Journal of Biomedical Science (2006) 13:59–72
DOI 10.1007/s11373-005-9035-9

Modular organization of SARS coronavirus nucleocapsid protein

Chung-ke Chang[1], Shih-Che Sue[1], Tsan-hung Yu[1], Chiu-Min Hsieh[1], Cheng-Kun Tsai[1,2], Yen-Chieh Chiang[3], Shin-jye Lee[1], Hsin-hao Hsiao[1], Wen-Jin Wu[1], Wei-Lun Chang[4], Chun-Hung Lin[4] & Tai-huang Huang[1,2,*]

[1]*Institute of Biomedical Sciences, Academia Sinica, Nankang, Taipei, Taiwan, ROC;* [2]*Department of Physics, National Taiwan Normal University, Taipei, Taiwan, ROC;* [3]*Department of Agronomy, National Taiwan University, Taipei, Taiwan, ROC;* [4]*Institute of Biological Chemistry, Academia Sinica, Nankang, Taipei, Taiwan, ROC*

Received 23 June 2005; accepted 12 September 2005

Key words: capsid protein, coronavirus, domain arrangement, intrinsically disordered protein, NMR, oligomerization, SARS

Abstract

The SARS-CoV nucleocapsid (N) protein is a major antigen in severe acute respiratory syndrome. It binds to the viral RNA genome and forms the ribonucleoprotein core. The SARS-CoV N protein has also been suggested to be involved in other important functions in the viral life cycle. Here we show that the N protein consists of two non-interacting structural domains, the N-terminal RNA-binding domain (RBD) (residues 45–181) and the C-terminal dimerization domain (residues 248–365) (DD), surrounded by flexible linkers. The C-terminal domain exists exclusively as a dimer in solution. The flexible linkers are intrinsically disordered and represent potential interaction sites with other protein and protein-RNA partners. Bioinformatics reveal that other coronavirus N proteins could share the same modular organization. This study provides information on the domain structure partition of SARS-CoV N protein and insights into the differing roles of structured and disordered regions in coronavirus nucleocapsid proteins.

Introduction

Coronaviruses are the causative agents of a number of mammalian diseases which often have significant economic and health-related consequences [1, 2]. Diseases such as transmissible gastroenteritis in pigs and avian infectious bronchitis in chicken often have great impact on the agricultural industry of a nation [3]. In humans, coronaviruses are often associated with mild respiratory illnesses, including common cold. However, a novel coronavirus has been identified as the etiology agent of severe acute respiratory syndrome (SARS), which has a case fatality rate of ca. 8% [4]. Sequence analysis reveals that SARS-CoV represents either a new coronavirus group or an outliner of group 2 coronaviruses [5–8].

The SARS CoV genome contains five major open reading frames that encode the replicase polyprotein, the spike protein (S), envelope (E), membrane glycoprotein (M), and the nucleocapsid protein (N). SARS-CoV is an enveloped virus with S, M and E proteins as the envelope proteins. The N protein binds to the viral RNA genome and forms the ribonucleoprotein core, which is presumed to be helical. The M protein may also be involved in the formation of the nucleocapsid through interaction with the N protein. Upon infection, the N protein enters the host cell with

*To whom correspondence should be addressed. Fax: +886-2-2788-7641; E-mail: bmthh@gate.sinica.edu.tw
CK Chang and SC Sue contributed equally to this project.

the ribonucleoprotein core and is able to interact with a number of host proteins [9]. The high abundance of the N protein makes it a major antigen, an attribute which has often been used in the development of rapid-diagnosis kits against SARS [10, 11].

The nucleocapsid protein is a 422 amino-acid protein, sharing only 20–30% homology with the N proteins of other coronaviruses [6, 7]. From genetic and bioinformatics studies, the N protein can be divided into three putative regions: an N-terminal domain, a RNA-binding domain (RBD) and a C-terminal domain [12, 13]. The N- and C-terminal domains are believed to play a role in interaction with other proteins. A number of recent studies have shown that part of the C-terminus in the N protein of SARS-CoV is involved in the oligomerization process of the protein [14, 15]. Rather surprising, the mid-portion of the protein has been shown to interact with the M protein and hnRNP A1 [16, 17], and structural studies have identified the region between amino acids 45–181 as the putative RNA-binding region, which is close to the N-terminus [18]. These discrepancies from the putative domain partition necessitate the determination of both the functional and structural organization of the protein. However, the structural organization of coronavirus N proteins in general remains largely unknown to this day.

We have employed a blend of experimental techniques and bioinformatics analyses to define the structural organization of SARS-CoV N protein. Through the power of nuclear magnetic resonance (NMR) spectroscopy, we present the first evidence that the SARS-CoV N protein consists of two independent structural domains. The first domain lies inside the putative RNA-binding domain identified in a previous report [18]. The second domain lies in the C-terminal half of the protein and is capable of forming dimers in solution. The rest of the protein is highly accessible to the solvent, and bioinformatics analysis predicts that they are intrinsically disordered. Other coronavirus N proteins share similar features of SARS-CoV N protein at the sequence level, implying functional significance. The elucidation of the modular organization of the SARS-CoV N protein, particularly the boundary between disordered and structured regions, facilitates future studies of this class of proteins at the functional and structural level.

Materials and methods

Sequence alignment, secondary structure and order–disorder prediction

The full-length sequences of SARS and other coronavirus N proteins were aligned using CLUSTALW version 1.83 with the slow algorithm, an identity matrix, a window of 4 amino acids and standard gap penalties [19]. The result was then edited with SeaView based on the position of the known structural domains of SARS-CoV N protein. The JPred server [20] was used for secondary structure prediction. Order–disorder prediction was obtained through sequence submission to the PONDR server (http://www.pondr.com) using the predictor VSL1, which is an implementation of the IST-Zoran predictor [21–23]. Access to PONDR® was provided by Molecular Kinetics (Indianapolis, IN, USA).

Plasmid construction

We cloned fragments spanning the different ordered and disordered regions of the SARS-CoV N protein (Figure 1B) based on PONDR information (Figure 1a) and reports in the literature [18]. SARS-CoV TW1 strain cDNA sequencing clones were kindly provided to us by Dr. P.-J. Chen of National Taiwan University Hospital [24]. Clones for SARS-CoV N protein fragments were obtained by polymerase chain reaction (PCR) on a RoboCycler Gradient 96 (Stratagene, CA) using appropriate primers. The resulting PCR fragments contained an *Nco*I site at one end and a *Bam*HI site at the other. After restriction enzyme digestion, the resulting fragments were cloned into pET6H (a gift from Prof. J.-J. Lin, National Yang Ming University, Taiwan) containing a His-tag coding region. Full-length SARS-CoV N protein construct was obtained by sequential ligation of the cloned PCR fragments using appropriate restriction enzyme sites. The sequences of all constructs were confirmed by DNA sequencing. The resultant protein fragments all include an extra MHHHHHHAMG sequence at the N-terminus.

