NMR Structure of the N-terminal Domain of the Nonstructural Protein 3 from the SARS Coronavirus.

Pedro Serrano, Margaret A. Johnson, Marcia S. Almeida, Reto Horst, Torsten Herrmann, Jeremiah S. Joseph, Benjamin W. Neuman, Vanitha Subramanian, Kumar S. Saikatendu, Michael J. Buchmeier, Raymond C. Stevens, Peter Kuhn, & Kurt Wüthrich, Departments of Molecular Biology, Molecular and Integrative Neurosciences, Cell Biology, and Chemistry, Skaggs Institute for Chemical Biology, and Joint Center for Structural Genomics, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, USA. Institute for Molecular Biology and Biophysics, ETH Zürich, CH-8093 Zürich, Switzerland.

* Corresponding author. Mailing address: Department of Molecular Biology, MB-44, The Scripps Research Institute, 10550 North Torrey Pines Rd., La Jolla, CA, 92037. Phone: (858) 784-8011. Fax: (858) 784-8014. E-mail: wuthrich@scripps.edu

Running Title: NMR structure of SARS-CoV nsp3a.

Abstract: 110 words

Text: 5,202 words
This paper describes the structure determination of nsp3a, the N-terminal domain of the SARS-CoV non-structural protein 3. Nsp3a exhibits an ubiquitin-like globular fold of residues 1–112, and a flexibly extended glutamic acid-rich domain of residues 113–183. In addition to the four β-strands and two α-helices that are common to ubiquitin-like folds, the globular domain of nsp3a contains two short helices representing a feature that has not previously been observed in these proteins. NMR chemical shift perturbations showed that these unique structural elements are involved in interactions with single-stranded RNA. Structural similarities with proteins involved in various cell-signaling pathways indicate possible roles of nsp3a in viral infection and persistence.
INTRODUCTION

Severe acute respiratory syndrome (SARS) is a viral infectious disease that has attracted worldwide attention since an outbreak in 2003 (26). It has been postulated that the SARS coronavirus (SARS-CoV) was introduced to the human population from animal coronaviruses (26). Coronaviruses comprise a large group of enveloped, positive-sense, single-stranded RNA viruses that have been classified in the Nidovirales order. There are three groups of CoVs, based on serological cross-reactivity and phylogenetic relatedness. The SARS-CoV is distantly related to the group 2 viruses and has been classified in group 2b (38).

The SARS-CoV represents one of the largest currently known RNA genomes. It is composed of at least 14 functional ORFs that encode three classes of proteins, i.e., structural proteins (the S, M, E, N, 3a, 7a and 7b proteins), non-structural proteins (nsp1 to nsp16), and the accessory proteins (3b, 6, 8, 9b, and 14) (38). With regard to the non-structural proteins, the translation of the SARS-CoV genome produces two large replicase polyproteins (pp1a, pp1ab), which are processed by two proteases to yield 16 mature nsps that mediate RNA replication and processing. Since the SARS outbreak in 2003, knowledge of the structure, activity and function of some of these proteins has increased considerably (30, 32, 35, 41, 46), however, the biological roles of many of the SARS-CoV proteins remain unknown. In this paper we describe the NMR structure determination and a preliminary functional characterization of nsp3a, the N-terminal domain of the largest of the non-structural proteins, nsp3.

SARS-CoV nsp3 is a 213 kDa polypeptide involved in RNA replication, and has been proposed to consist of seven domains, nsp3a-g, which have been identified based on
phylogenetic conservation and predicted amino acid secondary structure (38). The biological role of nsp3 is only partially understood, and so far structures have been determined only of the two domains nsp3b, which has been described as an ADP ribose-1”-phosphatase (34), and nsp3d, which is a papain-like protease (PLpro) involved in the proteolytic processing of pp1a and pp1ab. Nsp3d contains three domains, two of which are involved directly in proteolysis, while the third one has a ubiquitin-like fold (31).

Nsp3a exhibits less than 35% sequence identity with other known proteins, and the closest homologues are found in other coronaviruses. The alignment shown in Figure 1 indicates that group 2a CoVs (e.g., murine hepatitis virus and porcine hemagglutinating encephalomyelitis virus) exhibit higher similarity with nsp3a than proteins from groups 1 (e.g., human coronavirus 229E) and 3 (e.g., avian infectious bronchitis virus). The 183-residue nsp3a domain consists of a C-terminal subdomain of residues 113–183 that is rich in acidic residues (38% E and 12% D), and a 112-residue N-terminal subdomain with a more homogeneous content of amino acids (Fig. 1). This report presents a structural characterization of nsp3a(1–183), and the structure determination of the subdomain nsp3a(1–112) in solution by NMR spectroscopy.
MATERIALS AND METHODS

Production of nsp3a. Full length nsp3a (nsp3 residues 1–183) and a construct devoid of the residues 113–183, nsp3a(1–112), were cloned into the expression vector pMH1F (6xHis-tag; pBAD derivative), and expressed in DL41 E. coli cells with induction at 14°C in 2XYT medium. Each of the two constructs was shown, by 1D ¹H NMR, to form a folded globular domain (data not shown). To facilitate expression of samples suitable for NMR structure determination, both constructs were subcloned into pET-25b (Novagen). These plasmids were used to transform E. coli strain BL21-CodonPlus (DE3)-RIL (Stratagene). The expression of uniformly ¹³C,¹⁵N-labeled nsp3a(1–112) was carried out by growing freshly transformed cells in M9 minimal medium containing 1 g/L ¹⁵NH₄Cl and 4 g/L ¹³C₆-D-glucose as the sole nitrogen and carbon sources. Cell cultures were grown at 37 °C with vigorous shaking to an optical density at 600 nm of 0.8 to 0.9. The temperature was then lowered to 18 °C, and after induction with 1 mM isopropyl-β-D-thiogalactopyranoside, the cell cultures were grown for 18 h. The cells were harvested by centrifugation, resuspended in extraction buffer (50 mM sodium phosphate at pH 6.5, 150 mM NaCl, 0.1% Triton X-100, and Complete protease inhibitor tablets (Roche)), and lysed by sonication. The cell debris was removed by centrifugation (20000g, 20 min). For the first purification step, the soluble protein was loaded onto an anion exchange column (HiTrap Q FF, Amersham) equilibrated with 50 mM sodium phosphate buffer at pH 6.5 containing 150 mM NaCl. The proteins were eluted with a 150–1000 mM NaCl gradient. Fractions containing nsp3a(1–112) were pooled and concentrated to a volume of 10 mL using centrifugal ultrafiltration devices (Millipore). Subsequently the sample was loaded onto a size-exclusion column (Superdex 75, Amersham) equilibrated with 50 mM sodium
phosphate buffer at pH 6.5 containing 150 mM NaCl, and eluted with the same buffer.

