Functional and structural studies of Coronavirus ribonucleocapsid assembly

Fan, Hui

2007

Fan, H. (2007). Functional and structural studies of Coronavirus ribonucleocapsid assembly. Doctoral thesis, Nanyang Technological University, Singapore.

https://hdl.handle.net/10356/6573

https://doi.org/10.32657/10356/6573

Nanyang Technological University

FUNCTIONAL AND STRUCTURAL STUDIES OF CORONAVIRUS RIBONUCLEOCAPSID ASSEMBLY

FAN HUI

SCHOOL OF BIOLOGICAL SCIENCES

NANYANG TECHNOLOGICAL UNIVERSITY

2006

Acknowledgement

Firstly, I would like to thank my supervisor Dr. Julien Lescar for his more than three year guidance, encouragement and support to me and my project. Secondly, I would like to thank those people at Institute of Molecular and Cell Biology (IMCB) who used to help me and my project, including Dr. Ding Xiang Liu, Dr. Terje Dokland and Ms Sifang Wang for their contribution to this project. Thirdly, I would like to thank all the people in our lab, including Dr. Amy Ooi, ManoRavi, Mo Min, Lee Hooi Chen, Xu Ting and Yap Thai Leong, for their friendly support for me and my project. Additionally, I would like to thank my parents and Ms Cao Shenglan for their constant backing on me. Finally, I am full of gratitude to Singapore government to establish the School of Biological Sciences at Nanyang Technological University allowing the scholars and talented students from all over the world to gather and push forward the development of biological sciences.

Table of contents

Acknowledgement 1
Table of contents 2
Summary 4
List of figures 5
List of tables 7
Publication: 7
Abbreviations (alphabetically) 8
CHAPTER 1 INTRODUCTION 9
1. 1 Virus 9
1. 2. Coronavirus 11
1.2.1 Classification 12
1.2.2 Morphology and structure 14
1.2.3 RNA genome 15
1.2.4 Virus life cycle 18
1.3. Non-structural proteins in coronavirus 24
1.3.1 Precursor polyprotein 1a and 1ab 24
1.3.2 Papain-like protease 26
1.3.3 3C-like protease 28
1.3.4 Helicase 29
1.3.5 RNA-dependent-RNA-polymerase 29
1.3.6 Other non-structural proteins 30
1.4. Accessory proteins in coronavirus 34
1.5. Structural proteins in coronavirus 36
1.5.1 Spike (S) protein 36
1.5.2 Hemagglutinin-Esterase glycoprotein (HE) protein 38
1.5.3 Envelope (E) protein 39
1.5.4 Membrane (M) protein 40
1.5.5 Nucleocapsid protein (N) 42
1.6 Viral nucleocapsid architecture 44
1.7 RNA-binding proteins 50
1.7.1 The RNP motif 51
1.7.2 The arginine-rich motif (ARM) 53
1.7.3 The RGG box 53
1.7.4 The KH motif 54
1.7.5 The double-stranded RNA-binding motif (DSRM) 55
1.7.6 Zinc knuckle motif 56
1.8. Aims of research 57
CHAPTER 2. MATERIALS AND METHODS. 58
2.1 Cloning and expression of the IBV N protein 58
2.2 Purification of the IBV N protein 58
2.3 Dynamic light scattering 59

2.4 Negative staining electron microscopy of the IBV N protein..................... 60
2.5 Analysis of the IBV N protein and its two proteolytically stable fragments (the N-terminal domain IBV-N29-160 and the C-terminal domain IBV-N218-329) 61
2.6 Cloning and expression of IBV-N29-160 and IBV-N218-329................... 61
2.7 Size exclusion chromatography of the IBV N protein, IBV N29-160 and IBV N218-329... 62
2.8 Crosslinking experiments.. 63
2.9 RNA binding assay 63
2.10 Crystallization of IBV N29-160 and IBV N218-329............................... 64
2.11 Data Collection, Structure Determination, and Refinement 64
CHAPTER 3. RESULTS AND DISCUSSIONS.. 67
3.1 The IBV N protein 67
3.1.1 Expression and purification 67
3.1.2 Mass spectrometry 68
3.1.3 Dynamic Light Scattering (DLS)... 69
3.1.4 Electron microscopy 70
3.1.5 Size exclusion chromatography 72
3.1.6 Crosslinking 74
3.1.7 Discussion 75
3.2 Structure determination of the IBV N protein 76
3.2.1 Structural organization of the IBV N protein................................... 76
3.2.2 Identification of proteolytically stable N-terminal domain 80
3.2.3 Structure determination of the N-terminal domain and quality of the model... 83
3.2.4 Overall Structure of the N-terminal domain……………………… .89
3.2.5 Non-crystallography dimerization of the IBV N-terminal domain... 94
3.2.6 Identification of a proteolytically stable C-terminal domain............ 97
3.2.7 Structure determination of the C-terminal domain of the IBV N and quality of the model. 98
3.2.8 Overall structure of the C-terminal domain of the IBV N. 102
3.2.9 Super helical packing of the C-terminal domain of the IBV N 107
3.3 RNA binding of the IBV N protein..111
3.4 possible model for nucleocapsid formation 122
CHAPTER 4. DISCUSSIONS AND FUTURE WORK 128
4.1 The interaction between the N protein and RNA..................................... 128
4.2 The nucleocapsid formation... 130
4.3 The interaction between the N and M protein 131
4.4 The phosphorylation of the N protein.. 132
Reference 133

Summary

Coronaviruses cause a variety of respiratory and enteric diseases in animals and humans including the severe acute respiratory syndrome. Avian infectious bronchitis virus (IBV) is a major source of mortality in chickens worldwide and has a significant impact on the poultry industry. The viral particle consists of a nucleocapsid or core structure surrounded by a lipid envelope. The nucleocapsid (N) protein is a highly phosphorylated protein interacting intimately with the RNA genome to form nucleocapsid and is also implicated in viral genome replication and in modulating cell signaling pathways. We describe the structure of the two proteolytically resistant domains of the N protein from IBV. These domains are located at its N- and C-terminal ends, respectively. The N-terminal domain of the IBV N protein exhibits typical RNA-binding protein features, with a flexible hairpin loop rich in basic residues and a hydrophobic floor, providing a module for specific interaction with genomic viral RNA. The C-terminal domain forms a tightly intertwined dimer with an intermolecular four-stranded central β-sheet platform flanked by α helices. The C-terminal domain packing observed in one crystal form yields a super helical structure through dimerization and inter-dimer stacking, indicating that assembly of the C-terminal domain is likely to be a trigger for helical nucleocapsid formation.

List of figures

Fig. 1. 1. Morphology of the SARS coronavirus. 15
Fig. 1. 2. Comparison of coronavirus genome structures. 18
Fig. 1. 3. Coronaviruses life cycle. 23
Fig. 1. 4. Fragments of flexible helical nucleocapsids (NC) of Sendai virus 46
Fig. 1. 5. The helical structure of the rigid tobacco mosaic virus rod...................... 47
Fig. 1. 6. An icosahedron with symmetry axes. 48
Fig. 1. 7. 3D reconstruction of icosahedral viruses... 49

Fig. 3. 1. Expression and purification of the IBV N protein.................................. 67
Fig. 3. 2. Mass spectrometry of the IBV N protein.. 68
Fig. 3. 3. Dynamic light scattering of the IBV N protein 70
Fig. 3. 4. Electron microscopy of the IBV N protein.. 71
Fig. 3. 5. Standard protein marker for Superdex 200 column (a) and linear standard curve for calculating M.W. (b). 72
Fig. 3. 6. Size exclusion chromatography of the IBV N protein. 73
Fig. 3. 7. Crosslinking of the IBV N protein 74
Fig. 3. 8. Structural organization of the IBV N protein (a), the N-terminal domain crystal (b) and the C-terminal domain crystal (c). 77
Fig. 3. 9. Phosphorylation site prediction of the IBV N protein 79
Fig. 3. 10. Comparison of recombinant N protein to viral phosphorylated N protein and partially dephosphorylated viral N protein. 80
Fig. 3. 11. SDS-PAGE analysis of the full-length recombinant IBV-N protein and 14.7 kDa proteolytically stable fragment which can be crystallized. 82
Fig. 3. 12. Mass spectrometry of 14.7 kDa proteolytically stable fragment............ 83
Fig. 3. 13. Alignment of native IBV N with mutated strand. 85
Fig. 3. 14. Stereochemistry analysis of native IBV N protein N-terminal domain by Procheck.. 87
Fig. 3. 15. Overall Fold of the IBV-N N-terminal domain. 91
Fig. 3. 16. Structure based alignment of coronavirus nucleocapsid amino-acid sequences corresponding to the proteolytically stable N-terminal fragment.... 92
Fig. 3. 17. Superposition of the IBV-N N-terminal (yellow) domain and the SARS-N N-terminal domain (blue) 93
Fig. 3. 18. Dimer of the N-terminal domain at asymmetric unit. 94
Fig. 3. 19. Standard protein marker for Superdex 200 column (a) and linear standard curve for calculating M.W. (b). 95
Fig. 3. 20. Size exclusion chromatography elution profiles of IBV-N N-terminal domain (IBV N29-160) and the C-terminal domain (IBV-N 218-329). 96
Fig. 3. 21. Mass spectrometry of IBV N C-terminal domain.................................. 97

Fig. 3. 22. Stereochemistry analysis of the C-terminal domain structure by procheck 101
Fig. 3. 23. The C-terminal domain dimer structure represented by secondary structural elements and solvent accessible surface. 103
Fig. 3. 24. Structure-based alignment of coronavirus nucleocapsid amino acid sequences corresponding to the C-terminal dimerization domain. 104
Fig. 3. 25. Crosslinking of the C-terminal domain 105
Fig. 3. 26. The important role of α 4-β1-β2 in the C-terminal domain dimer formation. 106
Fig. 3. 27. Interface I and interface II in the C-terminal domain $P4_3$ crystal packing 108
Fig. 3. 28. The C-terminal domain packing in the tetragonal crystal forms two different interfaces. 109
Fig. 3. 29. Analysis of the RNA-binding activity of the full-length IBV-N protein, the N-terminal domain (IBV-N29-160) and the C-terminal domain (IBV-N218-329) fragments........113
Fig. 3. 30. Electrostatic surface of the N-terminal domain and proposed RNA binding residues. 115
Fig. 3. 31. Electrostatic surface of the C-terminal domain. 117
Fig. 3. 32. Crystal-packing interactions between the C-terminal dimers of IBV (Gray strain) N protein. 119
Fig. 3. 33. Comparison of the C-terminal domain to PRRSV N protein and MS2 coat protein. 122
Fig. 3. 34. Possible model for the IBV helical nucleocapsid formation. 127

List of tables

Table 1. Coronavirus groups, main representative, hosts and principal associated diseases 13
Table 2. Coronvavirus proteins 32
Table 3. RNA-binding motifs... 51
Table 4. The IBV N N-terminal domain data collection and phasing statistics....... 86
Table 5. The IBV N N-terminal domain refinement statistics 88
Table 6. The C-terminal domain crystallization and data collection 99
Table 7. Refinement statistics of the C-terminal domain...................................... 102

Publications:

1. Fan, H., A. Ooi, Y. W. Tan, S. Wang, S. Fang, D. X. Liu, and J. Lescar. 2005. The nucleocapsid protein of coronavirus infectious bronchitis virus: crystal structure of its N-terminal domain and multimerization properties. Structure 13:1859–1868.