Protein expression and purification

For biochemical studies, the SARS-CoV N protein clones were expressed in *Escherichia coli* BL21(DE3)

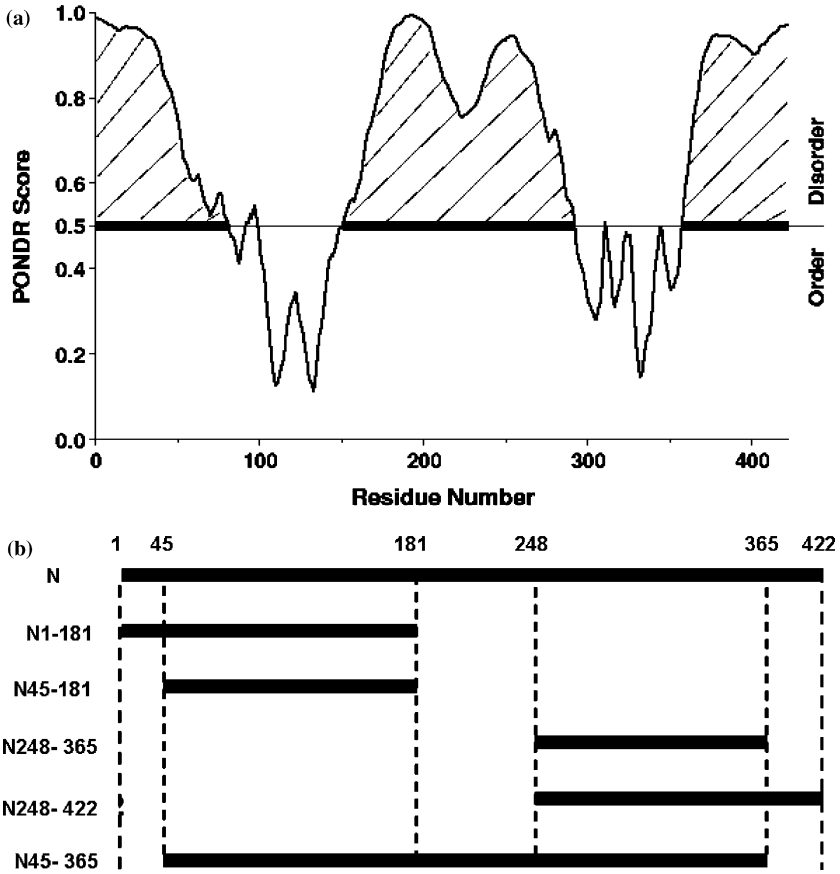

Figure 1. (a) SARS-CoV N protein fragments studied in this paper. Designations of the fragments are listed on the left. (b) PONDR prediction of the order–disorder regions of SARS-CoV N protein. Hatched regions represent PONDR scores higher than 0.5 and are considered disordered.

strain in Luria broth media using standard protocols. To prepare samples suitable for NMR studies, the cells were cultured in standard M9 media supplemented with $^{15}NH_4Cl$ (1 g/l) and ^{15}N-Isogro (0.5 g/l) (Isotec, OH, USA). The cells were then broken with a microfluidizer and the protein purified through a Ni-NTA affinity column (Qiagen, CA, USA) in buffer (50 mM sodium phosphate, 150 mM NaCl, pH 7.4) containing 7 M urea. The protein was then allowed to refold by gradually lowering the denaturant concentration through dialysis in liquid chromatography buffer (50 mM sodium phosphate, 150 mM NaCl, 1 mM EDTA, 0.01% NaN_3, pH 7.4). Renatured protein was loaded onto an AKTA-EXPLORER fast performance liquid chromatography (FPLC) system equipped with a HiLoad 16/60 Superdex 75 column (Amersham Pharmacia Biotech, Sweden). Complete Protease Inhibitor cocktail (Roche, Germany) was added to the purified protein. Protein concentration was determined with the Bio-Rad Protein Assay kit as per instructions from the manufacturer (Bio-Rad, CA, USA). The correct molecular weights of the expressed proteins were confirmed by mass spectroscopy.

Analytical gel-filtration chromatography

The experiments were conducted using a FPLC System (Pharmacia Biotech, Sweden) with a Hi-Load 16/60 Superdex 75 (prep grade) column at an elution rate of 1 ml/min. The molecular weights of the proteins were estimated from the elution profile calibrated with the LMW Gel Filtration Calibration Kit (Amersham, UK).

Chemical cross-linking

The homo-bifunctional amine cross-linker disuccinimidyl suberate was purchased from Sigma-Aldrich (MO, USA) and was dissolved in *N*,*N*-dimethylformamide (DMF) to a concentration of 25 mg/ml. Reactions were carried out in a final protein concentration of 0.35 mM and a final disuccinimidyl suberate concentration of 5 mM. Mock reactions were set up as controls which contained only the protein solution and DMF without disuccinimidyl suberate. The reaction mixtures in standard buffer were allowed to react for 1 h at 4 °C prior to quenching with 100 mM glycine (final concentration). The results were visualized on SDS-PhastGel minigels (Pharmacia Biotech, Sweden).

Sedimentation velocity analysis

Sedimentation velocity studies were carried out with a Beckman-Coulter XL-A analytical ultracentrifuge with an An60Ti rotor at 20 °C and 40,000 rpm. Protein samples were diluted to 0.40–0.75 mg/ml and loaded into standard double sector cells with aluminum or Epon charcoal-filled centerpieces. The UV absorption of the cells was scanned at 280 nm in continuous mode every 10 min for a period of 5 h. The data were analyzed with Sedfit version 8.9d. Collections of 10–15 radial scans were used for analysis, and 200 sedimentation coefficients between 2 and 10 S were employed in calculating the c(S) distribution. The positions of the meniscus and cell bottom were determined by visual inspection, and then refined in the final fit. The partial specific volumes for N45–181, N245–365 and N45–365 were calculated from the amino acid compositions to be 0.7192, 0.7244 and 0.7198 ml/g, respectively. The solvent density and viscosity were calculated with Sednterp version 1.08. All samples were visually checked for clarity after ultracentrifugation, and no indication of precipitation was found.

NMR spectroscopy

^{15}N-labeled protein samples were extensively exchanged with NMR buffer (100 mM sodium phosphate buffer, pH 6.0, containing 50 mM NaCl, 1 mM EDTA, 1 mM 2,2-dimethyl-2-silapentane-5-sulfonate, 0.01% NaN_3, 10% D_2O and Complete Protease Inhibitor cocktail) using an Amicon-15 concentrator (Amicon, MA, USA). The final concentrations of the samples were between 0.2 and 3 mM, depending on the solubility of the different fragments. All the NMR data were acquired at 27 and 30 °C on 500, 600 or 800 MHz Bruker AVANCE spectrometers equipped with a triple resonance (^{1}H, ^{13}C and ^{15}N) TXI probe with an actively shielded Z-gradient. Experimental parameters were set as described previously [25, 26]. CLEANEX-PM spectra, which only show resonances exchanging rapidly with the solvent ($k_{ex} > 2$ Hz), were obtained as described [27, 28]. Data were processed with the XWINNMR suite and AURELIA software (Bruker, Germany) on SGI workstations. The ^{1}H chemical shift was referenced to 2,2-dimethyl-2-silapentane-5-sulfonate at 0 ppm. The ^{15}N was referenced using the consensus ratio Ξ of 0.101329118 for ^{15}N/^{1}H [29].