The fractions containing nsp3a(1–112) were again pooled and concentrated to a final volume of 550 µL for a final protein concentration of 1.8 mM.

**Production of nucleic acid-free protein for NMR spectroscopy.** Nsp3a prepared as described in the preceding section co-purifies with nucleic acids, as was readily observed in the 1D $^1$H NMR spectrum (Fig.10a). Nucleic acid-free samples were obtained by the following modification of the purification procedure. After the anion-exchange chromatography, the sample was kept at 25 °C for 18 h. The protein solution was subsequently loaded onto a size-exclusion column (Superdex 75, Amersham) equilibrated with 50 mM sodium phosphate buffer at pH 6.5 containing 150 mM NaCl, and eluted with the same buffer. Under these conditions, the protein and the nucleic acid eluted separately. The fractions containing nucleic acid-free nsp3a(1–112) were again pooled and concentrated to a final volume of 550 µL for a final protein concentration of 1 to 2 mM. The 1D $^1$H NMR spectrum of the sample used for the NMR structure determination (Fig. 10b) confirms the absence of nucleic acids.

**NMR spectroscopy.** NMR measurements were performed at 298 K with Bruker Avance 600, DRX 700 and Avance 800 spectrometers (Bruker BioSpin, Billerica, MA), equipped with TXI HCN z- or TXI HCN xyz-gradient probeheads. Proton chemical shifts were referenced to internal 3-(trimethylsilyl)-1-propanesulfonic acid sodium salt (DSS). The $^{13}$C and $^{15}$N chemical shifts were referenced indirectly to DSS, using the absolute frequency ratios (42). The following NMR spectra were used to obtain sequence-specific backbone and side chain resonance assignments: 2D [$^{15}$N,$^1$H]-HSQC, 2D [$^{13}$C,$^1$H]-HSQC,
3D HNCA, 3D HNCA(C), 3D CBCA(CO)NH, 3D HNCO, 3D HC(C)H-TOCSY, 3D $^{15}$N-resolved [$^1$H,$^1$H]-TOCSY, and 2D [$^1$H,$^1$H]-NOESY.

Steady-state $^{15}$N-$[^1$H]-NOEs were measured using TROSY-based experiments (32, 45) on a Bruker Avance 600 spectrometer, with a saturation period of 3.0 s and a total interscan delay of 5.0 s. Diffusion experiments were recorded on a Bruker DRX700 spectrometer using a LED pulse scheme (1), with a diffusion time of 50 ms and sine shaped gradients of 4.5 ms. The data were processed with the TopSpin software (Bruker BioSpin, Billerica, MA).

The interaction of nsp3a(1–112) with ssRNA was evaluated by comparison of the 2D [$^{15}$N,$^1$H]-HSQC spectra of nsp3a(1–112) recorded at four nsp3a(1–112) : ssRNA ratios, i.e., 16 : 1, 8 : 1, 4 : 1 and 2 : 1. As controls, 2D [$^{15}$N,$^1$H]-HSQC spectra were obtained after addition of octa-U and octa-A in four-fold excess with respect to the protein concentration, using otherwise identical conditions. The weighted average of the $^1$H and $^{15}$N chemical shift differences, $\Delta \delta_{av}$, was calculated as $\Delta \delta_{av} = \{0.5[\Delta \delta (^{1}$H $N) ]^2 + (0.2\Delta \delta (^{15}$N)]^2\}^{1/2}$ (28).

**Structure determination.** The structure calculation was based on a 3D $^{15}$N-resolved [$^1$H,$^1$H]-NOESY spectrum, and on two 3D $^{13}$C-resolved [$^1$H,$^1$H]-NOESY spectra recorded with the carrier frequency in the aliphatic and the aromatic region, respectively. All three data sets were recorded with mixing times of 60 ms. In the input for the standalone version of the software package ATNOS/CANDID (9, 10), these NOE data were supplemented with the amino acid sequence and the chemical shift lists from the independently obtained sequence-specific resonance assignment (36). Seven cycles of automated NOESY peak picking and NOE cross peak identification with ATNOS (9),
automated NOE assignment with CANDID (10), and structure calculation with the
torsion angle dynamics algorithm of CYANA (8) were performed. In the second and
subsequent cycles, the intermediate protein structure was used as an additional guide for
the interpretation of the NOESY spectra. During the first six cycles,
ATNOS/CANDID/CYANA uses ambiguous distance restraints. In the final cycle, only
distance restraints which could be attributed to a single pair of hydrogen atoms were
retained. The 20 conformers with the lowest residual CYANA target function values
obtained from the seventh ATNOS/CANDID/CYANA cycle were energy-minimized in a
water shell with the program OPALp (18, 21), using the AMBER force field (5). The
program MOLMOL (19) was used to analyze the ensemble of 20 energy-minimized
conformers.

**Structure validation and data deposition.** Analysis of the stereochemical quality of the
molecular models was accomplished using the Joint Center for Structural Genomics
(JCSG) Validation Central Suite (http://www.jcsg.org), which integrates seven validation
tools: Procheck, SFcheck, Prove, ERRAT, WASP, DDQ, and Whatcheck. The $^1$H, $^{13}$C
and $^{15}$N chemical shifts have been deposited in the BioMagResBank (BMRB;
http://www.bmrb.wisc.edu) under the accession number 7029 (36). The atomic
coordinates of the bundle of 20 conformers used to represent the solution structure of
nsp3a(1–112), and of the conformer closest to the mean coordinates of the ensemble,
have been deposited in the Protein Data Bank (http://www.rcsb.org/pdb/) with the codes
2GRI and 2IDY, respectively.