2. Tan, Y.W., Fang, S., Fan, H., Lescar, J. and Liu, D.X. Amino acid residues critical for RNA-binding in the N-terminal domain of the nucleocapsid protein are essential determinants for the infectivity of coroavirus in cultured cells. Nucleic Acids Res 34:4816-25

3. Jayaram, H., Fan, H., Bowman, B.R., Ooi, A., Jayaram, J., Collisson, E.W., Lescar, J., Prasad, B.V. (2006) X-ray structures of the N- and C-terminal domains of a coronavirus nucleocapsid protein: implications for nucleocapsid formation. J Virol 80:6612-20

Abbreviations (alphabetically)

CCP4: collaborative computational project number 4

DNA: deoxyribonucleic acid

ds: double stranded

DLS: dynamic light scattering

DTT: dithiothreitol

E .coli: Escherichia coli

EM: electron micrograph

HIV human immunodeficiency virus

IPTG: isopropyl-β-D-1-thiogalactoside

LB: luria broth

MAD: multiple-wavelength anomalous dispersion

Ni-NTA: nickel-nitrilotriacetic acid

NMR: nuclear magnetic resonance

nt: nucleotides

ORF: open reading frame

PCR: polymerase chain reaction

r.m.s deviation: root mean square deviation

RNA: ribonucleic acid

RNase: Ribonuclease

RdRp: RNA dependent RNA polymerase

SDS-PAGE: Sodium dodecyl sulphate-polyacrylamide gel electrophoresis

SeMet: selenomethionine

ss: single stranded

CHAPTER 1 INTRODUCTION

1. 1 Virus

Viruses are small, subcellular agents that are unable to multiply outside a host cell (intracellular, obligate parasitism). The assembled virus (virion) is formed to include only one type of nucleic acid (RNA or DNA) and, in the simplest viruses, a protective protein coat. The nucleic acid contains the genetic information necessary for programming the synthetic machinery of the host cell for viral replication. The protein coat serves two main functions: first, it protects the nucleic acid from extracellular environmental insults such as nucleases; second, it permits attachment of the virion to the membrane of the host cell, the negative charge of which would repel a naked nucleic acid. Once the viral genome has penetrated and thereby infected the host cell, virus replication mainly depends on host cell machinery for energy and synthetic requirements.

Capsid proteins are coded for by the virus genome. Because of its limited size the genome codes for only a few structural proteins used in the construction of new virion and non-structural proteins (nsp) involved in virus replication. Capsids are formed as single or double protein shells and consist of only one or a few structural protein species. Therefore, multiple protein copies must self assemble to form the continuous three-dimensional capsid structure. Self assembly of virus capsids follows two basic patterns: helical symmetry, in which the protein subunits and the nucleic acid are arranged in a helix, and icosahedral symmetry, in which the protein subunits assemble into a symmetric shell that covers the nucleic acid-containing core.

Some virus families have an additional coat, called the envelope, which is usually derived from the host cell membranes. They consist of a lipid bilayer with membrane-associated proteins. In the exterior of the bilayer are inserted virus-coded, glycosylated (trans-) membrane proteins. Therefore, enveloped viruses often exhibit a fringe of glycoprotein spikes or knobs, also called peplomers. In viruses that acquire their envelope by budding through the plasma or another intracellular cell membrane, the lipid composition of the viral envelope closely reflects that of the particular host membrane. The outer capsid and the envelope proteins are responsible for virus binding to the cell membrane receptor(s). The receptor is a normal cell membrane component that participates in virus binding, facilitates viral infection, and is a determinant of virus host range, as well as tissue tropism.

In addition to virus-encoded envelope proteins, budding viruses may also carry certain host cell proteins as integral constituents of the viral envelope. For example, the HIV-1 envelope contains the host cell protein cyclophilin which is required for the efficient initiation of viral reverse transcription (Braaten *et al*., 1996).Virus envelopes can be considered an additional protective coat.

Viruses can be classified in several ways, such as by their geometry, by whether they have envelopes, by the identity of the host organism they can infect, by the mode of transmission, or by the type of disease they cause. The most common classification is probably according to the nucleic acid genome the virus contains and its mode of expression. This classification scheme was originally proposed by David Baltimore (Cann, 1997).

1. 2. Coronavirus

Coronaviruses are a group of enveloped, positive-stranded RNA viruses presently classified as a genus, which, together with the genus Torovirus, constitutes the family *Coronaviridae*. These viruses are grouped with two other families, the *Arteriviridae* and the *Roniviridae*, into the order *Nidovirales*. This classification is not based on structural similarities—in fact, the structure and composition of the viruses from the different families differ significantly—but on common features of genome organization and gene expression (Lai & Cavanagh, 1997).

In general, coronaviruses cause respiratory or intestinal infections in a wide variety of mammals as well as avian species, but some coronaviruses can also infect other organs (liver, kidney, and brain). Most coronaviruses were mainly of veterinary importance. The outbreak of atypical pneumonia in late 2002, which dramatically emphasized the potential relevance of coronaviruses for humans, was termed severe acute respiratory syndrome coronavirus (SARS-CoV). On the basis of antigenic and genetic relationships, the coronaviruses have been subdivided into three groups: group 1, 2 and 3 (Table 1); the taxonomic position of SARS-CoV has not been formally assigned (de Haan & Rottier, 2005).

1.2.1 Classification.

Coronaviruses were named after their corona crown-like appearance in the electron microscope, which is caused by the club-shaped peplomers that radiate outwards from the viral envelope (Fig. 1. 1) (Stadler *et al.*, 2003). At ～30,000 nucleotides (nt), their genome is the largest found in any of the RNA viruses. Coronaviruses are divided into three serotypes: group 1, 2, and 3 (Table 1). Phylogenetic analysis of coronavirus sequences also identified three main classes of viruses, corresponding to each of the three serotypes. Group 2 coronavirus contain a gene encoding hemagglutinin esterase (HE) that is homologous to that of influenza C virus. Groups 1 and 2 contain mammalian viruses, whereas group 3 contains only avian viruses. Within each group, coronaviruses are classified into distinct species by host range, antigenic relationships, and genomic organization. Coronaviruses typically have narrow host ranges and are selective to cell culture. The virus can cause a severe disease in many animals; and several viruses, including infectious bronchitis virus (IBV), feline infectious peritonitis virus, and transmissible gastroenteritis virus (TGEV), are important animal pathogens. Human coronavirus (HCoV) is found in both group 1 and group 2 and responsible for ～30% of mild upper respiratory tract illness, and is responsible for a large proportion of all common cold (de Haan & Rottier, 2005).

Table 1. Coronavirus groups, main representative, hosts and principal associated diseases

Group	Virus	Host	Disease
1	Feline coronavirus (FCoV)	Cat	Respiratory infection/enteritis/peritonitis/systemic enteritis
	Canine coronavirus (CCoV)	Dog	Enteritis
	Transmissible gastroentertitis virus (TGEV)	Pig	Enteritis
	Porcine epidemic diarrhea virus (PEDV)	Pig	Enteritis
	Porcine respiratory coronavirus (PRCoV)	Pig	Respiratory infection
	Human coronavirus (HCoV)-NL63	Human	Respiratory infection
	Human coronavirus (HCoV)-229E	Human	Respiratory infection
2	Murine hepatitis virus (MHV)	Mouse	Respiratory infection/enteritis/hepatitis/encephalitis
	Rat coronavirus (RCoV)	Rat	Respiratory infection
	Bovine coronavirus (BCoV)	Cow	Respiratory infection/enteritis
	Hemagglutinating encephalomyelitis virus (HEV)	Pig	Enteritis
3	Infectious bronchitis virus (IBV)	Chicken	Respiratory infection/enteritis
	Turkey coronavirus (TCoV)	Turkey	Enteritis
?	Severe acute respiratory syndrome-associated coronavirus (SARS-CoV)	Human	Respiratory infection/enteritis

(Table adapted from de Haan & Rottier, 2005).

In March 2003, the causative agent for the outbreak of atypical pneumonia with a high fatality rate was identified as SARS-CoV (Peiris *et al.*, 2003) and its genome was rapidly sequenced and characterized (Rota *et al.*, 2003; Marra *et al.*, 2003). Phylogenetic analyses and sequence comparisons showed that SARS-CoV is not close to any of the previously characterized coronaviruses. It was proposed that SARS-CoV can be defined as a fourth class of coronavirus (Rota *et al.*, 2003; Marra *et al*., 2003). The potential risks for public health posed by SARS-CoV and the current lack of specific antiviral agents or vaccine against this emerging pathogen have triggered a global research effort in order to characterize his family of viruses at the molecular level.

1.2.2 Morphology and structure

The virions of coronaviruses are spherical enveloped particles about 100 to 120 nm in diameter. Inside the virion is a single-stranded, positive-sense genomic RNA 27 to 32 kb in size, the largest of all RNA virus genomes. The RNA genome associates with N proteins to form nucleocapsid. At least two coronaviruses (TGEV and MHV) form a helical nucleocapsid enclosed within an "internal core structure", 65 nm that is spherical, and possibly icosahedral, in form. The core is composed of the M proteins and N proteins. It can be released from virions by NP-40 treatment (Risco *et al*., 1996). The virus core is enclosed by a lipoprotein envelope, which is formed during virus budding from intracellular membranes. Two types of prominent spikes line the outside of the virion. The long spikes (20 nm)，which consist of the S glycoprotein, are present on all coronaviruses (Fig. 1. 1); the short spikes, which consist of the HE (hemagglutinin-esterase) glycoprotein, are present in only some coronaviruses within group 2. The envelope also contains the M glycoprotein, which spans the lipid bilayer three times; thus, the M protein is apparently a component of both the internal core structure and envelope. The envelope also contains the E protein, which is present in much smaller amounts than the other viral envelope proteins (Fig. 1. 1a) (Vennema *et al.*, 1996).

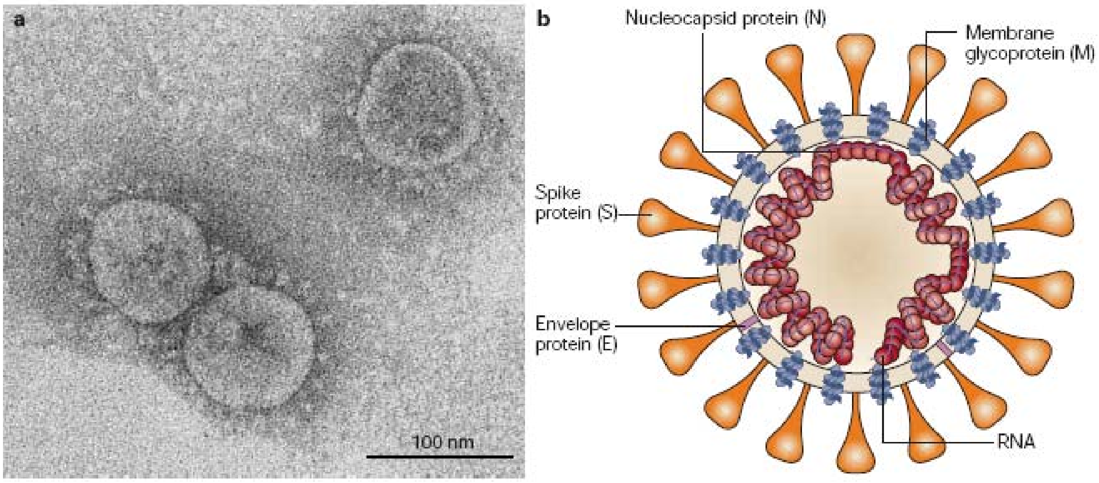


Fig. 1. 1. Morphology of the SARS coronavirus.

a. Electron micrographs of the virus that was cultivated in Vero cells (Image courtesy of Dr L. Kolesnikova, Institute of Virology, Marburg, Germany). Large, club-shaped protrusions consisting of S protein form a crown-like corona that gives the virus its name. b. Schematic representation of the virus. A lipid bilayer comprising the S protein, the M glycoprotein and the E protein cloaks the probably helical nucleocapsid, which consists of the N protein that is associated with the viral RNA. In the case of coronaviruses, the lipid envelope is derived from intracellular membranes (Figure adapted from Stadler *et al.*, 2003).