Results

SARS-CoV N protein contains two independent structural domains

A series of N protein fragments spanning different regions were constructed based on the PONDR prediction (Figure 1). We used a series of ^{15}N-HSQC spectra of these fragments to define the position of the structural domains of SARS-CoV N protein (Figure 2). NMR chemical shifts of amide resonances are sensitive to structural changes and the pattern of ^{15}N-HSQC spectrum has been commonly used to monitor order–disorder of proteins [30]. Well-dispersed spectra are indicative of structured protein whilst congested spectra having resonances clustered around a small region of 8.3 ± 0.5 ppm in the proton dimension are disordered. We observed that the resonances from residues N45–181 have good chemical shift dispersion (Figure 2a), indicating that the fragment has a structured character. The spectrum of N1-181 is a superposition of well-dispersed resonances and a cluster of overlapping resonances around 8.3 ± 0.4 ppm (Figure 2b). Comparing the spectra of N1–181 and N45–181 revealed that all resonances belonging to N45–181 were present in the spectrum of N1–181 with no change in resonance position. These results indicate that the N-terminal flanking region between amino acids 1–44 does not affect the structure of the N45–181 domain.

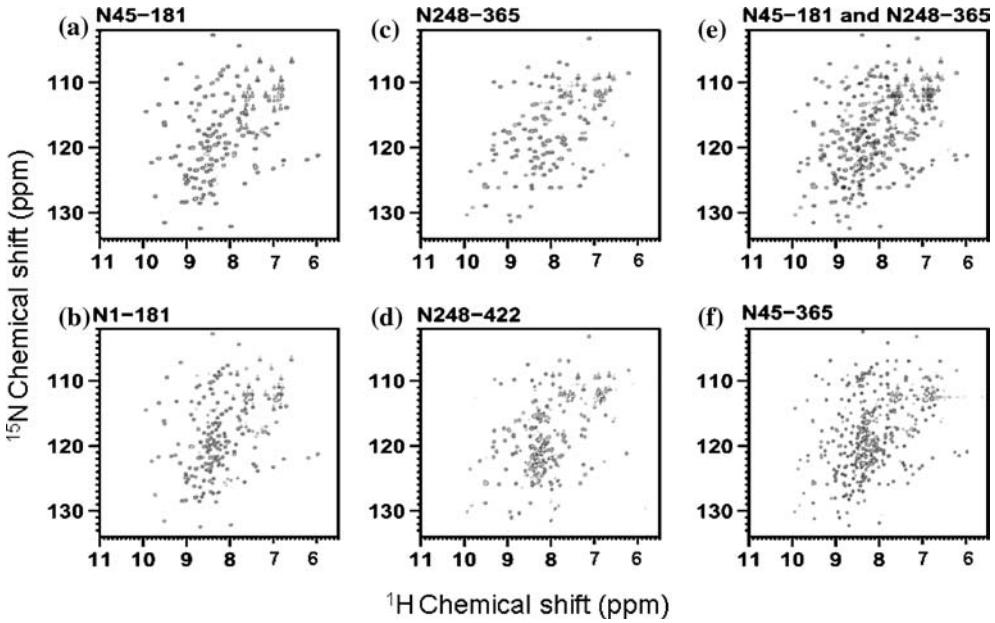

Figure 2. ^{15}N-HSQC spectra of the SARS-CoV N protein fragments. (a): u -^{15}N-N45–181. (b): u -^{15}N-N1–181. (c) u -^{15}N-N248–365. (d): u-(^{2}H, ^{15}N)-N248–422. (e): Overlay spectrum of u -^{15}N-N45–181 and u -^{15}N-N248–365. (f): u-(^{2}H, ^{15}N)-N45–365. All spectra were obtained on a Bruker AVANCE 800 MHz spectrometer at 27 °C. NMR sample contains 0.2–1 mM protein in 10 mM sodium phosphate buffer, 50 mM NaCl, 1 mM EDTA, 1 mM 2,2-dimethyl-2-silapentane-5-sulfonate, 0.01% NaN_3, pH 6.0 in 10% D_2O. The spectra shown at the bottom (b, d and f) are almost identical to those shown on the top (a, c and e) except that the bottom three spectra contain additional resonances in the 7.5–9 ppm (^{1}H dimension) region. These resonances arise from the additional disordered residues in the longer protein fragments.

To assess the structure of the C-terminal region several C-terminal fragments were prepared for the collection of ^{15}N-HSQC spectra. We found that the resonances from N248–365 are well-dispersed (Figure 2c), suggesting that N248–365 forms an ordered structure. To define the structural boundaries we constructed fragments containing N- and C-terminal extensions. Figure 2d shows the ^{15}N-HSQC spectrum of uniformly ^{15}N-labeled N248–422 sample. Comparing the spectrum of N248–422 with that of N248–365 (Figure 2c) we found that all resonances due to N248–365 can be identified in Figure 2d. These results indicate that residues from 365 to the C-terminal do not affect the structure of N248–365. Shortening the fragment to span amino acids 274–365 changes the ^{15}N-HSQC resonance pattern, which indicates that the 248–273 region is important for structure stabilization of this domain (data not shown).

To explore the structure of the region between residues 182–247 and their effect on the structure of N45–181 and N248–365, we constructed the fragment N45–365 which contains the two structured domains and the inter-domain residues. Comparing the ^{15}N-HSQC spectrum of N45–365 (Figure 2f) to that in Figure 2e, which is the overlay of the spectra from N45–181 (Figure 2a) and N248–365 (Figure 2c), we observed that the resonances from N45–181 and N248–365 overlap perfectly with the corresponding resonances from N45–365, indicating that the structures of the two domains, N45–181 and N248–365, are not altered in N45–365. The lack of resonance perturbation when the two domains are linked together suggests that interaction between these two domains is weak, if they interact at all. Our results conclude that SARS-CoV N protein contains two independent structural domains located at a.a. 45–181 and 248–365. These results are consistent with PONDR prediction.

SARS-CoV N protein contains three intrinsically disordered regions

PONDR predicts three intrinsically disordered regions in SARS-CoV N protein located at the N-terminus, the C-terminus and between the two

ordered regions (Figure 1b). We also observed additional resonances clustered around 8.3 ± 0.5 ppm in the proton dimension whenever the fragment was extended beyond the two structural domains (Figure 2). To test whether the residues beyond the structural domains are truly disordered, we employed the CLEANEX-PM experiment to identify solvent-accessible resonances [27]. The ^{15}N-HSQC spectrum obtained with CLEANEX-PM pulse sequence contains only resonances from solvent-exposed amide groups. When we compared the CLEANEX-PM spectrum of N1-181 (Figure 3b) with that of N45–181 (Figure 3a), we observed 40 resonances that only appeared in N1–181 but not in N45–181. This number agrees with that expected for the N-terminal region (5 prolines), indicating that all amide protons in the N-terminus of SARS-CoV N protein are exposed to the solvent. We counted 39 additional peaks in the CLEANEX-PM spectrum of N248–422 (Figure 3d) compared to that of N248–365 (Figure 3c) (51 expected since there are 6 prolines), suggesting that the majority of the C-terminal residues are also solvent-exposed. When we compared the CLEANEX-PM spectra of N45–181 (Figure 3a), N248–365 (Figure 3c) and N45–365 (Figure 3f), we observed the extra resonances representing the region between residues 182–247. A total of 27 additional peaks can be resolved, compared to 64 expected (2 prolines), indicating that about half of the linker region between residues 182–247 is exposed to the solvent. It should be noted here that due to resonance overlapping the numbers counted should be viewed as a lower limit for the number of solvent-exposed residues. Nevertheless we can conclude that all N-terminal residues are solvent exposed whilst most of the residues in the C-terminus and in the linker region between the two structural domains are exposed to the solvent as well. In conjunction with the observation that all additional resonances are observed in between 8.3 ± 0.5 ppm in the proton dimension