**Protein stoichiometry determination.** PFO-PAGE was performed according to the
method of Bear (30). Purified protein samples were mixed 1:1 with PFO loading buffer
containing 8% (w/v) perfluoro-octanoic acid (PFO), 100 mM Tris, 20% (v/v) glycerol and 0.05% (w/v) orange G. Samples with protein concentrations of 250 µM, 500 µM and 1 mM were loaded onto precast 4–20% Tris–glycine gels and electrophoresis was performed with a standard Tris–glycine running buffer (Invitrogen) to which 0.5% (w/v) PFO was added. Protein was detected by SYPRO-ruby post stain (Invitrogen).

**Electrophoretic mobility shift assay.** Protein samples (two-fold dilutions from 128 µM to 1 µM) were mixed with 0.8 µg of RNA substrate in 20 µl of assay buffer containing 150 mM NaCl, 50 mM Tris at pH 8.0 and 5 mM CaCl₂. The RNA sequences used included ssRNA1, AAAUACCUCUCAAAAAUAACACCACACCAUAUACCACAU, and ssRNA2, GGGGAUAAAA. Samples were incubated at 37 ºC for 1 h, and analyzed by native electrophoresis on precast 6% acrylamide DNA retardation gels (Invitrogen). RNA was detected by SYBR-gold post-stain and photographed using an ultraviolet light source equipped with a digital camera. Protein was then detected by SYPRO-ruby post-stain. Densitometric analysis was performed using a flatbed scanner with ImageJ software (NIH). The mobility shift of RNA at each protein concentration was calculated relative to the maximum shift observed in each experiment. \( K_d \) values were determined from the midpoints of the fitted titration data (37).

**Nuclease susceptibility assay.** Nsp3a(1–183) and nsp3a(1–112) were incubated with several different nucleases in order to characterize nucleic acids that co-purified with both proteins. RNAse-free DNase I (NEB), T7 endonuclease I (NEB), RNAse I (NEB), RNAse A (Invitrogen) and RNAse T1 (Ambion) cleavage assays were thus performed at 37ºC for 1 h with the manufacturer’s recommended buffer conditions. Digested samples
were analyzed by native electrophoresis on precast 6% acrylamide DNA retardation gels (Invitrogen). Nucleic acid was detected by SYBR-gold post-stain.
RESULTS

Nsp3a structure determination. The NOE cross peaks that were unambiguously assigned in the seventh cycle of the ATNOS/CANDID/CYANA calculation (see the Materials and Methods section for details) yielded 1888 meaningful upper distance limits, which were used as input for the final structure calculation with the program CYANA. The residual CYANA target function value of $1.88 \pm 0.28$ Å$^2$ and the average global root-mean-square deviation (RMSD) value relative to the mean coordinates of $0.77 \pm 0.09$ Å calculated for the backbone atoms of residues 20–108 in the bundle of Figure 2a (Table 1) represent a high-quality NMR structure determination.

Solution structure of nsp3a. Nsp3a(1–112) exhibits a ubiquitin-like fold with four helices and four β-strands arranged in the sequential order β1–α1–β2–α2–310–α3–β3–β4 (Figs. 1 and 2). The long helix α2 and the presence of the α1- and 310-helices, which have not been observed in other ubiquitin-like proteins, make the overall structure more elongated than other ubiquitin-related folds. The strand β1 spans the residues 20–24 and is connected via a well-defined nine-amino acid linker to the helix α1 containing the residues 34–37. A short turn then leads to β2 with residues 42–46. The helix α2 with residues 52–66 is followed by a short loop that leads to the 310-helix of residues 70–75, which is further connected by a short turn with the helix α3 of residues 79–84. The last two regular secondary structures, β3 with residues 89–91 and β4 with residues 101–106, form an antiparallel β-sheet, and they are connected to each other by a tight turn followed by an extended chain segment. The electrostatic potential surface of nsp3a(1–112) shows a pronounced polarity (Fig. 3), with the helices α2, α3 and 310...
exhibiting mainly negative charges to the solvent, while the strands β1 and β3, and the helix α1 contain primarily positive or hydrophobic surface residues.

This is the second domain with a ubiquitin-like fold found within full-length nsp3. Previously, the N-terminal 70-amino acid segment of the fourth domain of nsp3, the papain-like protease (nsp3d or PLpro), was found to have a ubiquitin-like fold (31). In Figure 4, regular secondary structure elements in the segment 20–104 of nsp3a have been superimposed with the corresponding polypeptide segments in the region of residues 725–777 of nsp3, which corresponds to the N-terminal domain of the papain-like protease (31). In as far as they overlap, the two structures share the same topology as canonical ubiquitin-like proteins, such as ISG15 (24) and Bacillus subtilis yukD (41). However, nsp3a also displays unique features (colored yellow in Fig. 4), i.e., the connection between the strands β1 and β2 in nsp3a is longer than that in nsp3d and includes the helix α1. nsp3a has two additional helices inserted between the strands β2 and β3, and the helix α2 is much longer in nsp3a than in nsp3d.