1.2.3 RNA genome

The coronavirus contains a positive-sense, single stranded RNA genome. The viral RNA functions as an mRNA and is infectious. It contains approximately 7-10 functional genes, 4 or 5 of which encode structural proteins. The genes are arranged in the order 5'-replicase -(HE)-S-E-M-N-3', with a variable number of genes that encode additional non-structural proteins interspersed among them (Fig. 1. 2). This gene arrangement also applied to toroviruses and arteriviruses. The 5' terminus of the coronavirus genome is capped, and the RNA starts with a leader sequence of

65-98 nucleotides, which is also present at the 5' end of mRNA, followed by a 200 to 400 nucleotide untranslated region (UTR). At the other end of the genome is a 3'UTR of 200-500 nucleotides followed by a poly (A) tail of variable length. The sequence of both the 3' and 5' UTR are important for RNA replication and transcription.

Almost two-thirds of the entire RNA is occupied by the polymerase gene, which comprises two overlapping ORFs, 1a and 1b. At the overlap region is a specific seven-nucleotide "slippery" sequence and a pseudoknot structure, characteristic of the ribosomal frameshifting signal (Brierley *et al*., 1987; Brierley *et al.*, 1989). A translational read-through by a -1 ribosomal frameshift mechanism allows the translation of the overlapping reading frames (1a, 1b) into a single polyprotein 1ab.

The coronavirus genome encodes additional non-structural proteins known as 'accessory genes'. Some of these molecules seem to be dispensable for virus viability both in vitro and in vivo; their deletion creates viruses that are attenuated (de Haan *et al*., 2002; Sama *et al*., 2002). The architecture of the nonstructural protein genes interspersed between the known structural protein genes varies significantly among different coronavirus species. In IBV, the two ORFs (5a, 5b) are inserted between the M and N protein genes (Fig. 1. 2). In SARS-CoV, there are six ORFs (6, 7a, 7b, 8a, 8b and 9b) between the M and N genes. The variability of gene structure indicates the plasticity of coronavirus RNA and the frequent occurrence of recombination also suggests that there is no strong conservation pressure on these non-structural proteins (Stadler *et al.*, 2003).

At the 3' end of the genome, a second 340-nucleotide untranslated region (UTR), which is followed by a poly(A) tract, is present. This 3′ UTR contains a 32-nucleotides stem–loop II-like motif (s2m), which has also been reported in astroviruses, in one equine rhinovirus and IBV (Jonassen *et al.*, 1998). Human astrovirus stem-loop II consists of a basal stem of 6 base pairs and a 31 nt loop region. In IBV this stem-loop II-like motif (s2m) can be folded exactly like the Human astrovirus stem-loop II, suggesting its presence in these very different viruses to be the result of a natural RNA recombination between an astrovirus and IBV (Jonassen *et al.*, 1998).

A typical feature of coronaviruses is the presence of a transcription-regulatory sequence (TRS) that is important in RNA transcription and regulation (Marra *et al.*, 2003). This short motif is usually found at the 3′ end of the leader RNA and, with a few exceptions, precedes each translated ORF. The TRSs are thought to regulate the discontinuous transcription of subgenomic mRNAs (Marra *et al.*, 2003). The TRSs include a partically conserved core sequence that in some coronaviruses. In SARS-CoV, Thiel and colleagues (2003) isolated one genomic and eight subgenomic RNAs from the SARS-CoV FRA strain and they identified a conserved sequence (5′ACGAAC3′) that was located in front of nine predicted ORFs, and which fitted the description of a TRS. By contrast, Marra et al and Rota et al (2003) proposed a different TRSs (5'CUAAAC3') in SARS sequence analysis, but these sequences do not precede all predicted genes and no experimental evidence for their function has been provided. Although the overall organization of the SARS-CoV genome is similar to other coronaviruses, the amino-acid conservation of the encoded proteins is usually low (Stadler *et al.*, 2003).

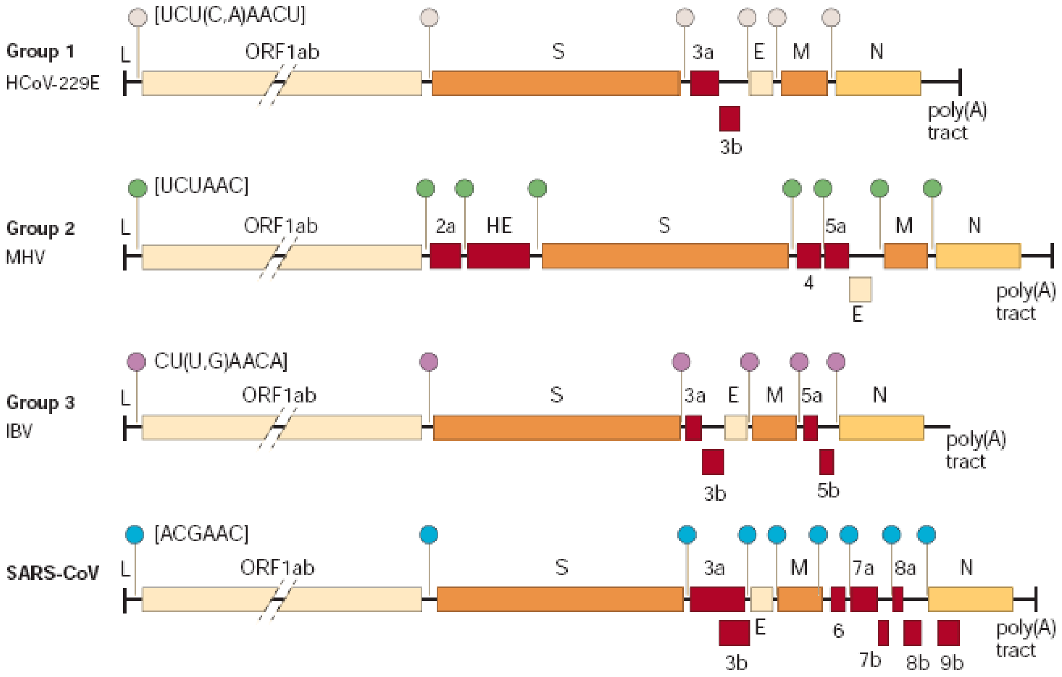


Fig. 1. 2. Comparison of coronavirus genome structures**.**
Genome organization of coronavirus representatives of group 1 (human coronavirus 229E, HCoV-229E), group 2 (mouse hepatitis virus, MHV) and group 3 (avian infectious bronchitis virus, IBV; SARS-CoV). Red boxes represent the accessory genes. The positions of the leader sequence (L) and poly(A) tract are indicated; circles of different colour represent group-specific transcription-regulatory sequences (TRS). (Figure adapted from Stadler *et al.*, 2003)

1.2.4 Virus life cycle

The life cycle of a coronavirus starts when the S protein, which forms the distinctive crown that is observed with coronaviruses, interacts with a receptor through its S1 domain (Bosch *et al.*, 2004). The viral entry, which is mediated by the S2 domain, occurs by membrane fusion. The coronavirus S protein has striking similarity with class I fusion proteins defined by Lescar *et al.,* (2001) like the influenza virus hemagglutinin (Skehel *et al.,* 2000), the HIV-1 gp41 (Weissenhom *et al.,* 1997) and the paramyxovirus F protein (Yin, *et al.,* 2005). The N-terminal half of the S protein (S1) contains the receptor-binding domain, whereas the C-terminal half (S2) is the membrane fusion subunit which is anchored to the membrane like other class I fusion protein (Lescar *et al.*, 2001). The interaction between the S1 protein and its receptor is the major determinant for virus entry and host range restriction. The SARS-CoV S1 protein interacts with the host cell receptor angiotension-converting enzyme 2 (ACE2) (Li *et al.*, 2005a). Receptors have so far been identified for the group 2 coronavirus MHV (CEACAM; Dveksler *et al.*, 1993); the group 1 coronaviruses TGEV and porcine respiratory coronavirus (PRCoV) (pAPN; Delmas *et al.*, 1992), FIPV (fAPN; Tresnan *et al.*, 1996), and HCoV-229E (hAPN; Yeager *et al.*, 1992).

The S2 protein contains an internal fusion peptide and two heptad repeat regions of which one (HR2) is located close to the transmembrane anchor; the other (HR1) is about 170 residues upstream of it. The putative fusion peptide has recently been identified upstream of and near to HR1 (Bartlam *et al.*, 2005). The presence of conserved HRs in S2 suggests, by analogy with other virus-encoded class I fusion glycoproteins, that this region is directly involved in the membrane fusion process The classical mechanism of enveloped virus and host cell membrane fusion mediated by class I fusion proteins was established by Wiley and colleagues in their comprehensive study of influenza hemagglutinin (Skehel *et al.*, 2000). Class I virus fusion proteins have a number of common structural features. They are type I membrane glycoproteins that fold into trimers and contain a protease cleavage site, a fusion peptide and at least two heptad repeat regions, one of which (here designated as HR1) is located downstream and in the vicinity of the fusion peptide, whereas the other (HR2) usually occurs adjacent to the transmembrane domain (Bosch *et al.*, 2004). The fusion proteins acquire a metastable state upon cleavage by cellular proteases. After binding of virus to the receptor or because of protonation during endocytosis, class I fusion proteins proceed through a series of conformational changes to mediate membrane fusion with the host cell. Initially, the fusion peptide located at or close to the N terminus of the membrane-anchored subunit becomes exposed and can insert into the cellular membrane. This is followed by further rearrangements within the protein trimer resulting in the formation of a six-helix bundle. In this structure, a homotrimeric coiled coil formed by HR1 is surrounded by three HR2 helices that pack against the HR1 coiled coil in an antiparallel manner. In the full-length protein, such a conformation leads to a close apposition of the fusion peptide (N-terminally of HR1), inserted in the cellular membrane, and the viral transmembrane segment (C-terminally of HR2), facilitating membrane fusion (Skehel *et al.*, 1998).

Despite these strong similarities to class I virus fusion proteins, coronavirus spike proteins have several characteristics that set them apart. First, unlike class I fusion proteins, cleavage is not essential for coronavirus infection; rather, group 1 coronaviruses are not cleaved at all. Second, although class I fusion proteins carry their fusion peptide at or close to the N terminus of the membrane anchored membrane fusion subunit, no such hydrophobic peptide occurs in this region of (cleaved) coronavirus spike proteins. Although the precise location of the fusion peptide still needs to be determined, it is clear that membrane fusion is mediated by an internal fusion peptide.

Peptides deriving from the HR2 domain can inhibit coronavirus infection, most likely by interfering with six-helix bundle formation, a process essential to drive the membrane fusion reaction and , thus, to initiate infection (Bosch *et al.*, 2004).