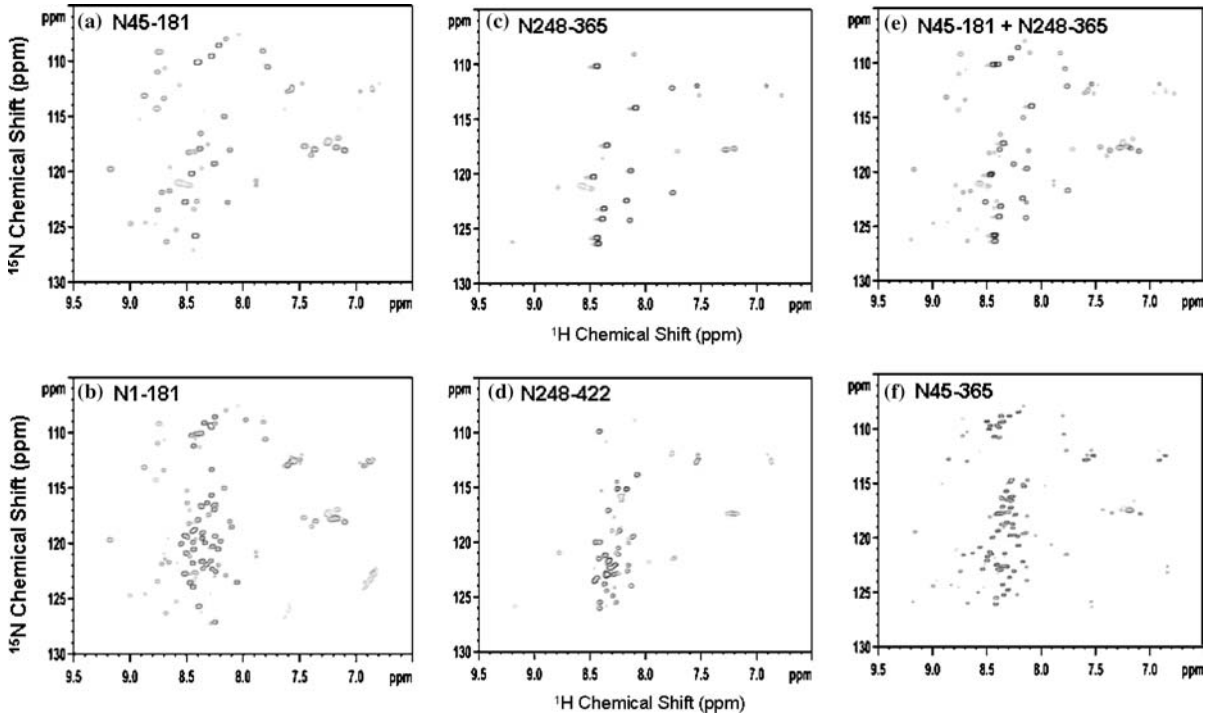

Figure 3. ^{15}N-edited CLEANEX-PM spectra of (a) u-^{15}N-N45–181, (b) u-^{15}N-N1–181, (c) u-^{15}N-N248–365, (d) u-(^{2}H, ^{15}N)-N248–422, (e) u-^{15}N-N45–181 and u-^{15}N-N248–365 overlaid on each other and (F) u-(^{2}H, ^{15}N)-N45–365. Spectra were obtained on a Bruker AVANCE 800 MHz spectrometer at pH 6.0 and 27 °C. The extra resonances are mostly clustered between 7.5 and 9 ppm in the ^{1}H-dimension. The numbers of resonances are much larger in the spectra (b), (d), and (f) that contain the disordered regions.

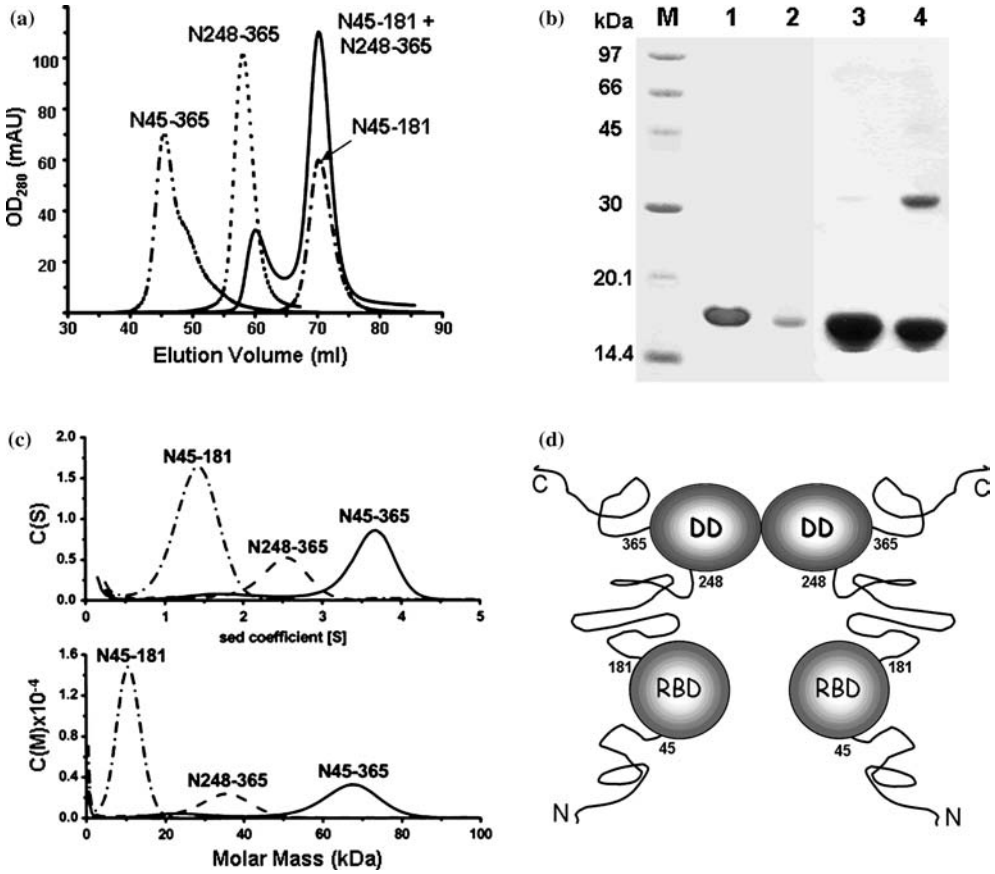

Figure 4. (a) Analytical gel-filtration chromatography of SARS-CoV N protein fragments. Fragments employed for obtaining the traces are indicated. (b) SDS-PAGE results of N45–181 (Lanes 1 and 2) and N248–365 (Lanes 3 and 4). M: molecular weight marker. Lanes 1 and 3 are mock reactions without disuccinimidyl suberate. The corresponding gel traces of N248–365 and N45–181 after reacting with disuccinimidyl suberate for 1 h at 4 °C are shown on lanes 2 and 4, respectively. (c) Sedimentation velocity studies of SARS-CoV N protein fragments. The distribution of the sedimentation coefficient (top) and molecular mass (bottom) of SARS-CoV N protein fragments N45–181, N248–365 and N45–365. (d) A model of the overall structure of SARS-CoV N protein. The two solids represent the two structural domains, the RNA-binding domain (RBD) and the dimerization domain (DD). The wavy lines represent disordered segments.

and PONDR results, we conclude that amino acids 1–44, 182–247 and 366–422 are disordered. The long disordered linker between the two structural domains is consistent with the observation that there is little interaction between the two domains. However, the number of counted peaks in the CLEANEX-PM spectra of the C-terminus and the linker region are less than that expected, so it is likely that parts of these regions are solvent-protected, possibly through the formation of transient structures. Attempt to obtain a spectrum of the linker region alone was unsuccessful due to the extremely poor protein expression of the clone harboring the linker sequence.