Characterization of flexible regions in nsp3a. Mobility in the two nsp3a protein constructs was investigated by heteronuclear $^{15}$N–$^{1}$H-NOE experiments (Figs. 5 and 6). Figure 5 shows the values of the steady-state $^{15}$N–$^{1}$H-NOE for each $^{15}$N–$^{1}$H moiety in nsp3a(1–112). The residues 20–108, for which the mobility of the backbone $^{15}$N–$^{1}$H moieties is essentially limited to the overall tumbling of the molecule, have positive NOE values of about 0.8. In contrast, the residues 1–19 and 110–112 have values in the range of -0.4 to 0.5, indicating increased mobility for these polypeptide segments, which are also visibly less well defined in the structure (Fig. 2a).
In order to investigate the structural role of the Glu-rich subdomain of residues 113–183, two nsp3a variants were generated which differ in the presence or absence of the C-terminal Glu-rich region, and the 2D $^{15}$N,$^1$H-HSQC spectra of the two proteins were then compared (Fig. 6a). There are no significant changes of the chemical shifts of the resonances of residues 2–112 in the two proteins, which indicates that both variants contain a similarly structured globular domain. These data also show for the full-length nsp3a (blue in Fig. 6a) that most of the peaks from residues 113–183 are in the random coil chemical shift region ($^1$H shifts between 7.5–8.5 ppm). These chemical shifts, together with the higher intensity of these resonances when compared with the peaks from the globular region (Fig. 6, a and b), are indicative of a flexibly extended polypeptide segment. This is confirmed by the fact that the $^{15}$N,$^1$H-NOE values for most of the peaks corresponding to residues 113–183 are negative (pink in Fig. 6c). Thus, the C-terminal Glu-rich subdomain is best described as a flexible tail of residues 113–183 attached to the globular domain of residues 1–112.

**Nsp3a(1–112) is a monomer in solution.** During the purification of nsp3a(1–112) we noticed that the retention volume (Vr) of nsp3a by size exclusion chromatography (Superdex 75, Amersham) was lower than expected for a globular protein with 12.6 kDa molecular weight. In view of the implications for the structure determination and the biological activity of the protein, we decided to further investigate the oligomeric state of nsp3a (1–112) in solution using NMR diffusion experiments and PFO-PAGE. In diffusion NMR experiments, the decay of the signal intensity ($\ln I/I_0$) versus the square of the magnetic field gradient ($G^2$) was used to estimate the translational diffusion properties of the proteins (40). In Figure 7a we compare data obtained for 1 mM solutions...
of nsp3a(1–112), ribonuclease A and chymotrypsinogen, which have molecular weights of 12.6 kDa, 13.7 kDa and 25.0 kDa, respectively. The nsp3a(1–112) intensity decay curve is located between the two standards, which is indicative of the presence of the monomeric form, since the elongated shape of nsp3a(1–112) should result in a lower diffusion coefficient than near-spherical proteins with similar molecular weights. A PFO-PAGE gel also indicates that nsp3a(1–112) exists predominantly in the monomeric form at room temperature. The assays performed at three different protein concentrations, of 1 mM, 500 µM and 250 µM (Fig. 7b) show that even at 1 mM concentration the monomeric form predominates, and only a small amount of the dimeric form can be observed.

**Nsp3a(1–112) binds single-stranded RNA.** In the initial nsp3a(1–112) purification assays (see Materials and Methods) the protein co-purified with small fragments of single-stranded RNA (ssRNA). These fragments were readily detected in the 1D 1H NMR spectra (Fig. 10a), and were subsequently also observed by native PAGE analysis. In addition to preparing the nucleic acid-free protein for the NMR structure determination (described in Materials and Methods), we also investigated the nature of the co-purifying nucleic acids. To this end a sample of nucleic acid-loaded nsp3a(1–112) was unfolded in 6 M guanidinium-HCl solution, and the mixture was subsequently loaded onto a Superdex 75 size-exclusion column (Fig. 10c). The mass spectrometry analysis of the isolated fragments allowed us to identify a RNA component with a molecular weight of 1327.3 (Fig. 11). The different peaks found in this spectrum are consistent with the sequences AU, GAU and GAUA, with the longest component corresponding to GAUA.
Nuclease digestion assays of protein samples containing co-purifying nucleic acids further revealed that the major species associated with nsp3a(1–183) was DNA, which could be completely removed by DNAse I treatment (Fig. 8a), and that the shorter form of the protein retained a much smaller nucleic acid species that was partly susceptible to RNase A digestion and was not susceptible to RNase I or T1 digestion (Fig. 8a). The incomplete digestion by RNase A and the lack of cleavage by RNase I or T1 were interpreted as an indication of the formation of a robust protein–RNA complex.

We then went on to study the binding of exogenous ssRNA substrates to nsp3a(1–112), starting from the aforementioned observation that the endogenous RNA contained the predominant trinucleotide sequence AUA. We thus designed two AUA-containing ssRNA fragments for further studies, ssRNA1 with the sequence AAAUACCUCUCUCAAAAAUAACACCACCAUAUACCACAU, and ssRNA2 with the sequence GGGGAUAAAA. The binding of nsp3a(1–112) to these two ssRNAs was assessed by electrophoretic mobility shift assay (EMSA), using RNA-free protein prepared by as described in Materials and Methods. EMSA showed that nsp3a(1–112) bound to the two ssRNA substrates with similar affinity (Fig. 8b). Measurement of the percentage of bound ssRNA1 at variable concentrations of nsp3a(1–112) (Fig. 8c) allowed to estimate the dissociation constant of the nsp3a(1–112)–ssRNA1 complex to be approximately 20 µM. In control experiments, no binding was observed with several ssDNA sequences, or with double-stranded RNA sequences containing the motif AUA (Fig. 9). Furthermore, no binding to smaller ssRNA forms, such as fragments containing only G and U, and octa-A, octa-C, octa-G and octa-U could be detected. Thus, RNA binding by nsp3a is consistent with the profile of a sequence-sensitive ssRNA-binding
Following up on these results, NMR chemical shift perturbation studies were performed in order to map the regions of nsp3a(1–112) that are affected by the interaction with ssRNA2. Figure 12, a and b, shows the effect of the addition of ssRNA2 on the chemical shift for each residue in nsp3a(1–112). The residues with large chemical shift perturbations are all located on the same surface area of the protein. It is rather surprising that this contact region comprises the two loops linking β3 and β4, and β1 and α1, and the helices α1 and 3₁₀, which contain a surplus of negatively charged amino acid side chains (Fig. 3). There is thus an indication that these chemical shift perturbations might result primarily from long-range effects on the protein conformation, rather than from direct protein–RNA contacts.