After membrane fusion, the viral genome RNA is then released into the cytoplasm where replication takes place. The host translation machinery translates the overlapping open reading frames ORF1a and ORF1b by a ribosomal frame-shifting mechanism to produce a single polyprotein 1ab. Polyproteins 1a and 1ab are cleaved by virally encoded proteinases (3C-like protease and papain like protease) to yield the components that are necessary to assemble the viral replication complex, which synthesizes full-length negative-strand RNA. In addition, a discontinuous transcription strategy during negative-strand synthesis produces a nested set of sub-genomic negative-sense RNAs (Fig. 1. 3). In this process, TRS, which is present at the 3’ end of the leader sequence and, with a few exceptions, upstream of each translated gene, is important. It is postulated to fuse the 3’ ends of the nascent subgenomic minus strands to the antisense leader sequence. These discontinuously synthesized minus strands then act as templates for the synthesis of positive-sense mRNAs. An alternative hypothesis proposes that these mRNA molecules are generated by discontinuous transcription during positive-strand synthesis (the figure shows the positive-sense mRNA products). In almost all cases, only the most 5′ ORF is translated. N proteins and genomic RNA assemble in the cytoplasm to form the nucleocapsid. This core structure acquires its envelope by budding through intracellular membranes between the endoplasmic reticulum (ER) and the Golgi apparatus. The M, E and S proteins, all of which will be accommodated by the lipid bilayer, are transported through the ER to the budding compartment, where the nucleocapsid probably interacts with the M protein to trigger assembly. The interaction between the M and N protein was detected by coimmunoprecipitation in infected cells (Narayanan *et al*, 2000). However, the M and N protein interaction did not occur in cells coexpressing M protein and N protein alone. These data indicates that some MHV function(s) is necessary for the initiation of the M-N protein interaction (Narayanan *et al*, 2000). During the transport of the virus through the Golgi apparatus, sugar moieties are added and in some, but not all, coronaviruses the S protein is cleaved into S1 and S2 domains. Any S protein that is not incorporated into the virions is transported to the cell surface. Finally, the virus is released from the host cell by fusion of virion-containing vesicles with the plasma membrane (Fig. 1. 3).

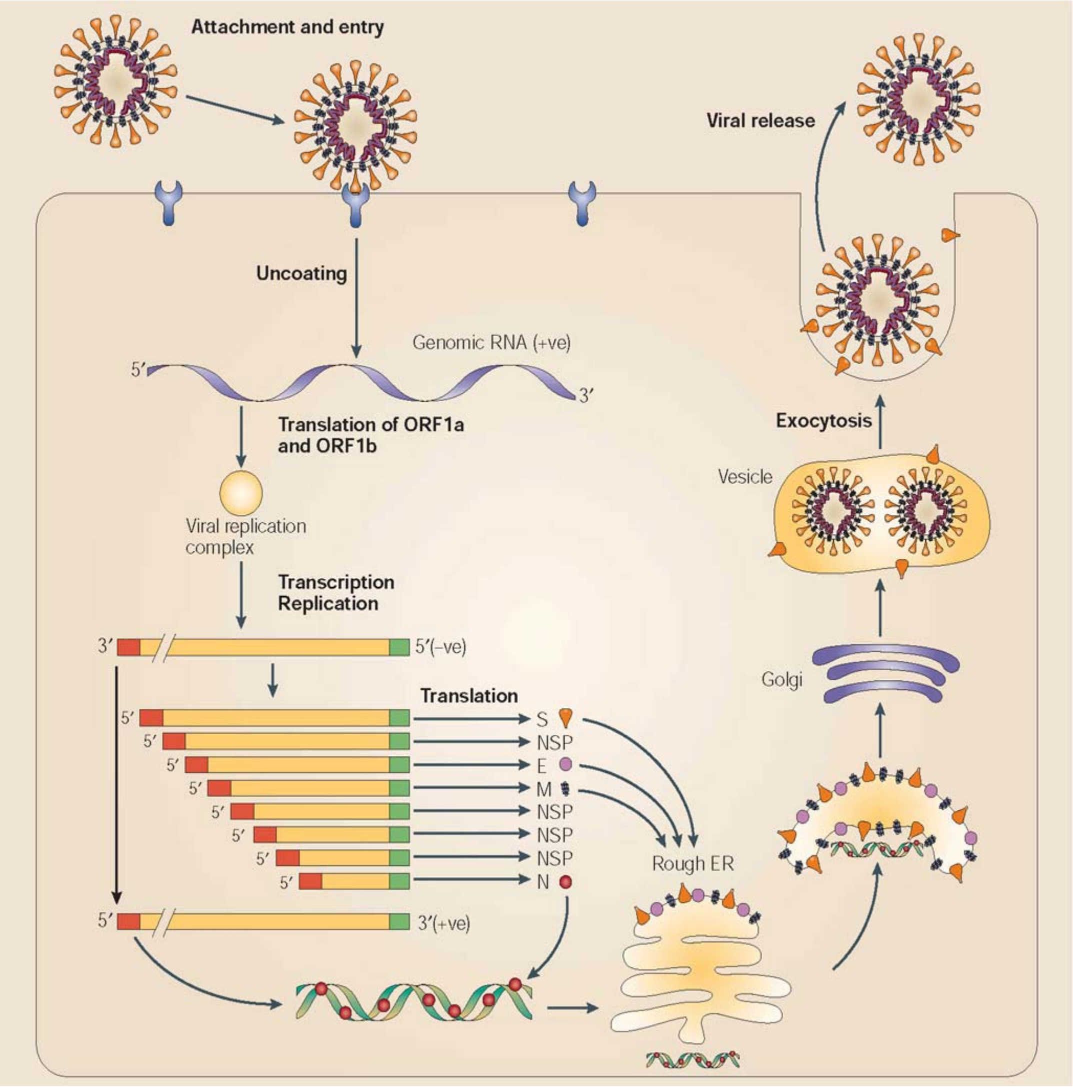


Fig. 1. 3. Coronaviruses life cycle.

The coronavirus enters host cell via S protein receptor-mediated endocytosis. Overlapping open reading frames 1a and 1ab are translated by a ribosomal frame-shifting mechanism to produce polyprotein 1a and 1ab, which are subsequently cleaved by virally encoded proteases to yield the components necessary for the assembly of the virus replication complex. A discontinuous transcription strategy during negative-strand RNA synthesis produces a nested set of sub-genomic negative-sense RNAs. These negative-sense RNAs then act as templates for the synthesis of the positive-sense mRNA. The common leader sequence on the 5'end of each mRNA is shown in red. For the virus assembly and release please see text. (Figure adapted from Stadler *et al.*, 2003).

1.3. Non-structural proteins in coronavirus

1.3.1 Precursor polyprotein 1a and 1ab

Coronavirus genome expression starts with the translation of two large replicative polyproteins, polyprotein 1a and 1ab, which are encoded by the viral replicase gene that comprises ORFs 1a and 1b. The replicase gene accounts for approximately two-thirds of the genome. Two ORFs 1a and 1b, overlap by a few dozen nucleotides. The ORF 1b is in -1 reading frame with respect to the upstream ORF 1a and is translated following ribosomal frameshifting. During -1 ribosomal frameshifting, the ribosome is forced to shift one nucleotide backwards into an overlapping reading frame and to translate an entirely new sequence of amino acids. This process is indispensable in the replication of numerous viral pathogens, including human immunodeficiency virus (HIV) and the coronavirus associated with severe acute respiratory syndrome. -1 ribosomal frameshifting depends on an mRNA signal composed of two essential elements: a heptanucleotide 'slippery' sequence and an adjacent mRNA secondary structure, most often an mRNA pseudoknot (Brierley *et al.*, 1989). Namy *et al.* (2006) proposed a mechanical explanation for -1 frameshifting. When the elongating ribosome encounters the pseudoknot structure, the pseudoknot can stall the elongating ribosome and form a frameshifting intermediate with ribosome. During translocation, the movement of tRNA through the ribosome is resisted by tension developed in the mRNA strand by the pseudoknot blockages. The mRNA in turn is connected to the tRNA by means of the codon–anticodon interaction. Because the tRNA is prevented from returning to the A-site by the presence of eEF2, the ribosome, in attempting to translocate the anticodon into the authentic P-site, places strain on the tRNA that results in the adoption of a bent conformation. The opposing actions of translocation, catalysed by eEF2, and pulling from the mRNA strand account for the bending of the tRNA (Namy *et al.* 2006), spring-like, in a (+) sense (3' direction) and the movement of the tRNA into the P-site. These opposing forces place a strain on the codon–anticodon interaction that promotes breakage. Subsequently, the anticodon-codon interaction breaks over the slippery sequence, allowing a spring-like relaxation of the tRNA in a (-) sense direction (5' direction), allowing the tRNA to repair with the mRNA in the -1 position (Namy *et al.*, 2006). -1 ribosomal frameshift mechanism allows the translation of the overlapping ORFs 1a and 1b into a single polyprotein (Thiel *et al.*, 2003).

The translation strategy is expected to yield two extremely large polyproteins, 1a and 1ab, of about 450 and 750 kDa. However, proteins of these sizes have not been detected in coronavirus-infected cells because translation of 1a and 1ab is coupled with proteolytic processing by viral proteinases, namely the papain-like cysteine protease (PLpro) and the 3C-like cysteine protease (3CLpro). The polyproteins 1a and 1ab are cleaved into individual polypeptides, which provide all the proteins required for replication and transcription (Thiel *et al.*, 2001).

A particular phenomenon in RNA viruses is the use of multiple start codons within a gene which gives rise to different protein products (Meier, *et al.*, 2006). In SARS-CoV, an additional protein is synthesized (called ORF-9b) from an alternative reading frame of the N protein gene (Meier, *et al*., 2006). The ORF-9b plays a role in virus assembly via membrane association (Meier, *et al*., 2006).

1.3.2 Papain-like protease

The coronavirus polyprotein 1ab can be divided into an N-terminal region that is processed by one or two papain like proteases and a C-terminal region that is processed by the 3C-like protease (Ziebuhr, 2001). The ORF 1a encodes two proteases: a PLpro and a 3CLpro. Except for IBV and SARS-CoV which contain only one PLpro, all previously characterized coronaviruses encode two (paralogous) PLpro (PL1pro and PL2pro) which cleave the N-proximal polyprotein region at three sites (Liu *et al*., 1995; Lim et al, 2000; Thiel *et al.*, 2003). SARS-CoV PLpro, similar to IBV, cleaves two sites in N-proximal region. In contrast to other coronaviruses (e.g. HCoV and MHV), which cleave three sites by using two PLpro activity: PL1pro and PL2pro (Thiel *et al.*, 2003).

Baker *et al* (1993) suggest that homologous residues (Cys-1121 and His-1272 for MHV) can be defined as the catalytic dyad of this protease. IBV encodes only one PLpro that is responsible for processing Gly ↓Gly cleavage sites both upstream and downstream of the proteinase domain and autocatalytically releasing protease itself (Kanjanahaluethai & Baker, 2000).

SARS coronavirus, like IBV, contains only one papain-like protease (PLpro) that has activity consistent with both PL1pro and PL2pro. The SARS coronavirus replicase polyproteins pp1a and pp1ab are predicted to be processed into 16 nonstructural proteins (nsp) by the two distinct proteases: PLpro and 3CLpro.The PLpro can act in *trans* to process the N-terminal end of the polyprotein 1a at three sites: Gly180↓Ala181 (nsp 1/2 cleavage site), Gly818↓Ala819 (nsp 2/3 cleavage site), and Gly2740/Lys2741 (nsp3/4 cleavage site) (Harcourt *et al.*, 2004). Residues Cys-1651 and His1812 are required for processing, because substitution of these residues to Ala abolishes activity (Harcourt *et al.*, 2004). Currently, the function of these nonstructural proteins (nsp1, nsp2 and nsp3) is not known. However, biochemical and colocalization studies of a murine coronavirus indicate that these nonstructural proteins are part of a membrane-associated replication complex (Brockway *et al.*, 2003).