The C-terminal structural domain is sufficient for dimerization

N45–181 has been identified as an RNA-binding domain. The function of the N248–365 is not clear, but many reports have identified the C-terminal half of SARS-CoV N protein to be involved in oligomerization [14, 15]. To test this possibility, we have applied analytical gel-filtration chromatography, chemical cross-linking and analytical ultracentrifugation to assay the self-association property of the N protein fragments. As shown in Figure 4a, N45–181 elutes out at a molecular weight of 18 kDa and N248-365 elutes out as a 28-kDa molecule, suggesting that N45–181 exists

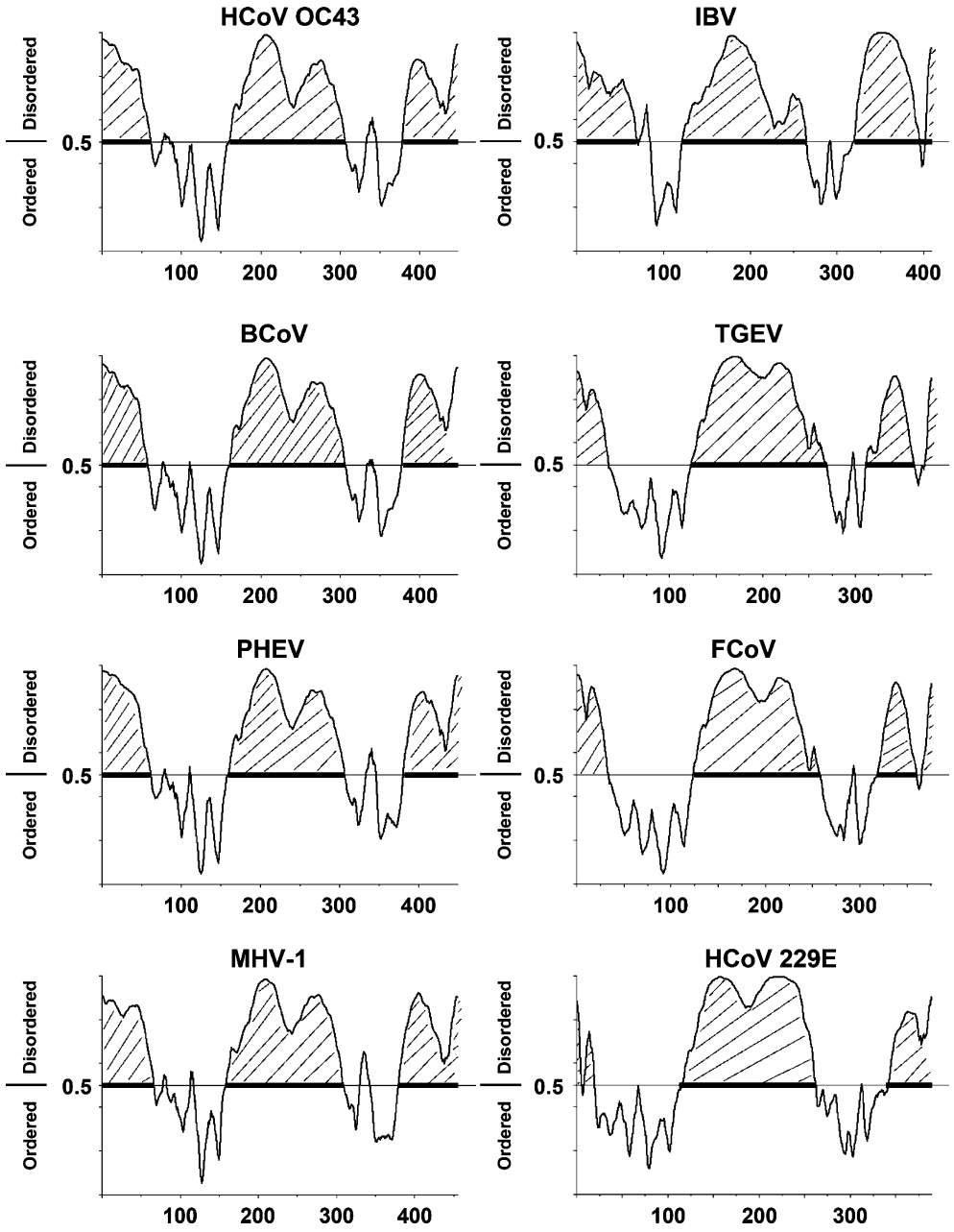

Figure 5. Order–disorder prediction of coronavirus N proteins through the PONDR server. Swiss-Prot/TrEMBL accession codes are included in parentheses. Hatched regions represent disordered segments. HCoV OC43: Human coronavirus strain OC43 (P33469); BCoV: Bovine coronavirus strain Quebec (P59712); PHEV: Porcine hemagglutinating encephalomyelitis virus (Q8BB23); MHV-1: Mouse hepatitis virus (P18446); IBV: Avian infectious bronchitis virus (P32923); TGEV: Porcine transmissible gastroenteritis virus (P05991); FCoV: Feline coronavirus (O12298); HCoV 229E: Human coronavirus strain 229E (P15130). All coronavirus N proteins in this study share the same order–disorder profile.

as a monomer and N248–365 exists as a dimer. The self-association between the two N248–365 monomers is very strong, since we could not detect any monomeric fraction. Similarly, N45–365 eluted out at molecular weight of ~70 kDa, suggesting that N45–365 also exists as a dimer. Furthermore, when N45–181 sample was mixed with N248–365 sample two peaks at 18 and 28 kDa were observed in the elution profile, demonstrating that the two fragments do not interact with each other.