Nsp3a did not interact with other ssRNA species tested. For example, the superposition of the [¹⁵N,¹H]-HSQC spectra of nsp3a in the presence and absence of octa-U (Fig. 12c) does not show any significant chemical shift differences, indicating that octa-U does not bind to the protein and supporting that the interaction of nsp3a(1–112) with ssRNA is sequence-specific.
DISCUSSION

Nsp3a is well conserved within different SARS-CoV sequences but exhibits low sequence identity (<35%) to other coronavirus nsp3 proteins. The closest sequence homologies with the globular domain of nsp3a prevail for the replicase polyproteins of porcine hemagglutinating encephalomyelitis virus (PHEV) and murine hepatitis virus (MHV) (Fig. 1). For example, the sequences in the strands $\beta 3$ and $\beta 4$ are well conserved among all group 2 coronaviruses, including SARS-CoV, while the region containing the $3_{10}$- and $\alpha 3$-helices is less well conserved, and the helix $\alpha 3$ actually appears to be absent in the group 1 and 3 CoVs. Additionally, the regions corresponding to $\beta 1$ and $\alpha 1$ in nsp3a exhibit a high number of conservative amino acid substitutions. It is worth mentioning that $\beta 1$, $\alpha 1$ and $\beta 4$ define the positively charged surface areas of nsp3a (Fig. 3, right-hand panel). The $\alpha 1$- and $3_{10}$-helices, which have not been observed in other ubiquitin-like proteins, seem to be important for the interaction of nsp3a with ssRNAs, since they exhibit extensive chemical shift perturbations upon ssRNA interaction, and since other ubiquitin homologues do not exhibit RNA binding activity.

Although the observed affinity of nsp3a for ssRNA cannot by itself define a unique biological function, it seems to be important for the overall nsp3 biological role. As indicated above, nsp3 is a large multidomain protein and only two of its domains, nsp3b and nsp3d, have been structurally and functionally characterized to date. The analysis of these domains indicates that nsp3 is a multifunctional protein involved in multiple biological processes, such a proteolysis (31) and RNA processing (34). The fact that the presently studied N-terminal region of nsp3 and two of its other domains, nsp3c and nsp3e, exhibit RNA binding activity (Neuman, B. W. et al., manuscript in preparation)
together with the ADP-ribose-1\'-phosphate dephosphorylation activity of nsp3b (34) suggest that this protein could be also involved in the replication and processing of viral RNA. Although the short sequences AUA and GAUA are common in the genome, a possible biological function for the sequence-specific RNA-binding activity observed for nsp3a might be in binding to the 5’end of the SARS-CoV genome. The sequence AUA occurs several times in the 5’ untranslated region (UTR) of the genome, including at the extreme 5’end. Proteins that specifically recognize the 5’ UTR might function in cap-dependent translation or, alternatively, in genome replication or subgenomic RNA synthesis.

The observation of two ubiquitin-like structures within nsp3 (nsp3a and the N-terminal domain of nsp3d) has important implications in attempting to assign its likely biological function. Apart from nsp3d being a cysteine protease, it is also a potent de-ubiquitinating enzyme that has been extensively studied (2, 3, 31). It has been speculated earlier that the ubiquitin-like (ubl) domain of SARS-CoV nsp3d might act as a decoy for cellular ubiquitinating enzymes, thereby protecting nascently synthesized viral proteins from proteasome-mediated degradation. Alternatively, the two ubl domains might be involved in modulation of protein–protein interaction pathways of cellular immunomodulators, such as interferons and ISGylating enzymes. This view is reinforced by the structural similarity of the two ubl domains of nsp3 with ISG15, an interferon-stimulated gene that is induced as a primary response to diverse stimuli, including viral infections. The SARS-CoV proteins 3b, 6 and the nucleocapsid protein have recently been shown to function as effective interferon antagonists (16).
It seems possible that other SARS-CoV proteins, such as nsp1 (13) (and possibly host proteins as well) might also be part of these pathways, acting at either the RNA or protein levels. Several studies probing the intricate interplay of viral and host proteins during the progression of the SARS-CoV viral cycle have been reported (22, 33, 39). Since the biological role of nsp3a still remains unclear, structural homology studies could at this point provide insights into the potential function of this domain and its role within the viral cycle.

**Nsp3a exhibits three-dimensional structure similarity with Ras-interacting domains.**

Many of the structure homologues of nsp3a interact with other polypeptides to regulate processes such as protein degradation, cell signaling (12) and antiviral response (24). It seems significant that five of them are Ras-interacting proteins. Based on the primary sequences, the ubiquitin $\alpha/\beta$-roll superfold comprises five families (14). Members of three of these families, RA (RalGDS/AF6 Ras-association domain), RBD (Raf-like Ras-binding domain) and PI3K_rbd (Ras-binding domain of PI3Kinase-like proteins) interact with Ras (14). A large fraction of the structure homologues of nsp3a(1–112) identified using the software DALI are members of these families. The protein nsp3a(1–112) has the highest structure similarity with the Ras-interacting domain (RID) of RALGDS, a member of the RA family with which it shares the topology of the ubiquitin-like fold (Fig. 13d). This effector of Ras is a stimulator of the guanine nucleotide dissociation mechanism specific for Ral. RID-RalGDS binds Ras through its C-terminal domain and presents low sequence identity with other Ras interacting proteins, but similar hydrophobic profiles (12). The superposition of the three-dimensional structures of RID-RalGDS and nsp3a(1–112) reveals a region with conserved residues located in the $\beta1$-
strand of nsp3a(1–112) (Fig. 13, a and b) which is intimately involved in the Ras contact interface. Similarly, the Ras-binding domain of the AF6 protein (29), which is also a member of the RA family, shows three-dimensional structure homology with nsp3a(1–112) (Fig. 13d) and similar residues located in the β1 region (Fig. 13b). Both Ral-GDS and AF6 are known as Ras effectors. Similar patterns are also found in other RA domains with significant levels of structural homology with nsp3a(1–112), e.g., the human Grb7 protein and the guanine nucleotide exchange factor for Rap1 (25).

