., Clamp, M. E., Siddiqui, A. S., Finlay, M. & Barton, G. J. (1998). *Bioinformatics*, **14**, 892–893.
Delaglio, F., Grzesiek, S., Vuister, G. W., Zhu, G., Pfeifer, J. & Bax, A. (1995). *J. Biomol. NMR*, **6**, 277–293.
Dessen, A., Volchkov, V., Dolnik, O., Klenk, H. D. & Weissenhorn, W. (2000). *EMBO J.* **19**, 4228–4236.
Emsley, P. & Cowtan, K. (2004). *Acta Cryst.* D**60**, 2126–2132.
Feldmann, H., Klenk, H. D. & Sanchez, A. (1993). *Arch. Virol. Suppl.* **7**, 81–100.
Gomis-Rüth, F. X., Dessen, A., Timmins, J., Bracher, A., Kolesnikowa, L., Becker, S., Klenk, H. D. & Weissenhorn, W. (2003). *Structure*, **11**, 423–433.
Hartlieb, B., Muziol, T., Weissenhorn, W. & Becker, S. (2007). *Proc. Natl Acad. Sci. USA*, **104**, 624–629.
Holm, L., Kääriäinen, S., Rosenström, P. & Schenkel, A. (2008). *Bioinformatics*, **24**, 2780–2781.
Huang, Y. J., Acton, T. B. & Montelione, G. T. (2014). *Methods Mol. Biol.* **1091**, 3–16.
Johnson, B. A. (2004). *Methods Mol. Biol.* **278**, 313–352.
Kelley, L. A. & Sternberg, M. J. (2009). *Nature Protoc.* **4**, 363–371.
Kuhn, J. H. *et al.* (2013). *Arch. Virol.* **158**, 301–311.
Langer, G., Cohen, S. X., Lamzin, V. S. & Perrakis, A. (2008). *Nature Protoc.* **3**, 1171–1179.
Lee, J. E., Fusco, M. L., Hessell, A. J., Oswald, W. B., Burton, D. R. & Saphire, E. O. (2008). *Nature (London)*, **454**, 177–182.
Leung, D. W., Prins, K. C. *et al.* (2010). *Nature Struct. Mol. Biol.* **17**, 165–172.
Leung, D. W., Shabman, R. S., Farahbakhsh, M., Prins, K. C., Borek, D. M., Wang, T., Mühlberger, E., Basler, C. F. & Amarasinghe, G. K. (2010). *J. Mol. Biol.* **399**, 347–357.
Linding, R., Russell, R. B., Neduva, V. & Gibson, T. J. (2003). *Nucleic Acids Res.* **31**, 3701–3708.
Negredo, A. *et al.* (2011). *PLoS Pathog.* **7**, e1002304.
Noda, T., Hagiwara, K., Sagara, H. & Kawaoka, Y. (2010). *J. Gen. Virol.* **91**, 1478–1483.
Noda, T., Watanabe, S., Sagara, H. & Kawaoka, Y. (2007). *J. Virol.* **81**, 3554–3562.
Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.
Paessler, S. & Walker, D. H. (2013). *Annu. Rev. Pathol.* **8**, 411–440.
Rice, P., Longden, I. & Bleasby, A. (2000). *Trends Genet.* **16**, 276–277.
Rudolph, M. G., Kraus, I., Dickmanns, A., Eickmann, M., Garten, W. & Ficner, R. (2003). *Structure*, **11**, 1219–1226.
Salvaggio, M. R. & Baddley, J. W. (2004). *Dermatol. Clin.* **22**, 291–302.
Sanchez, A., Kiley, M. P., Holloway, B. P., McCormick, J. B. & Auperin, D. D. (1989). *Virology*, **170**, 81–91.
Sayle, R. A. & Milner-White, E. J. (1995). *Trends Biochem. Sci.* **20**, 374–376.
Schneider, T. R. & Sheldrick, G. M. (2002). *Acta Cryst.* D**58**, 1772–1779.
Sheffield, P., Garrard, S. & Derewenda, Z. (1999). *Protein Expr. Purif.* **15**, 34–39.
Sheldrick, G. M. (2002). *Z. Kristallogr.* **217**, 644–650.
Shen, Y. & Bax, A. (2013). *J. Biomol. NMR*, **56**, 227–241.
Sherwood, L. J. & Hayhurst, A. (2013). *PLoS One*, **8**, e61232.
Watanabe, S., Noda, T. & Kawaoka, Y. (2006). *J. Virol.* **80**, 3743–3751.
Wit, E. de, Feldmann, H. & Munster, V. J. (2011). *Genome Med.* **3**, 5.
Xu, Q. & Dunbrack, R. L. Jr (2011). *Nucleic Acids Res.* **39**, D761–D770.
Ye, Y. & Godzik, A. (2004). *Protein Sci.* **13**, 1841–1850.
Zhang, A. P. P., Bornholdt, Z. A., Liu, T., Abelson, D. M., Lee, D. E., Li, S., Woods, V. L. Jr & Saphire, E. O. (2012). *PLoS Pathog.* **8**, e1002550.