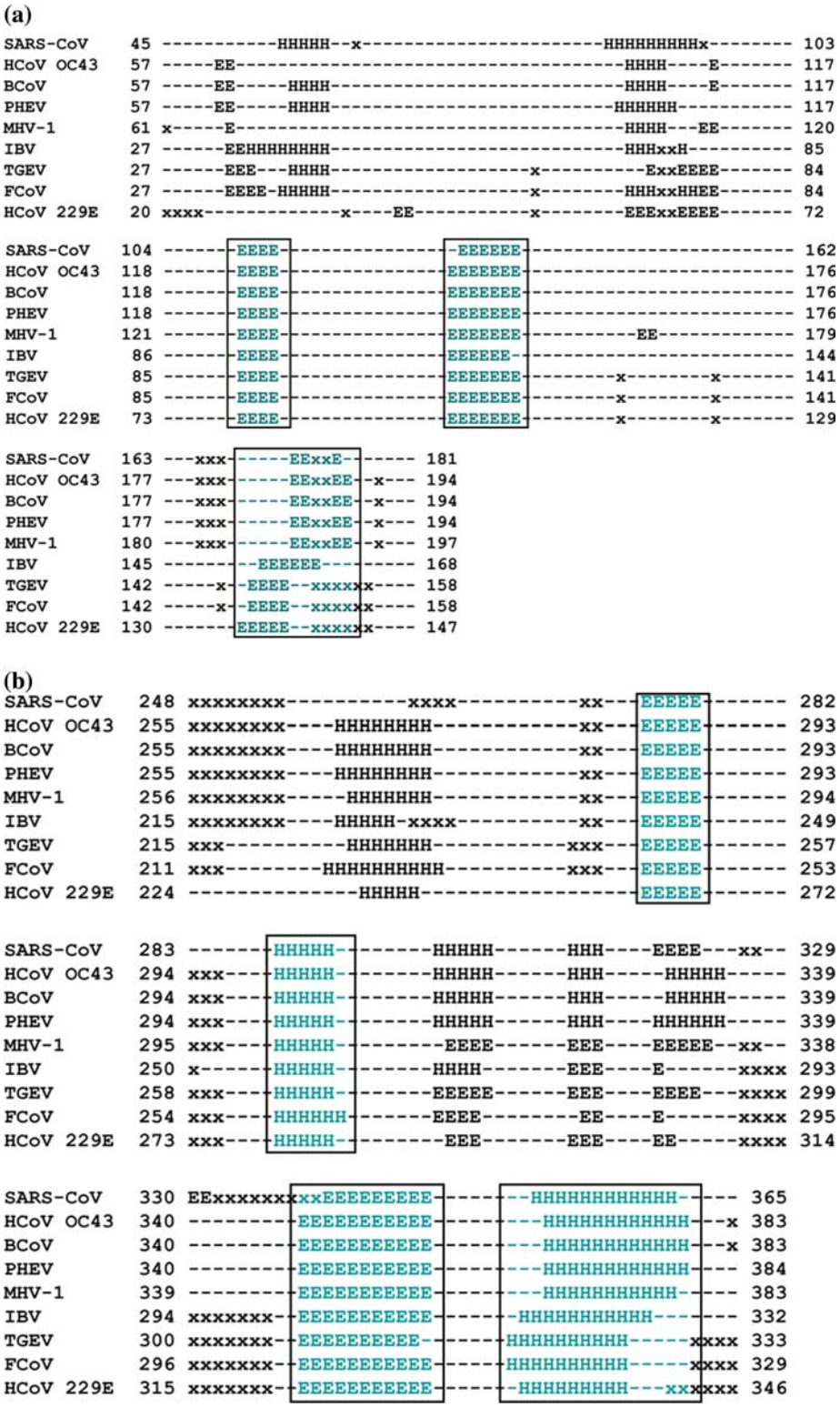

Figure 6. Predicted secondary structure alignment of the putative structural domains of coronavirus N proteins. Virus denotations are the same as in Figure 5. Residue numbers are listed. H: α-helix; E: extended β-strand. (a) Alignment of the N-terminal domain. The three conserved β-strands are enclosed. (b) Alignment of the C-terminal domain. Conserved secondary structure elements are enclosed.

Cross-linking experiments shown in Figure 4b detected the presence of only monomer for N45–181 and both monomer and dimer for N248–365.

The quaternary structures of N45–181, N248–365 and N45–365 fragments were further examined by analytical ultracentrifugation. Only one major peak was detected for each of these three protein fragments, indicating that they are structurally homogeneous in solution. The results of data analysis with Sedfit version 8.9d showed that protein fragments N45–181, N248–365 and N45–365 sediment at 1.4 S, 2.6 S and 3.7 S (Figure 4c), corresponding to a molecular mass of 10, 36 and 68 kDa, respectively. These results confirmed that N45–181, N248–365 and N45–365 exist as a monomer, dimer and dimer, respectively, in agreement with the results of gel-filtration chromatography and chemical cross-linking. Taking together all three results indicate that N45–181 exists as a monomer and N248–365 as a dimer. The fact that dimerization occurs through a structural domain strongly suggest that the process is dependent on the structure. A model of the SARS-CoV N protein interaction based on our current results is shown in Figure 4d. It is interesting to note that we did not observe the formation of higher-order multimer in our studies, which may be important for the formation of the ribonucleoprotein complex within the virion. A possible explanation is that multimer formation may require additional factors, such as the presence of RNA or other parts of the N protein that were not present in our samples. Also we can not exclude the possibility that multimers do form at much higher protein concentrations than the ones used in these studies. We suggest that the dimeric form represents a basic building block of the nucleocapsid of SARS-CoV.

Order–disorder profiles are conserved among coronavirus N proteins

Since coronavirus N proteins belong to the same protein family, it is probable that they share similar structural features. Comparison of the order–disorder profile of these proteins (Figure 5) shows that they all share the same disordered regions (hatched regions). There are two long disordered regions in the middle and at the C-termini of the proteins, whereas the length of the N-terminal disordered region shows more variability. Two ordered regions are located between the disordered regions, and their locations generally match those of the structural domains in SARS-CoV N protein.

Disordered regions are often involved in biomolecular interactions. The C-terminus of MHV N protein, which is disordered, has been shown to interact with hnRNP A1 [31], whereas the disordered region in the middle is responsible for its RNA-binding activity [13, 32]. In SARS-CoV, the disordered region in the middle of the N protein has been implicated in N-protein self-interaction [33], interaction with the M protein [16] and hnRNP A1 interaction [17]. These experimental observations suggest that disordered regions of coronavirus N proteins are probable interaction sites with functional implications.

Ordered regions of coronavirus n proteins share similar secondary structure profiles

Secondary structure alignment of coronavirus N protein sequences based on the two structural domains of SARS-CoV N protein show that they share very similar secondary structure profiles (Figure 6). The N-terminal domain has three conserved β strands which have been implicated in RNA binding in SARS-CoV [18]. The C-terminal domain is also mostly conserved in terms of secondary structure position within the sequence. The extensive secondary structure and high similarity suggests that the two structural domains observed in SARS-CoV N protein also exist in the N proteins of other coronaviruses.

The results from the order–disorder prediction and secondary structure prediction coupled with sequence alignment suggest that coronavirus N proteins all share the same modular organization. The two structural domains are connected by a disordered linker and capped by disordered N-terminal head and C-terminal tail.

Discussion

Role of the structural domains of SARS-CoV N protein

The two structural domains of SARS-CoV N protein carry out two distinct functions. The

N-terminal domain is able to bind RNA, whereas the C-terminal domain acts as a dimerization domain. The ability of the N-terminal domain to bind RNA is closely related to its structure. Although the structure of the C-terminal domain has not been determined, we suggest that dimerization is also structure-dependent. A number of experimental observations support our hypothesis: First, it has been found that oligomer dissociation and protein unfolding of SARS-CoV N protein occur simultaneously [34]; second, most self-interaction studies have mapped the oligomerization domain to regions containing the structural domain [14, 15]. The structural domains may also serve additional functions. For example, a putative loop between W302 and P310 in the C-terminal domain has been suggested to bind to cyclophilin A [35]. These additional functions may also be dependent on the structure of the protein.

Although the two structural domains do not interact with each other, we cannot discount the possibility that the two domains could act in concert to carry out important biological functions. The long flexible linker between the two domains provides enough freedom to make this scenario possible. Previously, the lack of information on structural organization precluded the study of multiple-domain interactions. Now our findings provide a structural framework to perform such studies.