In general, Ras domains contain a combination of hydrophobic and acidic residues that interact with hydrophobic and positive groups on RIDs. Both nsp3a and the different aforementioned Ras-interacting proteins exhibit these characteristics (Fig. 13c). In nsp3a, the conserved basic residue R23 is located in the β1-strand, which exhibits high consensus sequence with the RA family (Fig. 13b). This suggests that nsp3a could interfere in biological processes that involve Ras. Given its high similarity with RA family proteins, there might be a potential interaction of nsp3a with human Ras proteins during SARS-CoV infection. Ras family proteins (RFPs) act as molecular switches that cycle between inactive GDP- and active GTP-bound states. In this manner, RFPs control cell growth, motility, intracellular transport and differentiation. The fundamental role of Ras in the cell cycle progression from phase G0 to G1 has been extensively reported (6, 27). Molecular interactions that result in Ras inactivation prevent cell progression to the G1 phase. In this context, murine hepatitis virus is able to induce cell cycle arrest in the G0/G1 phase during the lytic infection cycle (4). It has also been shown that some SARS-CoV proteins are able to induce apoptosis or G0/G1 arrest in transfected cells (17, 43, 44).
Overall, the structural similarity of nsp3a(1–112) and RIDs thus leads us to hypothesize that nsp3a may have a physiological role in cell cycle arrest.

**Structure and potential functional role of the C-terminal Glu-rich subdomain.** The Glu-rich C-terminal polypeptide segment 113–183 of nsp3a shows less than 25% sequence identity with the corresponding acidic regions in other coronavirus genomes, whereby the SARS-CoV protein contains overall a somewhat higher percentage of acidic residues than the acidic regions of other CoVs. Similar motifs are found in some eukaryotic proteins (11, 23). In mammals, these acid-rich polypeptide segments are mainly involved in the transport of mRNA from the nucleus to the cytoplasm by association with RNA binding proteins. For example, the pp32/leucine-rich acidic protein (pp32/LAMP) associates with HuR, which binds to AU-rich elements of mRNAs to export mRNAs from the nucleus to the cytoplasm (11). Interestingly, Higashino *et al.* reported that several viruses interfere with this transport, in order to increase the production of their virions by the cellular machinery (11).

The NMR data of the figures 5 and 6 now show that the Glu-rich region of nsp3a forms a flexible tail attached to the globular region of residues 1–112. Although sequence similarity to other proteins is not identifiable, several cellular proteins also contain regions with high percentages of acidic residues. The homopolymer of glutamic acid, poly-L-Glu, is unstructured at pH 8, but can adopt helical structures at pH 5 (15). Some acidic regions of polypeptides exhibit well-defined regular secondary structure when interacting with other proteins. The structure of the RanGAP (35) complex with RanBP1 and RanGAP presents examples of both situations. Ran is a nuclear Ras-related protein that regulates both transport between nucleus and cytoplasm, and formation of the mitotic
spindle or nuclear envelope in dividing cells (35). The C-terminal region of RanGAP, which is important for binding affinity, exhibits an acidic motif that is flexibly disordered in both the complexed and the uncomplexed forms of RanGAP, whereas other acid-rich segments of this protein comprise folded secondary structure elements. Taking into account the high similarity of nsp3a(1–112) with some of these other polypeptides involved in protein–protein interaction processes, as well as its location in the large multidomain protein nsp3, it could well be that the long, flexibly extended Glu-rich segment could have an important role in interactions with other SARS-CoV or host cell molecules, and that this domain might adopt a well-defined fold during interactions with other polypeptides. Overall, based on the structural data reported in this paper, there are indications that the globular and non-globular subdomains of nsp3a are important for SARS-CoV infection and persistence, and thus represent new potential targets for therapeutic intervention.
ACKNOWLEDGEMENTS

This study was supported by the NIAID/NIH contract #HHSN266200400058C “Functional and Structural Proteomics of the SARS-CoV”, and the Joint Center for Structural Genomics through the NIH/NIGMS grant #U54-GM074898. P. S. was supported by a fellowship from the Spanish Ministry of Science and Education and by the Skaggs Institute of Chemical Biology. M. S. A. was supported by the PEW Latin American Fellows Program in the Biological Sciences and by the Skaggs Institute for Chemical Biology. M. A. J. was supported by a fellowship from the Canadian Institutes of Health Research and by the Skaggs Institute for Chemical Biology. Kurt Wüthrich is the Cecil H. and Ida M. Green Professor of Structural Biology at TSRI and a member of the Skaggs Institute for Chemical Biology.
REFERENCES


43. Yee-Joo, T., B. C. Fielding, P.-Y. Goh, S. Shen, T. H. P. Tan, S. G. Lim, and W. Hong. 2004. Overpression of 7a, a protein specifically encoded by the severe...


Table 1: Input for the structure calculation and characterization of the bundle of 20 energy-minimized CYANA conformers that represent the NMR structure of nsp3a(1–112)

<table>
<thead>
<tr>
<th>Quantity</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOE upper distance limits</td>
<td>1888</td>
</tr>
<tr>
<td>Intraresidual</td>
<td>400</td>
</tr>
<tr>
<td>Short-range</td>
<td>637</td>
</tr>
<tr>
<td>Medium-range</td>
<td>491</td>
</tr>
<tr>
<td>Long-range</td>
<td>360</td>
</tr>
<tr>
<td>Dihedral angle constraints</td>
<td>118</td>
</tr>
<tr>
<td>Residual target function value, Å²</td>
<td>1.88 ± 0.28</td>
</tr>
<tr>
<td>Residual NOE violations</td>
<td></td>
</tr>
<tr>
<td>Number ≥ 0.1Å</td>
<td>22 ± 4</td>
</tr>
<tr>
<td>Maximum, Å</td>
<td>0.13 ± 0.01</td>
</tr>
<tr>
<td>Residual dihedral angle violations</td>
<td></td>
</tr>
<tr>
<td>Number ≥ 2.5, deg</td>
<td>1 ± 1</td>
</tr>
<tr>
<td>Maximum, deg</td>
<td>2.44 ± 0.82</td>
</tr>
<tr>
<td>Amber energies, kcal/mol</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>-3102.71 ± 80.00</td>
</tr>
<tr>
<td>van der Waals</td>
<td>-254.89 ± 15.32</td>
</tr>
<tr>
<td>Electrostatic</td>
<td>-3679.82 ± 82.77</td>
</tr>
<tr>
<td>rmsd from ideal geometry</td>
<td></td>
</tr>
<tr>
<td>Bond lengths, Å</td>
<td>0.0078 ± 0.0002</td>
</tr>
<tr>
<td>Bond angles, deg</td>
<td>2.086 ± 0.029</td>
</tr>
<tr>
<td>rmsd to the mean coordinates, Å²</td>
<td></td>
</tr>
<tr>
<td>bb (20—108)</td>
<td>0.77 ± 0.09</td>
</tr>
<tr>
<td>ha (20—108)</td>
<td>1.02 ± 0.10</td>
</tr>
<tr>
<td>Ramachandran plot statistics</td>
<td></td>
</tr>
<tr>
<td>Most favored regions (%)</td>
<td>73</td>
</tr>
<tr>
<td>Additional allowed regions (%)</td>
<td>24</td>
</tr>
<tr>
<td>Generously allowed regions (%)</td>
<td>3</td>
</tr>
<tr>
<td>Disallowed regions (%)</td>
<td>0</td>
</tr>
</tbody>
</table>