The crystal structure of the SARS-CoV PLpro catalytic core (Barretto *et al.*, 2006) shows that it includes an intact catalytic triad, a zinc-binding domain, an N-terminal ubiquitin-like (Ub1) domain, and an overall resemblance to structures of known deubiquitinating binding domains (DUBs) such as USP14 and HAUSP (Barretto *et al.*, 2006). It was demonstrated that SARS-CoV PLpro has deubiquitinating activity. The purified catalytic domain of PLpro can efficiently disassemble diubiquitin and branch polyubiquitin chains. However, the role of these deubiquitinating activities in the virus replication cycle is unclear (Barretto *et al.*, 2006). The active site of PLpro consists of a catalytic triad of cysteine 112, histidine 273, and aspartic acid 287 residues which cleaves the polyprotein 1a at three sites at the N-terminus.

1.3.3 3C-like protease

The 3C-like protease (3CLpro, also called main proteinase) has autocatalytic proteolytic activity. Not only does it cleave the boundaries of itself but also the polyprotein 1a/1ab central and C-proximal region at 11 conserved sites (Ziebuhr *et al.,* 2000), releasing the key replicative functions, such as RNA-dependent RNA polymerase and the helicase, from the polyprotein precursors (Ziebuhr *et al.,* 2001). The Human-CoV and SARS-CoV 3CL (Xu *et al*., 2005) proteinases show that the molecule comprises three domains. Domains I and II are six-stranded antiparallel barrels and together resemble the architecture of chymotrypsin and of picornavirus 3C proteinases. The substrate-binding site is located in a cleft between these two domains. A long loop (residues 184 to 199) connects domain II to the C-terminal domain (domain III). This latter domain, a globular cluster of five helices, has been implicated in the proteolytic activity of 3CL proteinase (Anand *et al.,* 2003). 3CL-proteinase forms a tight dimer in coronavirus, including TGEV and SARS-CoV (Xu *et al*., 2005). In the active site of SARS-CoV 3CL proteinase, a catalytic dyad consists of Cys and His which are fully conserved among all the coronaviruses (Yang *et al.,* 2003). 3CL protease's substrate specificity was defined as P4 (Ser, Thr, Val, Pro, Ala), P2 (Leu, Ile, Val, Phe, Met), P1 (Gln) and P1' (Ser, Ala, Gly, Asn, Cys) residues (Ziebuhr *et al.,* 2000). The notation of Pn-P1-P1'-Pn' for the residues of substrate for inhibitor is that of Schechter and Berger (1967). P1-P1' are the scissile bond residues. Pn and Pn' are the upstream and downstream residues of the scissile bond, respectively.

1.3.4 Helicase

The ORF 1b encodes two functional domains associated with RNA synthesis, including an RNA-dependent RNA polymerase (RdRp), a helicase with ATPase and DNA duplex-unwinding activities (Thiel *et al*., 2003). Computational analysis assigned an mRNA cap-1 methyltransferase to nsp 13 in SARS-CoV (von Grotthuss *et al*., 2003).

The helicase domain is part of non-structural protein 13 (nsp 13), a cleavage product that is released from pp 1ab by the activity of 3CL protease. HCoV-229E nsp13 and SARS-CoV nsp13 have a variety of enzymatic functions, including NTPase, dNTPase, RNA 5'- triphosphatase, RNA helicase, and DNA helicase activities (Ivanov & Ziebuhr, 2004).

1.3.5 RNA-dependent-RNA-polymerase

Following the virus penetration and uncoating, the viral genome is used to translate polyprotein 1a and 1ab using the host translation machinery. Like helicase, the RdRp is encoded by ORF 1b and released by 3CL protease processing from polyprotein 1ab. The RdRp uses the genomic RNA as a template to synthesize negative-stranded RNAs, which are, in turn, used for genomic RNA synthesis (replication) and subgenomic mRNAs synthesis (transcription). In MHV, the best-studied coronavirus in terms of molecular biology, the majority of the viral replicase subunits were found to be associated with intracellular membranes, a feature encountered in many positive-stranded RNA viruses. Coincidentally, experiments on MHV have shown the RdRp co-immunoprecipitates with nsp 8, nsp 9, nsp 5 (main protease, also called 3CL protease) and nsp 13 (helicase), which also implies an interaction between RdRp and the nsp 7 and nsp 8 hexadecamer (Zhai *et al.*, 2005). During the peak of MHV RNA synthesis, key nonstructural proteins such as RdRp and helicase, de novo-synthesized RNA, and also the viral nucleocapsid protein were found to colocalize and associate with the viral RNA replication-transcription machinery (Ivanov *et al.*, 2004). However, the precise intracellular location of coronavirus replication complexes has remained elusive. Results of ultrastructural studies of the replicase complex in MHV-infected suggest that MHV replication occurs on late endosomal and/or lysosomal membranes (van der Meer *et al.*, 1999).

1.3.6 Other non-structural proteins

The coronavirus genome encodes three classes of proteins: structural proteins (the S, M, E, HE and N proteins), non structural proteins involved in viral RNA synthesis (the nsp or replicase proteins), and proteins that are thought to be non-essential for replication in tissue culture but clearly provide a selective advantage *in vivo* (the nsp or accessory proteins) (Marra *et al.*, 2003; Rota *et al.*, 2003). Both replicase and accessory proteins belong to the non structural proteins, but the replicase and accessory proteins are encoded by different genes and translated by different mechanisms. Coronavirus genome has a typical gene order as follows 5’-replicase-(HE)-S-E-M-N-3’. The virus replicase proteins are encoded by replicase genes. They are translated from genomic RNA and are initially synthesized as large polyproteins (1a and 1ab) that are extensively processed by virus-encoded proteases to produce a functional replicase-transcriptase complex. In addition to these replicase and structural protein genes, several open reading frames encoding additional non-structural proteins are called “accessory genes” which are located between the replicase and S genes, between the S and E genes, between the M and N genes (Fig. 1.2). Unlike the replicase proteins which are translated from genomic RNA, the accessory proteins and structural proteins are translated from subgenomic mRNAs (Sutton *et al.*, 2004). In recent years, many studies have focused on a newly identified coronavirus: SARS-CoV. Its genome is predicted to contain 14 functional ORFs, encoding 16 replicase proteins, 4 structural proteins and 8 accessory proteins. The replicase proteins are expected to have multiple enzymatic activities. The activities of a papain-like protease (PLpro, also known as nsp 3), a main protease (also known as 3CL or nsp 5), an RNA-dependent RNA polymerase (RdRp, also known as nsp 12), a superfamily 1-like helicase (also known as nsp 13) and a uridylate-specific endoribonuclease (NendoU, also known as nsp 15) were recently identified and characterized (Zhai *et al.*, 2005). In addition, nsp 3, nsp 14 and nsp 16 were predicted to have ADP-ribose 1’-phosphatase, 3’-5’ exonuclease and 2’-O-ribose methyltransferase domains, respectively (Table 2) (Snijder *et al.*, 2003).

Table 2. Coronvavirus proteins

Protein	ORF	Function	Structure determined	Reference
Structural proteins				
Spike (S) protein	ORF2	Spike	x-ray	Xu *et al*, 2004
Hemagglutinin-Esterase glycoprotein (HE) protein	ORF2b	Viral membrane	No	
Envelop (E) protein	ORF4	Viral membrane	No	
Membrane (M) protein	ORF5	Viral membrane	No	
Nucleocapsid (N) protein	ORF9	Nucleocapsid	NMR, x-ray	Huang *et al*, 2004 Fan *et al*, 2006
Non-structural proteins (nsps)				
Nsp1	ORF1a		No	
Nsp2	ORF1a		No	
Nsp3	ORF1a	PLpro, ADPR	x-ray	Ratia *et al.*, 2006
Nsp4	ORF1a		No	
Nsp5	ORF1a	3C-like protease	No	Xu *et al.*, 2005
Nsp6	ORF1a		No	
Nsp7	ORF1a	RNA-binding	x-ray	Zhai *et al.*, 2005
Nsp8	ORF1a	RNA-binding	x-ray	Zhai *et al.*, 2005
Nsp9	ORF1a	ssRNA-binding	x-ray	Egloff *et al.*, 2004
Nsp10	ORF1a		No	
Nsp11	ORF1a		No	
Nsp12	ORF1b	RdRp	No	
Nsp13	ORF1b	RNA Helicase, ATPase, NTPase	No	
Nsp14	ORF1b	Endonuclease	No	
Nsp15	ORF1b	Endonuclease	x-ray	Ricagno *et al*, 2006
Nsp16	ORF1b	2'-O-MT	No	
Accessory proteins				
Orf3a	ORF3a		No	
Orf3b	ORF3b		No	
Orf5a	ORF5a		No	
Orf5b	ORF5b		No	
Orf6	ORF6		No	
Orf7a	ORF7a	Ig like	x-ray	Nelson *et al,* 2005
Orf7b	ORF7b		No	
Orf8a	ORF8a		No	
Orf8b	ORF8b		No	
Orf9b	ORF9b	Lipid binding	x-ray	Meier *et al,* 2006

ADPR, ADP-ribose 1'-phosphatase; 2'-O-MT, 2'-O-ribose methyltransferase; PLpro, papain-like protease; RdRp, RNA-dependent RNA polymerase;

The SARS-CoV nsp 7 and nsp 8 can form a hexadecameric supercomplex, the structure of which was solved by crystallography (Zhai *et al*., 2005). The structure of the supercomplex resembles a hollow cylinder with a central channel and two handles protruding from opposite sides. The electrostatic properties and dimensions of the nsp 7-nsp 8 supercomplex imply that its role is to bind nucleic acids. The inner channel is coated by positive potential, whereas the outer surface of the cylinder is mainly covered by negatively potential. This bipartite charge distribution ensures that the phosphate backbone of the nucleic acids can pass through the channel without electrostatic repulsions, as with other DNA/RNA-binding proteins such as the eukaryotic DNA polymerase processivity factor PCNA (Krishna *et al*., 1993) and the β subunit of *E. coli* DNA polymerase III holoenzyme (Kong *et al*., 1992). Imbert *et al* (2006) discovered that nsp 8 is a second, non-canonical RNA-dependent RNA polymerase (RdRp) in SARS-CoV. This enzyme may catalyze the synthesis of RNA primers for the primer-dependent nsp 12 RdRp (Imbert *et al*. 2006).

The structure of SARS-CoV nsp 9 has a central core comprised of a six-stranded barrel, flanked by a C-terminal helix and N-terminal extension. The topology of the protein most closely resembles the domains of the chymotrypsin-like proteases (members of the serine protease superfamily), which have two domains comprising a six-stranded barrel motif for each domain (coronavirus has a third α-helical domain) (Yang *et al*., 2003). In cells infected by the related coronavirus MHV, nsp 9 is localized in the perinuclear region, together with three other proteins of the replication complex. Also, the polymerase (nsp 12) has been shown to coimmunoprecipitate with 3CL protease (nsp 5), nsp 8 and nsp 9 (Brockway *et al.*, 2003). Analytical ultracentrifugation experiment further indicates that nsp 8 interacts with nsp 9. On the basis of the nsp 7-nsp 8 supercomplex structure, the most probable nsp 9 binding site should be in the region located at the entrance of the channel and has high flexibility (Zhai *et al.*, 2005). In addition, the nsp 9 is capable to bind RNA and this binding is not strongly RNA sequence specific.

No functions have been definitively assigned to other replicase proteins.

1.4. Accessory proteins in coronavirus

The coronavirus genome has a principal organization with 'conserved' open reading frames (ORFs) in 5'-replicase-S-E-M-N-3' arrangement. In addition to these conserved genes, several open reading frames encoding additional non-structural proteins are known as 'accessory genes'. Some of these proteins seem to be dispensable for the virus viability both *in vitro* and *in vivo*. Their deletion creates viruses that are attenuated (Stadler *et al.*, 2003). These accessory genes differ significantly among coronavirus groups. They are also referred to as group-specific genes. The SARS-CoV genome contains eight novel ORFs (3a, 3b, 6, 7a, 7b, 8a, 8b and 9b) (Table 2). To date, the functions of these genes remain largely unknown (Table 2), although their absence from other genomes suggests unique functions that might be associated with SARS-CoV replication, assembly or virulence.