The flexible linker as an interaction hotspot

The flexible linker between the two structural domains is largely disordered. This disordered region may enable transient interactions with several structurally distinct partners. It has been shown that the M protein of SARS-CoV binds to this region between a.a. 168–208 [16]. Interestingly, human cellular hnRNP A1 has also been shown to bind to almost the same region between a.a. 161–210 [17]. The disordered state of this region potentially allows it to interact with different partners depending on context, e.g. with the M protein during virus assembly and with hnRNP A1 during host cell infection. The exact mechanism by which this occurs is not known, but it could involve different induced folding pathways, which has been shown to occur in other disordered proteins [23, 36, 37].

The same phenomenon is observed in other coronavirus N proteins. In mouse hepatitis virus (MHV), the region corresponding to the flexible linker in its N protein is involved in RNA binding [13, 32]. The same region has also been shown to bind murine hnRNP A1 in infected cells [31]. It seems that the coronavirus N proteins share the common theme of using the flexible linker as an interaction "hotspot", and use characteristics of disordered regions to achieve multiple functions within a limited sequence length.

Disordered regions are potential phosphorylation sites

Phosphorylation is one of the most important regulatory post-translational modification in proteins. SARS-CoV N protein has been shown to get serine-phosphorylated by multiple kinases and phosphorylation is proposed to be a possible mechanism for nucleocytoplasmic shuttling of the N protein [38]. Disordered regions represent potential sites for phosphorylation. The flexible linker of SARS-CoV N protein contains an SR-rich region, which is targeted by a number of kinases [39]. In fact, this region can be phosphorylated in vitro (Dr. W.-Y. Tarn, personal communication). Recent *in silico* prediction suggested that most of the potential phosphorylation sites fall in the disordered regions, although the exact phosphorylations sites have not been identified experimentally [38]. Although the exact role of phosphorylation has not been elucidated, it could be related to regulate functions such as RNA-binding and localization within the host cell.

The phosphorylation patterns of other coronavirus N proteins which have been studied also fall in the disordered regions. In avian infections bronchitis virus (IBV), the phosphorylation sites of the N protein have been mapped to a.a. 186–198 and 367–394 [40]. These two regions are all located in the disordered region as predicted by PONDR (Figure 5). Phosphorylation of transmissible gastroenteritis virus (TGEV) N protein has also been mapped to residues 9, 156, 254 and 256, which are at or close to the disordered regions [41]. Phosphorylation in disordered regions of structural proteins is also observed in other virus families, such as in *Paramyxovirinae* [42]. Coronavirus N proteins seem to employ a widespread property to

allow for modification. Whether or not such modification affects the folding or structural properties of the protein and how these properties affect its function remain to be determined.

Implications for structural and functional studies of coronaviral N proteins

Identification of the disordered regions of SARS-CoV N protein provides a blueprint for structural studies of the protein. The structural domains are logical candidates for structural determination through X-ray crystallography or solution NMR studies. However, structure determination of the full-length protein is hindered by the disordered regions, which often interfere with crystallization [43]. The large size of the dimeric protein (ca. 90 kDa) also makes full-length structure determination through NMR extremely difficult due to T2 issues. The fact that the two structural domains do not interact provides a handle to solve this problem. The two structural domains can be solved independently and still provide fair representation of the full-length protein.

The modular organization of SARS-CoV N protein is shared among other coronavirus. The relative positions of the two structural domains are fairly conserved in all coronavirus N proteins, making them excellent targets for comparative structural studies. The structures of the N-terminal domains would be of special interest since in SARS-CoV it has been identified as an RNA-binding domain, whereas in other coronaviruses the exact function is not yet known. Of special note is the RNA-binding domain of MHV, which has been mapped to the flexible linker region instead of the N-terminal structural domain. At present the molecular mechanism involving N protein/RNA interaction is still not fully understood and the RNA binding site(s) have not been unequivocally defined. It is possible that the N-terminal structural domain folds into different tertiary structures and plays different roles in different coronavirus N proteins. It is also possible that the linker region may also be involved in RNA binding. Another interesting point that needs further study is the role of the C-terminal structural domain. It is not yet known whether it plays the same dimerization role in other coronavirus as in SARS-CoV, although there are hints in the literature [44].

In summary, we have the following conclusions: (1) The N protein of SARS-CoV is a di-domain protein connected by a flexible linker. The protein is capped by disordered N-terminal head and C-terminal tail. (2) The C-terminal structural domain is sufficient for dimerization, implying a structural role in the process. (3) Based on findings by other groups and our structural data, disordered regions of SARS-CoV N protein are potentially important interaction sites with functional implications. However, the exact roles of the disordered regions are yet to be defined. (4) The modular organization of SARS-CoV N protein is likely shared by the N proteins of other coronavirus. Our conclusions open up new venues for the study of coronavirus N proteins on a domain basis, including the study of complex interactions involving the different domains.

Acknowledgements

This work was supported in part by the Academia Sinica and by grants (NSC 92-2113-M-001-056 and NSC 92-2751-B-001-020-Y) (to THH) from the National Science Council of the Republic of China. The NMR spectra were obtained at the High-field Biomacromolecular NMR Core Facility, National Research Program for Genomic Medicine (NRPGM), Taiwan, Republic of China.

References

1. Lai M.M. and Cavanagh D., The molecular biology of coronaviruses. Adv. Virus Res. 48: 1–100, 1997.
2. Strauss J.H. and Strauss E.G. Viruses and Human Diseases. Academic Press, San Diego, 2002.
3. Mullan B.P., Davies G.T. and Cutler R.S., Simulation of the economic impact of transmissible gastroenteritis on commercial pig production in Australia. Aust. Vet. J. 71: 151–154, 1994.
4. Lai M.M.C., SARS virus: the beginning of the unraveling of a new coronavirus. J. Biomed. Sci. 10: 664–675, 2003.
5. Gorbalenya A.E., Snijder E.J. and Spaan W.J., Severe acute respiratory syndrome coronavirus phylogeny: toward consensus. J. Virol. 78: 7863–7866, 2004.
6. Rota P.A. *et al.*, Characterization of a novel coronavirus associated with severe acute respiratory syndrome. Science 300: 1394–1399, 2003.
7. Marra M.A. *et al.*, The genome sequence of the SARS-associated coronavirus. Science 300: 1399–1404, 2003.
8. Snijder E.J. *et al.*, Unique and conserved features of genome and proteome of SARS-coronavirus, an early