- a. Except for the six top entries, the average value for the 20 energy-minimized conformers with the lowest residual CYANA target function values and the standard deviation among them are given.
- b. bb indicates the backbone atoms N, Cα, C’; ha stands for “all heavy atoms”. The numbers in parentheses identify the residues for which the rmsd was calculated.
- c. As determined by PROCHECK (20).
FIGURE LEGENDS

FIG. 1. (a) Sequence alignment of human SARS-CoV nsp3a(1–112) and the homologous regions from bat SARS-CoV (AAZ67050), murine hepatitis virus (MHV-A59, NP_740609), porcine hemagglutinating encephalomyelitis virus (HEV-VW572, YP_459949), human coronavirus (hCoV 229E, NP_835345) and avian infectious bronchitis virus (IBV-Cal99, AAS00078). The residue numbers at the top correspond to the sequence of the human SARS-CoV and do not account for the insertions shown in the drawing. In each sequence the conserved residues relative to SARS-CoV nsp3a are in bold. The regular secondary structure elements of SARS-CoV nsp3a are indicated by boxes. (b) Sequence of the subdomain of residues 113–183 of human SARS-CoV.

FIG. 2. NMR structure of nsp3a(1–112). (a) Stereoview of the polypeptide backbone of a bundle of 20 energy-minimized CYANA conformers superimposed for minimal RMSD value of the backbone atoms of residues 20–108. The N-terminal segment of residues 1–19 is flexibly disordered (see Fig. 5 below). (b) Stereoview of a ribbon representation of the conformer with the smallest RMSD relative to the mean coordinates of the ensemble of Fig. 2a. In (a) and (b), β-strands are cyan and helices red. Selected residue positions are indicated in (a), and the regular secondary structures are identified in (b).

FIG. 3. Electrostatic surface potential of nsp3a(1–112). Positive and negative electrostatic potential is represented in blue and red, respectively. On the left we show the surface of the helices α2, α3 and 3₁₀, and of the loop between the strands β3 and β4, which contain a high density of acidic residues (Fig. 1). On the right there is the surface of helix α1 and
the strands β1, β2 and β4, which contain mainly neutral and basic residues. Positions of
selected charged residues are indicated.

FIG. 4. Superposition of nsp3a(1–112) (color code: regular secondary structures that
superimpose with nsp3d, green; segments not present in nsp3d, yellow; other segments,
gray) and the ubiquitin-like domain of nsp3d (31) (PDB code: 2FE8) (regular secondary
structures that superimpose with nsp3a, red; other segments, gray). The structure
superposition was performed using the SSM module of Coot (7). 30 Cα-atoms were
superimposed with a RMSD value of 2.22 Å, i.e., from nsp3a(1–112) the residues 20–26,
40–46, 49–54, 87–91 and 100–104, and from nsp3d the residues 725–731, 739–745,

FIG. 5. 15N-{1H}-NOE values plotted as relative intensities, I_{rel}, versus the sequence of
nsp3a(1–112). Diamonds represent experimental measurements, which are linked by
straight lines along the sequence. Gaps represent proline residues, which lack a backbone
{1H} atom, or overlapping residues in the 15N-{1H} correlation spectrum that could not be
integrated accurately. The experiment was recorded at a {1H} frequency of 600 MHz, using
a saturation period of 3.0 s and a total interscan delay of 5.0 s.

FIG 6. (a) Superposition of the 2D [15N,1H]-HSQC spectra of nsp3a(1–183) (blue) and
nsp3a(1–112) (red). (b) High contour level presentation of a 2D [15N,1H]-HSQC
spectrum of nsp3a(1–183). (c) Heteronuclear NOE experiment with nsp3a(1–183), using
a saturation period of 3.0 s and an interscan delay of 5.0 s. Negative and positive peaks are colored pink and green, respectively.

FIG. 7. Study of the oligomeric state of nsp3a(1–112). (a) Data obtained from NMR diffusion experiments at 700 MHz. The relative NMR signal intensity, ln I/I₀, is plotted versus the square of the gradient field strength, G². ◇, nsp3a(1–112); □, ribonuclease A; ▲, chymotrypsinogen. (b) PFO-PAGE of nsp3a(1–112); the size of the protein complexes was estimated from the Benchmark protein ladder shown on the left (Invitrogen). The protein concentration increases from right to left in three steps of 250 μM, 500 μM and 1 mM. The filled arrowheads denote the positions of the monomeric (12.6 kDa) and dimeric (25.2 kDa) forms of nsp3a(1–112).