IBV (in group 3) contains three ORFs (3a, 3b and 3c) between the S and E protein genes (Fig. 1. 2). ORF 3c encodes the E protein, which is a viral structural protein, while the ORFs 3a and 3b encode two accessory proteins (Liu *et al.*, 1991). Moreover, IBV is unique in that it has two ORFs (5a and 5b) between the M protein gene and the N protein genes, which encode proteins of 7.4 and 9.5 kDa, respectively (Liu & Inglis, 1992). All these accessory proteins have been detected in very small amounts in virus-infected cells.

In the group 1 coronavirus genome, for example HCV-229E, there are two ORFs (3a and 3b) between the S and E protein genes (Fig. 1. 2). Members of the group 1 coronavirus exhibit great heterogeneity with respect to the number, size and mechanism of expression of ORFs between the S and E genes. These accessory proteins probably are not required for viral replication (Duarte *et al.*, 1994).

In MHV (coronavirus group 2) genome, there are two genes located between the polymerase and S genes (Fig. 1. 2). Gene 2-1 encodes the HE protein, while gene 2a encodes an ns protein of unknown function. Additionally, there is gene 5 between the S and E genes and gene 5 of the MHV has two ORFs, 5a and 5b (Fig. 1. 2.). The latter encodes the structural E protein, whereas the 5a gene encode an ns protein without known functions (Lai & Cavanagh, 1997).

1.5. Structural proteins in coronavirus

In contrast to most of the nonstructural proteins, which are expressed from the genome RNA, the coronavirus structural proteins and a number virus-specific accessory proteins are expressed from an extensive nested set of 3'-coterminal sub-length mRNAs that possess a common 5' leader sequence derived from the 5' end of the genome (Spaan *et al.*, 1983).

1.5.1 Spike (S) protein

The Spike (S) protein is a large type I glycoprotein of 1160- to 1452-amino acid-long (Bosch *et al.*, 2004; Lescar *et al.*, 2001) with a cleavable N-terminal signal sequence and a membrane-anchoring sequence followed by a short hydrophilic carboxyl-terminal tail of about 30 residues. When comparing primary sequences, the S protein shows two faces: an amino-terminal half with hardly any sequence similarities and the carboxyl-terminal half in which regions with significant conservation can be observed (de Haan & Rottier, 2005).

The S protein can be cleaved into an amino-terminal S1 subunit and a membrane anchored S2 subunit. A basic amino acid sequence resembling the furin consensus sequence motif (RXR/KR) occurs approximately in the middle of the protein and was shown to be the target of a furin-like enzyme in the case of MHV-A59 (de Haan *et al.*, 2004). Cleavage has been demonstrated for S proteins from coronavirus groups 2 and 3, but not for S proteins from group 1 viruses or from SARS-CoV (de Haan *et al.*, 2004). The coronavirus S protein has two functions, which appear to be spatially separated. The S1 subunit (or the equivalent part in viruses with the uncleaved S protein) is responsible for receptor binding, and the S2 subunit is responsible for membrane fusion. The interaction between the S protein and its receptor is the major determinant for virus entry and host range restriction. The SARS-CoV S1 protein recognizes host cell angiotension-converting enzyme 2 (ACE2) as its cell receptor (Li *et al.*, 2005a).

The S protein on mature SARS-CoV virions does not appear to be cleaved, and the sequence that aligns with the MHV cleavage site lacks the essential residues for furin susceptibility. The uncleaved S1 and S2 subunits of the SARS-CoV S protein contain 666 and 583 amino acid residues, respectively.

The structures of refolded heptad repeat fragments of S2 from the mouse hepatitis coronavirus (MHV) (Xu *et al.*, 2004) and from SARS-CoV (Supekar *et al.*, 2004) show that postfusion conformation has the trimer-of-hairpins organization characteristic of “class 1” fusion proteins (Lescar *et al.*, 2001), such as those of HIV (Weissenhom *et al.*, 1997), influenza virus (Skehel *et al.*, 2000), and Ebola virus (Skehel *et al.*, 1998). The S2 contains an internal fusion peptide and has two hydrophobic (heptad) repeat regions, designated HR1 and HR2. Both MHV and SARS-CoV fusion core structures exhibit a six-helix bundle in which three HR1 helices form a central coiled coil surrounded by three HR2 helices in an oblique antiparallel manner. HR2 peptides pack into the hydrophobic grooves of the HR1 trimer in a mixed extended and helical conformation; this represents a stable post-fusion structure, similar to that observed for HIV-1 gp41 (Bartlam *et al.*, 2005).

1.5.2 Hemagglutinin-Esterase glycoprotein (HE) protein

Viruses in group 2 express and incorporate into their particles an additional membrane protein, HE protein. Only coronaviruses belonging to the group 2 possess the HE gene. Although all group 2 viruses contain an HE gene, the protein is not expressed by all MHV strains (Yokomori *et al.*, 1991), indicating that HE is a non-essential in these viruses. The HE gene encodes a type I membrane protein of 424–439 residues that contains a cleavable signal peptide at its amino terminus (Hogue *et al.*, 1989) and a transmembrane domain close to its carboxyl terminus, leaving a short cytoplasmic tail of about 10 residues. The ectodomain contains 8–10 putative N-linked glycosylation sites.

Little is still known about the function(s) of the coronavirus HE protein. The protein contains hemagglutinin and acetyl esterase activities. While the HE proteins of BCoV, HEV, and HCoV-OC43 hydrolyze the 9-O-acetyl group of sialic acid and therefore appear to function as receptor-destroying enzymes (Schultze *et al.*, 1991; Vlasak *et al.*, 1988), the HE proteins of MHV like coronaviruses function as sialate-4-O-acetylesterases (Klausegger *et al.*, 1999; Regl *et al.*, 1999; Wurzer *et al.*, 2002).

1.5.3 Envelope (E) protein

Coronavirus envelope (E) protein is a small integral membrane protein with multi-functions in virion assembly. The E protein from different coronaviruses share very low homology in the primary amino acid sequence. The E protein contains a relatively large hydrophobic region in its amino-terminal half, followed by a cysteine-rich region, an absolutely conserved proline residue, and a hydrophilic tail (de Haan and Rottier, 2005). A striking feature of E protein is that different coronavirus E proteins assume distinct membrane topologies. Studies on SARS-CoV E protein membrane topology show that both the N- and C-termini of the E protein are exposed to the cytoplasmic side of the membranes (Yuan *et al*., 2006). In contrast, parallel experiments showed that the E protein from the infectious bronchitis virus (IBV) spanned the membranes once, with the N-terminus exposed luminally and the C-terminus exposed cytoplasmically (Corse & Machamer, 2000).

Most coronavirus E protein could be translocated to the cell surface to facilitate budding and release of progeny viruses (Liu & Inglis, 1991). In addition, the E protein of SARS-CoV and MHV was found to modify membrane permeability, allowing entry of small molecules into cells and leading to cell lysis (Liao *et al*., 2004; Liao *et al*., 2006). A minor proportion of the SARS-CoV E protein was found to be post-translationally modified by N-linked glycosylation on the asparagines 66 residue (Yuan *et al*., 2006). The function and effects of this modification is not currently known.

1.5.4 Membrane (M) protein

The M protein, ranging from 225-260 residues in length, is the most abundant envelope protein in coronavirus. The topology of M protein is characterized as having three domains: a short N-terminal ectodomain, a triple transmembrane domain, and a C-terminal endodomain (Naryanan *et al*., 2000). The TGEV M protein adopts two topologies in the virus envelope. The two-thirds of the molecules that are in a Nexo-Cendo conformation (with their carboxyl termini embedded within the virus core) interact with the internal core, and remaining third of the molecules, whose carboxyl termini are in a Nexo-Cexo conformation with both the amino and carboxyl termini exposed to the virion surface (Escors *et al*., 2001). The M proteins with the Nexo-Cexo conformation were removed when the virus envelope was disrupted during the core purification. In contrast, the M protein with the Nexo-Cendo conformation interact with the viral nucleocapsid through M protein C-terminal endodomain in the core and stabilize the viral core (Escors *et al*., 2001). In addition, the cryoelectron microscopy studies on TGEV core showed that the C-terminus of the M protein is the main component in the viral spherical, probably icosahedral internal core. After removing the virus envelope by NP-40 treatment and core purification, the M proteins remain tightly associated with the viral cores in large amounts, covering most of their surface. The existence of the M protein in the internal core can be detected by anti-M monoclonal antibodies (Escors *et al.*, 2001). Since the M protein is an integral membrane protein of the virion envelope, its presence on the surface of the internal core might be the result of stabilizing the interactions from the intravirion domains of the M protein (Escors *et al.*, 2001).

In MHV, the M protein interacts with the N protein in a pre-Golgi compartment, which is part of the MHV budding site. Coimmunoprecipitation analyses further reveal that the M protein interacted with only genomic-length MHV mRNA, mRNA1. These data indicate that the M protein interacts with the nucleocapsid, consisting of the N protein and mRNA1. The M protein and nucleocapsid interaction occurred in the absence of the S and E proteins. Intracellular M- N protein interaction was maintained after removal of viral RNAs by RNase treatment. This data indicated M-N protein interaction is independent of viral RNA (Narayanan *et al.*, 2000). A 16-amino-acid domain (residue 237 to 252) located in the C-terminal domain of the TGEV M protein, has been identified as being responsible for binding the M protein to the nucleocapsid (Escors *et al.*, 2001). Moreover, the C-terminal 45 residues of the N protein in MHV interact with the M protein C-terminal domain (Hurst *et al.*, 2005).

The M protein plays a predominant role in the assembly of virus particle, for which the S protein appears not to be required. Growth of coronaviruses in the presence of tunicamycin gave rise to the production of spikeless, noninfectious virions (de Hann *et al*., 1998). These particles were devoid of the S protein but contained the M protein. Interaction between the M and S proteins has been identified by coimmunoprecipitation and sedimentation analysis. Only the M and E proteins are required for virus like particle formation. The S protein is dispensable but was incorporated when present (Vennema *et al*., 1996).

1.5.5 Nucleocapsid protein (N)

The N protein is a 45-60 kDa phosphoprotein which, together with the genomic RNA, forms a nucleocapsid (RNP). The N proteins vary from 377-455 amino acids in length, are highly basic, and have a high (7-11%) serine content which are potential target for phosphorylation. The amino acid sequences of the N proteins are quite similar within the groups, whereas the homology between proteins from different coronavirus groups is rather limited (30-35%).

Consistent with the N proteins' role as nucleic acid-binding proteins, they are all highly basic because of the abundance of arginine and lysine residues. The abundance of basic residues is also reflected in the calculated overall isoelectric points of the N proteins, the values of which range between 9.7-10.1. The N protein's specific packaging to viral genome is usually performed via the recognition of a particular nucleotide sequence. Such "packaging signals" have been identified at the 3' end of the viral genomes of mouse hepatitis virus (MHV) (Fosmire *et al.*, 1992) and bovine coronavirus (BCV) (Cologna & Hogue, 2000). In elegant structural studies performed on other viral families with RNA genomes, such as human immunodeficiency virus (HIV) (De Guzman *et al.*, 1998) and the MS2 bacteriophage (Valegard *et al.*, 1997), the packaging signals were seen to form a stem-loop structure that is recognized by the nucleocapsid protein.