split-off from the coronavirus group 2 lineage. J. Mol. Biol. 331: 991–1004, 2003.
9. Ng M.L., Tan S.H., See E.E., Ooi E.E. and Ling A.E., Early events of SARS coronavirus infection in vero cells. J. Med. Virol. 71: 323–331, 2003.
10. He Q. *et al.*, Development of a Western blot assay for detection of antibodies against coronavirus causing severe acute respiratory syndrome. Clin. Diagn. Lab. Immunol. 11: 417–422, 2004.
11. Lin Y. *et al.*, Identification of an epitope of SARS-coronavirus nucleocapsid protein. Cell Res. 13: 141–145, 2003.
12. Parker M.M. and Masters P.S., Sequence comparison of the N genes of five strains of the coronavirus mouse hepatitis virus suggests a three domain structure for the nucleocapsid protein. Virology 179: 463–468, 1990.
13. Masters P.S., Localization of an RNA-binding domain in the nucleocapsid protein of the coronavirus mouse hepatitis virus. Arch. Virol. 125: 141–160, 1992.
14. Surjit M., Liu B., Kumar P., Chow V.T. and Lal S.K., The nucleocapsid protein of the SARS coronavirus is capable of self-association through a C-terminal 209 amino acid interaction domain. Biochem. Biophys. Res. Commun. 317: 1030–1036, 2004.
15. Yu I.M., Gustafson C.L., Diao J, Burgner J.W. II, Li Z., Zhang J. and Chen J., Recombinant Severe Acute Respiratory Syndrome (SARS) Coronavirus Nucleocapsid protein forms a dimer through its C-terminal domain. J. Biol. Chem. 280: 23280–23286, 2005.
16. He R. *et al.*, Characterization of protein–protein interactions between the nucleocapsid protein and membrane protein of the SARS coronavirus. Virus Res. 105: 121–125, 2004.
17. Luo H., Chen Q., Chen J., Chen K., Shen X. and Jiang H., The nucleocapsid protein of SARS coronavirus has a high binding affinity to the human cellular heterogeneous nuclear ribonucleoprotein A1. FEBS Lett. 579: 2623–2628, 2005.
18. Huang Q. *et al.*, Structure of the N-terminal RNA-binding domain of the SARS CoV nucleocapsid protein. Biochemistry 43: 6059–6063, 2004.
19. Thompson J., Gibson T., Plewniak F., Jeanmougin F. and Higgins D., The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 25: 4876–4882, 1997.
20. Cuff J., Clamp M., Siddiqui A., Finlay M. and Barton G., JPred: a consensus secondary structure prediction server. Bioinformatics 14: 892–893, 1998.
21. Romero P., Obradovic Z. and Dunker A.K., Identifying disordered regions in proteins from amino acid sequences. Proc. I.E.E.E. Int. Conf. Neural Networks 1: 90–95, 1997.
22. Iakoucheva L.M. *et al.*, Identification of intrinsic order and disorder in the DNA repair protein XPA. Protein Sci. 10: 560–571, 2001.
23. Iakoucheva L.M., Brown C.J., Lawson J.D., Obradovic Z. and Dunker A.K., Intrinsic disorder in cell-signaling and cancer-associated proteins. J. Mol. Biol. 323: 573–584, 2002.
24. Yeh S.-H. *et al.*, Characterization of severe acute respiratory syndrome coronavirus genomes in Taiwan: molecular epidemiology and genome evolution. PNAS 101: 2542–2547, 2004.
25. Sue S.C., Chang J.Y., Lee S.C., Wu W.G. and Huang T.-h., Solution structure and heparin binding site of Hepatoma-derived growth factors. J. Mol. Biol. 343: 1365–1377, 2004.
26. Lin T.H., Chen C.P., Huang R.F., Lee Y.L., Shaw J.F. and Huang T.H., Multinuclear NMR resonance assignments and the secondary structure of Escherichia coli thioesterase/protease I: A member of a new subclass of lipolytic enzymes. J. Biomol. NMR 11: 363–380, 1998.
27. Hwang T.L., ven Zijl P.C.M. and Mori S., Accurate quantitation of water-amide exchange rates using the phase-modulated CLEAN chemical exchange (CLEANEX-PM) approach with fast-HSQC (FHSQC) detection scheme. J. Biomol. NMR 11: 221–226, 1998.
28. Hwang T.L. and Shaka A.J., Multiple-pulse mixing sequences that selectively enhance chemical exchange or cross-relaxation peaks in high-resolution NMR spectra. J. Magn. Reson. 135: 280–287, 1998.
29. Wishart D.S., Bigam C.G., Yao J., Abildgaard F., Dyson H.J., Oldfield E., Markley J.L. and Sykes B.D., ^{1}H, ^{13}C, ^{15}N chemical shift referencing in biomolecular NMR. J. Biomol. NMR 6: 135–140, 1995.
30. Garcia P., Serrano L., Rico M. and Bruix M., An NMR view of the folding process of a CheY mutant at the residue level. Structure (Camb) 10: 1173–1185, 2002.
31. Wang Y. and Zhang X., The nucleocapsid protein of coronavirus mouse hepatitis virus interacts with the cellular heterogeneous nuclear ribonucleoprotein A1 in vitro and in vivo. Virology 265: 96–109, 1999.
32. Nelson G.W. and Stohlman S.A., Localization of the RNA-binding domain of mouse hepatitis virus nucleocapsid protein. J. Gen. Virol. 74(Pt 9) 1975–1979, 1993.
33. He R. *et al.*, Analysis of multimerization of the SARS coronavirus nucleocapsid protein. Biochem. Biophys. Res. Commun. 316: 476–483, 2004.
34. Luo H. *et al.*, In vitro biochemical and thermodynamic characterization of nucleocapsid protein of SARS. Biophys. Chem. 112: 15–25, 2004.
35. Luo C. *et al.*, Nucleocapsid protein of SARS coronavirus tightly binds to human cyclophilin A. Biochem. Biophys. Res. Commun. 321: 557–565, 2004.
36. Romero P., Obradovic Z., Li X., Garner E.C., Brown C.J. and Dunker A.K., Sequence complexity of disordered protein. PROTEINS: Struct. Funct. Gen. 42: 38–48, 2001.
37. Dyson H.J. and Wright P.E., Coupling of folding and binding for unstructured proteins. Curr. Opin. Struct. Biol. 12: 54–60, 2002.
38. Surjit M., Kumar R., Mishra R.N., Reddy M.K., Chow V.T.K. and Lal S.K., The Severe Acute Respiratory Syndrome Coronavirus Nucleocapsid Protein is phosphorylated and localizes in the cytoplasm by 14-3-3-mediated translocation. J. Virol. 79: 11476–11486, 2005.
39. Yeakley J.M., Tronchere H., Olesen J., Dyck J.A., Wang H.Y. and Fu X.D., Phosphorylation regulates *in vivo* interaction and molecular targeting of serine/arginine-rich pre-mRNA splicing factors. J. Cell Biol. 145: 447–455, 1999.
40. Chen H., Gill A., Dove B.K., Emmett S.R., Kemp C.F., Ritchie M.A., Dee M. and Hiscox J.A., Mass spectroscopic characterization of the coronavirus infectious bronchitis virus nucleoprotein and elucidation of the role of phosphorylation in RNA binding by using surface plasmon resonance. J. Virol. 79: 1164–1179, 2005.
41. Calvo E., Escors D., Lopez J.A., Gonzalez J.M., Alvarez A., Arza E. and Enjuanes L., Phosphorylation and subcel-

lular localization of transmissible gastroenteritis virus nucleocapsid protein in infected cells. J. Gen. Virol. 86: 2255–2267, 2005.
42. Karlin D., Ferron F., Canard B. and Longhi S., Structural disorder and modular organization in Paramyxovirinae N and P. J. Gen. Virol. 84: 3239–3252, 2003.
43. Dale G.E., Oefner C. and D'Arcy A., The protein as a variable in protein crystallization. J. Struct. Biol. 142: 88–97, 2003.
44. Tang T.K. *et al.*, Biochemical and immunological studies of nucleocapsid proteins of severe acute respiratory syndrome and 229E human coronaviruses. Proteomics 5: 925–937, 2005.