FIG. 8. Association of nsp3a(1–183) and nsp3a(1–112) purified from E. coli with nucleic acids. (a) Nucleic acid was visualized with SYBR-gold staining before or after digestion with nucleases specific to DNA (DNase I, T7 endonuclease) or RNA (RNase I, RNase A, RNase T1). Cleavage assays were performed at 37°C for 1 h and digested samples were analyzed by native electrophoresis on precast 6% polyacrylamide gels. Open arrowheads denote co-purified nucleic acid species associated with nsp3a(1–112) or nsp3a(1–183), respectively. (b) Electrophoretic mobility shift assays performed to estimate the RNA binding affinity of nsp3a(1–112). Samples containing ssRNA1 or ssRNA2 were incubated at 37°C for 1 h, with variable concentrations of protein, and analyzed by native electrophoresis on precast 6% polyacrylamide gels. RNA was detected by SYPRO-gold post-stain, and the fraction of bound RNA was calculated...
relative to the maximum binding observed in each experiment. Lane markings denote protein only (P), ssRNA only (O), ssRNA with two-fold dilutions of protein from a final concentration of 128 µM to 2 µM for ssRNA1 (1–7 in the left panel), and with four-fold dilutions of protein from 64 µM to 1 µM for ssRNA2 (2–8 in the right panel). Electrophoretic mobilities of free (f) and bound (b) forms of each ssRNA species are indicated with arrowheads. (c) ssRNA1-binding at variable concentrations of nsp3a(1–112), as calculated from the EMSA data shown in panel (b).

FIG. 9. EMSA experiments performed to evaluate the affinity of nsp3a(1–112) for different nucleic acid species. (a) Gels obtained after loading mixtures of nsp3a(1–112) with ten different ssDNA fragments (1–10). Lanes labeled P and M correspond to nucleic acid-free protein and nucleic acid marker, respectively. Comparison of the two gels, using nucleic acid (left) and protein (right) specific stains, indicates that nsp3a(1–112) does not exhibit affinity for ssDNAs. (b) Gels containing decreasing concentrations (100-1.6 µM) of nsp3a(1–112), in the presence of 800 ng of a ssRNA 40-mer lacking the sequence AUA (left), a dsRNA 20-mer (center) and a ssDNA 40-mer (right). In lanes labeled N, only nucleic acid species were loaded. In none of the cases, an interaction of nsp3a(1–112) and nucleic acids (NA) is observed. All experiments were performed after incubation of nsp3a(1–112) and the corresponding nucleic acid fragment for 1 h at 37 °C.

FIG. 10. (a) 1D ¹H NMR spectrum of nsp3a(1–112) before removal of co-purifying nucleic acids. Spectra were measured at 25 °C with water presaturation on a Bruker DRX700 spectrometer. 64 scans were accumulated. The presence of characteristic
nucleic acid signals in the area from 4.8 to 6.4 ppm (*) is readily apparent (1'H, 2'H, 3'H, 4'H, 5'H, 5''H of all nucleotides and pyrimidine 5H are typically observed in this spectral region). (b) 1D $^1$H NMR spectrum of the nucleic acid-free nsp3a(1–112) sample used for the NMR structure determination (see Materials and Methods). The weak peaks between 4.8 and 6.4 ppm are part of the protein spectrum. (c) Isolation of RNA that co-purified with nsp3a(1–112). The chromatogram was obtained after loading a sample of unfolded nsp3a(1–112) in 6 M guanidinium-HCl onto a size-exclusion column. Absorbance at 280 nm and conductivity are shown in blue and brown, respectively. The protein and ssRNA absorption peaks are labeled, the high conductivity observed after 320 minutes is due to guanidinium-HCl.

FIG. 11. Mass spectrum of the isolated ssRNA fragment. The proposed structures for the main peaks are presented together with their corresponding molecular weight and atom composition.

FIG. 12. (a) Superposition of the $[^{15}\text{N},^{1}\text{H}]-$HSQC spectra of nsp3a(1–112) in the absence (blue) and presence (red) of a four-fold excess of the exogenous ssRNA2 (see text). (b) Plot versus the amino acid sequence of the chemical shift changes in the backbone $^{1}\text{H}N^{15}\text{N}$ moieties of nsp3a(1–112) due to ssRNA2 binding. $\Delta\delta_{av}$ is a weighted average of the $^{1}\text{H}$ and $^{15}\text{N}$ chemical shift differences, $\Delta\delta_{av} = \{0.5[\Delta\delta^{1}\text{H}\text{N}^2] + (0.2\Delta\delta^{15}\text{N})^2]\}^{1/2}$, determined from comparison of the $[^{15}\text{N},^{1}\text{H}]-$HSQC spectra shown in panel (a). (c) Superposition of the $[^{15}\text{N},^{1}\text{H}]-$HSQC spectra of nsp3a(1–112) in the absence (blue) and presence (red) of a four-fold excess of octa-U.
FIG. 13. Comparison of nsp3a with Ras-interacting proteins. (a) In a complex consisting of a Ras dimer (grey) bound to two RID-RalGDS subunits (yellow) (PDB: ILFD), nsp3a(1–112) (red) is superimposed on one of the two RID subunits. The residues used for the superposition were identified using the software DALI with the NMR structure of nsp3a(1–112) and the X-ray structure of the Ras-RID-RalGDS complex (12). (nsp3a(1–112): residues 17–29, 33–37, 41–63, 83–87, 88–94, 95–98 and 101–108; RID-RalGDS: residues 14–26, 27–31, 32–54, 55–59, 63–69, 74–77 and 93–100). The C\text{\textalpha}-atoms of these residues could be superimposed with a RMSD of 2.3 Å. (b) Sequence alignment of a dodecapeptide containing the \beta1-strand of nsp3a (box) with the corresponding segments in some members of the Ras-interacting protein family, with the residue numbers of nsp3a indicated. (c) Electrostatic potential surfaces of nsp3a(1–112), RID-RalGDS and Ra-AF6. The positions of the conserved residues corresponding to R23 in nsp3a(1–112) are indicated. (d) Ribbon presentations of the same structures as in (c).
Figure 1

(a) Amino acid sequences of trimeric spike proteins from various sources.

(b) Expanded view of the ECD region with residues colored by secondary structure.
Figure 2
Figure 9

(a) Nsp3a (1–112) (100 μM to 1.6 μM)

(b) 800 ng ssRNA

Figure 9
Figure 10
Figure 11
Figure 12
Figure 13