Although the primary function of the N protein is to form tubular or helical nucleocapsid, as shown for the TGEV (Escors *et al.*, 2001), several studies indicate the protein to be multifunctional. Immunofluorescence microscopy has shown the N protein to be localized in a particulate manner throughout the cytoplasm of coronavirus-infected cells. Although the protein lacks a membrane-spanning domain it was found in association with membranes (Anderson & Wong, 1993; Sims *et al.*, 2000; Stohlman *et al.*, 1983). N protein is a likely component of the coronavirus replication and transcription complex. Its presence is not an absolute requirement for replication and transcription because a human coronavirus (HCoV) RNA vector containing the complete pol1ab gene appeared to be functional in the absence of the N protein (Thiel *et al.*, 2003). However, the efficiency of the system was much enhanced when the protein was present. In addition to its cytoplasmic localization, the N proteins of IBV, MHV, and TGEV have also been demonstrated to localize to the nucleolus both in coronavirus-infected cells and when expressed independently (Hiscox *et al.*, 2001; Wurm *et al.*, 2001). Interactions that have been observed between the N protein and leader/TRS sequences (Baric *et al.*, 1988; Nelson *et al.*, 2000; Stohlman *et al.*, 1988) and between N protein and the 30 UTR (Zhou *et al.*, 1996) suggest a role for the N protein in the discontinuous transcription process. Furthermore, the N protein was also shown to interact with cellular proteins that play a role in coronavirus RNA replication and transcription (Choi *et al.*, 2002; Shi *et al.*, 2000).

Interestingly, TGEV N protein has RNA chaperone activity (Zuniga *et al.*, 2007), which can promote annealing of DNA and viral transcription-regulation sequence (TRS) RNAs in vitro. However, the precise role of the N protein in vivo as an RNA chaperone has not been determined.

1.6 Viral nucleocapsid architecture

The nucleic acid of a virion is enclosed within a protein coat, or capsid, composed of multiple copies of one protein or a few different proteins. A capsid and the enclosed nucleic acid is called a nucleocapsid. The simplest virions consist of two basic components: nucleic acid (single- or double-stranded RNA or DNA) and a protein coat, the capsid, which functions as a shell to protect the viral genome from nucleases and which during infection attaches the virion to specific receptors exposed on the prospective host cell. Capsid proteins are coded for by the virus genome. Because of its limited size the genome codes for only a few structural proteins (besides non-structural regulatory proteins involved in virus replication). Capsids consist of only one or a few structural protein species instead of a large number of different proteins. Therefore, multiple protein copies must self assemble to form the continuous three-dimensional (3D) capsid structure. Self assembly of virus capsids can follow two basic patterns: helical symmetry, in which the protein subunits and the nucleic acid are arranged in a helix, and icosahedral symmetry, in which the protein subunits assemble into a symmetric shell that covers the nucleic acid-containing core.

Electron microcopy (EM) is a primary tool for classifying viruses. In addition, the development of efficient algorithms for processing EM micrographs to produce 3D structures of the viral capsids has revealed more details about these viral architectures.

In the replication of viruses with helical symmetry (for example Sendai virus), identical protein subunits (protomers) self-assemble into a helical array surrounding the nucleic acid, which follows a similar spiral path. Such nucleocapsids form rigid, highly elongated rods or flexible filaments; in either case, details of the capsid structure are often discernible by EM (Fig. 1. 4).

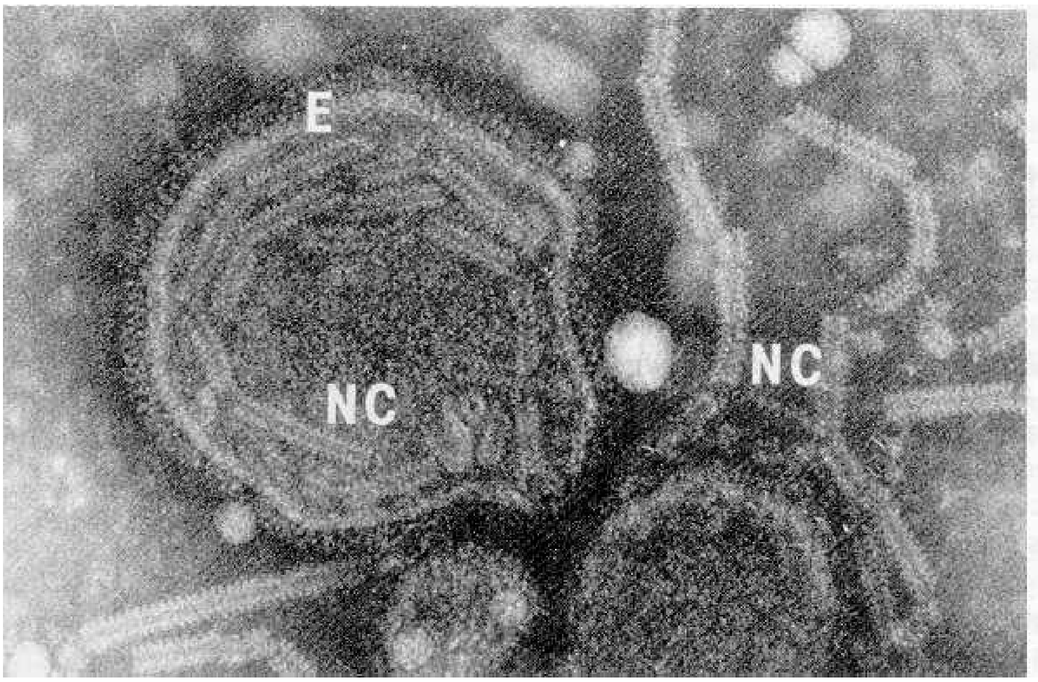


Fig. 1. 4. Fragments of flexible helical nucleocapsids (NC) of Sendai virus, a paramyxovirus, are seen either within the protective envelope (E) or free, after rupture of the envelope. The intact nucleocapsid is about 1,000 nm long and 17 nm in diameter; its pitch (helical period) is about 5 nm. (Figure adapted from Egelman *et al*., 1989)

The helical nucleocapsids are characterized by length, width, pitch of the helix, and number of protomers per helical turn. The most extensively studied helical virus is tobacco mosaic virus (Fig.1.5). Many important structural features of this plant virus have been revealed by x-ray diffraction studies (Bhella *et al*, 2002).

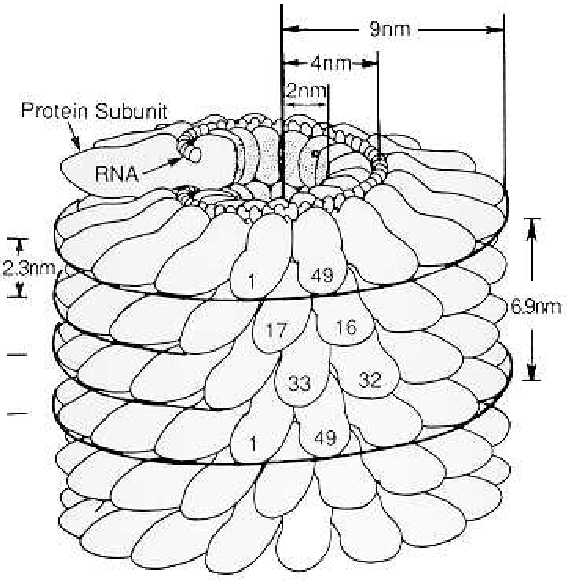


Fig. 1. 5. The helical structure of the rigid tobacco mosaic virus rod
Individual 17.4 kDa protein subunits (protomers) assemble in a helix with an axial repeat of 6.9 nm (49 subunits per three turns). Each turn contains a non-integral number of subunits (16-1/3), producing a pitch of 2.3 nm. The RNA ($2\times1O^6$ Da) is sandwiched internally between adjacent turns of capsid protein, forming a RNA helix of the same pitch, 8 nm in diameter, that extends the length of virus, with three nucleotide bases in contact with each subunit. Some 2,130 protomers per virion cover and protect the RNA. The complete virus is 300 nm long and 18 nm in diameter with a hollow cylindrical core 4 nm in diameter.(Figure adapted from Bhella *et al*, 2002)

The other major structural class of viruses, called icosahedral or quasi-spherical viruses, is based on the icosahedron, a solid object built of 20 identical faces, each of which is an equilateral triangle. In the simplest icosahedral virion, each of the 20 triangular faces is constructed of three identical capsid protein subunits, making a total of 60 identical subunits per capsid (T=1). Any icosahedron has a defined set of exact symmetry elements: 6 fivefold axes through the 12 vertices, 10 threefold axes through the 20 triangular face, and 15 twofold axes through the edges (Fig. 1. 6)

(Baker *et al*., 1999). An icosahedron with fivefold, threefold, and twofold axes of rotational symmetry is defined as having 5-3-2 symmetry.

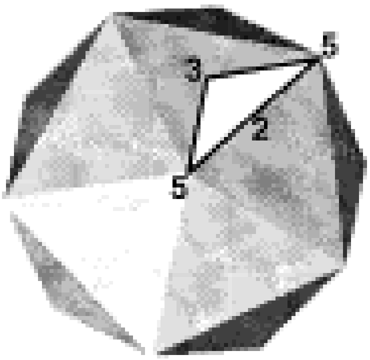


Fig. 1. 6. An icosahedron with its symmetry axes.

The numbers (5, 3 and 2) indicate the positions of some of its symmetry axes.

Viruses were first found to have 5-3-2 symmetry by x-ray diffraction studies and subsequently by EM with negative-staining techniques. Most virus capsids consist of multiples of 60 copies of subunits that obey icosahedral symmetry (T>1). The quasi-equivalence theory of Caspar & Klug (1962) explained the arrangement of these subunits by proposing that they are held together by interactions that are quasi similar to the ones found between the subunits related by the strict icosahedral symmetry. They introduced the T (triangulation) number ($T=h^2+hk+k^2$, where h and k are integers), which corresponds to the number of unique quasi-equivalent environments present in a given surface lattice. Quasi-equivalence can be defined as the extent of similarity between these structurally unique environments occupied by the chemically identical subunits in the virus capsid (Caspar & Klug, 1962).

Negative staining and Cryo-EM in combination with image reconstruction provides a direct, objective way to determine triangulation numbers of spherical viruses and also allows direct determination of the number of subunits in the virion when individual units are resolved. For example, satellite tobacco mosaic virus, cowpea chlorotic mottle virus, Norwalk virus and Nudaurelia capensis virus are icosahedral viruses as shown in Fig. 1. 7 (Natarajan *et al*., 2005). 3D reconstruction identified not only the T numbers of these viruses but also the numbers of subunits and their environments.

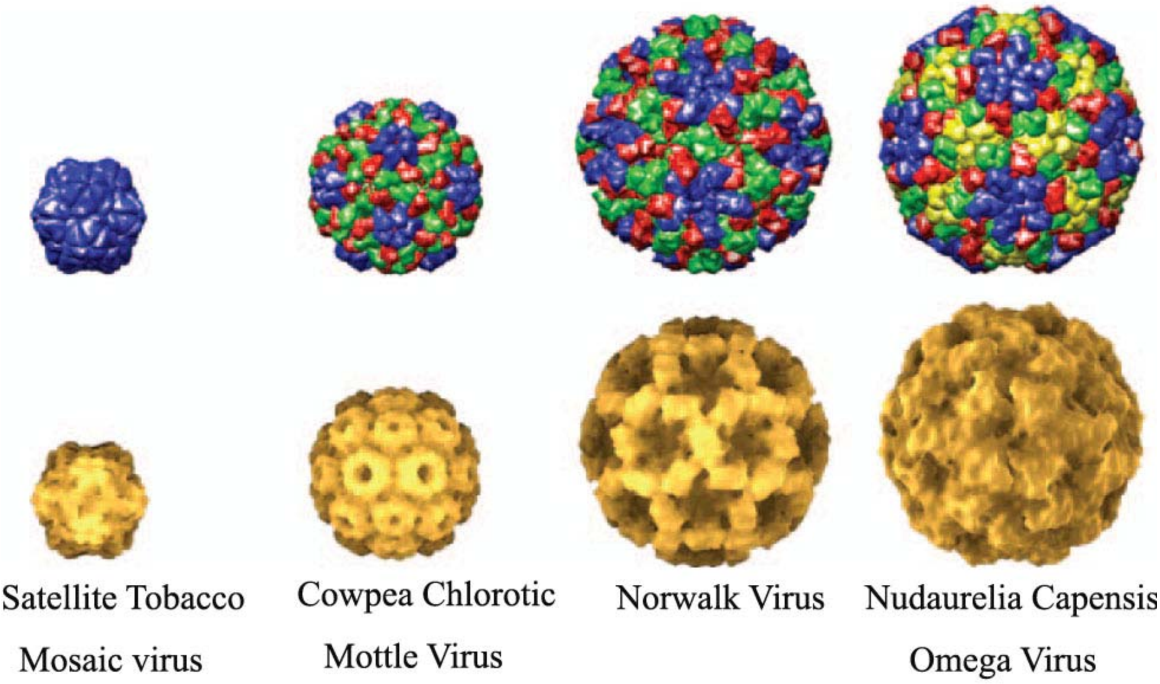


Fig. 1. 7. 3D reconstruction of icosahedral viruses.
The viruses shown (left to right) are viruses with triangulation number T=1 (satellite tobacco mosaic virus), T=3 (cowpea chlorotic mottle virus), T=3 (Norwalk virus) and T=4 (Nudaurelia capensis ω virus). From top to bottom are shown a 5-Å-resolution rendering of the subunit coordinates shown by different colors, a radial rendering view of the particle. (Figure adapted from Nataranjia *et al*., 2005)

Except in helical nucleocapsids, little is known about the packaging or organization of the viral genome within the core. Small virions are simple nucleocapsids containing 1 to 2 protein species. The larger viruses contain in a core the nucleic acid genome complexed with basic protein(s) and protected by a single- or double layered capsid (consisting of more than one species of protein) or by an envelope.

1.7 RNA-binding proteins

A multitude of RNA-binding proteins play key roles in post-translational regulation of gene expression in eukaryotic cells. The study of proteins that bind to RNA has led to the identification of several protein families. Each family is characterized by a common sequence motif that is either known or supposed to interact directly with RNA. RNA is normally single stranded but often forms secondary structures such as hairpins (stem loops), bugles and internal loops through pairing of complementary stretches within the strand (Chastain & Tinoco, 1991). These secondary structural elements are often used as binding sites for proteins. Loop and bulge structures allow many opportunities for specific recognition, since they expose the RNA backbone and bases to interaction with protein groups. RNA also folds into more complex three-dimensional structures through tertiary interactions, as observed in tRNA and group I introns (Nagai, 1996).

The most common of these RNA-binding motifs are the ribonucleoprotein (RNP) motif, the arginine-rich motif (ARM), the arginine-glycine-glycine (RGG) box, the K-homologous (KH) motif, the double-stranded RNA-binding motif (DSRM) and the zinc knuckle motif (Table 2). We describe each of the major RNA-binding motifs and provide examples of how they may function below.

Table 3. RNA-binding motifs

Motif name	Length (amino acid)	Examples	Reference
RNP motif	～80	U1 snRNA (U1 A)	**Nagai *et al,* 1990**
ARM	10−20	HIV Rev and Tat	**Malim *et al.*, 1989; Karn & Graeble, 1992**
RGG box	20-25	hnRNP U protein	**Kiledjian & Dreyfuss, 1992**
KH motif	～50	human hnRNP K protein	**Siomi *et al.*, 1993**
dsRNA-binding	～70	interferon-induced protein kinase (DAI)	**Mathews & Shenk, 1991**
Zinc knuckle	～14	TFIIIA	**Theunissen *et al.*, 1992**

1.7.1 The RNP motif

The RNP motif is most widely found and characterized RNA-binding motif. It is composed of about 80 amino acids which form an RNA-binding domain (RBD) that is present in over 200 ribonucleoproteins (RNPs) involved in RNA processing, transport and metabolism(Mattaj, 1993). The identifying feature of the RNP motif is the RNP consensus sequences, which is composed of two short sequences, RNP1 and RNP2, and a number of other, mostly hydrophobic conserved amino acids interspersed throughout the motif (Dreyfuss *et al*., 1993). The structural details of the RNP motif are available from the three-dimensional structures of the NH2-terminal RBD of U1 snRNA (U1 A) (Nagai *et al,* 1990) and the single RBD of hnRNP C (Kenan *et al*., 1991). The RBDs show a similar four-stranded antiparallel β sheet flanked on one side by two α helices. The RNP1 and RNP2 motifs are located in the two middle β sheets and play a crucial role in RNA binding (Oubridge *et al*., 1994). Charged and aromatic side chains of RNP1 and RNP2 are solvent exposed and make direct contact with bound RNA through hydrogen bonds (Nagai *et al,* 1990; Hoffman *et al,* 1991). RNA binding studies indicate that β sheets of RBD, which contain many of the most highly conserved residues of the domain, constitute a general RNA-binding surface but probably do not distinguish between different RNA sequences. Major determinants of RNA-binding specificity reside in the most variable regions of the RNP motif (Burd & Dreyfuss, 1994).

One RNA-binding protein may contain multiple RBDs which can bind different RNA sequences simultaneously. U1A, for example, binds to U1 snRNA through its first RBD and to pre-mRNA sequences through its second RBD (Lutz & Alwine, 1994). The structure of the RNP motif RBD when bound to RNA is nearly identical to the unbound structure. The exposed β sheet RNA-binding surface engages RNA as an open platform rather than buries the RNA in a binding crevice (Gorlach *et al*., 1992)

1.7.2 The arginine-rich motif (ARM)

Short (10 to 20 amino acids) arginine-rich sequences in viral, bacteriophage and ribosomal proteins mediate RNA binding. Structure-function information is available for human immunodeficiency virus (HIV) Rev, a regulatory RNA-binding protein that facilitates the export of unspliced HIV pre-mRNA from nucleus (Malim *et al*., 1989). Rev binds with high affinity to an internal bulged loop (Rev responsive element, RRE) found in all intron-containing viral mRNAs (Malim *et al*., 1989). Another HIV-encoded ARM protein, Tat, binds trans-acting responsive element of HIV mRNAs and functions in transcription (Karn & Graeble, 1992).

The arginines in ARMs are essential for specificity (Tan *et al*., 1993). The RNA binding sites of ARM proteins consist of stem-loops (nucleocapsid protein), internal loops (Rev), or bulges (Tat), and their structure rather than particular sequence, may be the major binding determinant (Malim *et al*., 1989). Rev and Tat bind to the RNA base and phosphoribose backbone probably through hydrogen bonds (Puglisi *et al*., 1992). The RNA-binding and modeling studies of Rev and Tat suggest that RNA-protein interactions may induce conformation changes in RNA (Puglisi *et al*., 1992).

1.7.3 The RGG box

The RGG box is a 20 to 25 amino acids long RNA-binding motif, which is defined as closely spaced Arg-Gly-Gly (RGG) repeats interspersed with other, often aromatic, amino acids (Kiledjian & Dreyfuss, 1992). The RGG box was initially identified as RNA-binding motif in hnRNP U protein (Kiledjian & Dreyfuss, 1992). The minimal number of RGG repeats necessary for RNA-binding is not known. hnRNP A1 protein has as few as six whereas yeast GAR1 protein has up to 18 (Burd & Dreyfuss, 1994). The high density of glycine within the motif suggests that it is not a rigid protein structure. RGG boxes usually occur in proteins that also contain other types of RBDs. In nucleolin (with four RNP motifs and an RGG box), specific binding to pre-ribosomal RNA requires the four RNP motifs, and the RGG box increase overall RNA affinity 10-fold (Ghisolfi *et al*., 1992). This suggests that RNA binding by RGG box is relatively sequence nonspecific.

1.7.4 The KH motif

Three copies of a short conserved sequence were identified in the human hnRNP K protein (Siomi *et al*., 1993). The sequence, now known as the K homology (KH) motif, has been found in divergent organisms such archaebacteria, the yeast alternative splicing factor and several human RNA-binding proteins (Gibson *et al*., 1993). The widespread presence of KH motifs in diverse organisms suggests that it is an ancient protein structure with important cellular functions. All KH motif proteins of known function are associated with RNA and many bind RNA in vitro (Gibson *et al*., 1993; Siomi *et al*., 1993). Like many other RNA-binding motifs, KH motifs are found in one or multiple copies [14 copies in chicken vigilin (Gibson *et* *al*., 1993)]. Each motif is necessary for in vitro RNA binding activity, suggesting they may function cooperatively. An NMR study on the structure of the KH domain of human vigilin, a protein thought to be involved in tRNA transport, has shown that it contains a three-stranded β-sheet and three helices (Nagai, 1996).

1.7.5 The double-stranded RNA-binding motif (DSRM)

The DSRM is 70 amino acid region that binds double-stranded RNA (dsRNA) (Green & Mathews, 1992). Conserved positions, including many basic (both arginine and lysine) and hydrophobic amino acids, are scattered throughout the DSRM. RNA-binding experiments show that mutations of nearly each of the conserved region in DSRM affect RNA binding (Green & Mathews, 1992).

DSRM proteins are involved in diverse cellular functions. For example, an important component of the response mammalian cells to viral infection is the interferon-induced protein kinase (DAI) (Mathews & Shenk, 1991). This enzyme contains two DSRMs, binds dsRNA, and shut off host protein synthesis when activated by dsRNA (viral RNA). In an apparent attempt to abrogate this host response, some viruses encode DSRM proteins that block the activation of DAI (Johnson *et al*., 1992).

1.7.6 Zinc knuckle motif

A generalized zinc knuckle motif can be written as $CX_{2\text{-}5}CX_{2\text{-}4}$ C/H $X_{2\text{-}4}$ C/H (in which X represents any amino acid). The spaced cysteine-histidine residues in zinc knuckle motif are similar to the zinc finger family of DNA-binding proteins. The zinc knuckle motif was found in a small number of RNA-binding proteins, including retroviral nucleocapsid proteins, RNA polymerases and yeast RNA-binding proteins. The best characterized zinc knuckle motif is TFIIIA, a nine-zinc figure protein that binds both 5S rRNA gene and 5S RNA. The middle three knuckles are primarily responsible for RNA binding (Theunissen *et al*., 1992).

Finally, most unstructured proteins undergo some degree of folding upon binding to their partners, a process termed "induced folding" (Longhi *et al*., 2003). Disordered basic region devoid of defined structure can participate in condensing the viral nucleic acid genome. The nucleoprotein of measles virus consists of an N-terminal moiety, N_{CORE}, resistant to proteolysis and a C-terminal moiety, N_{TAIL}, hypersensitive to proteolysis. The N_{TAIL} undergoes such an unstructured-to-structured transition upon binding to its physiological partner, the phosphoprotein (Longhi *et al*., 2003).

1.8. Aims of research

In addition to their association with the RNA genome, the IBV N protein is thought to be a multi-functional proteins involved in viral RNA replication and transcription. Specific interactions between the N protein and viral genome were seen via the recognition of a particular nucleotide sequence by the N protein. Such particular nucleotide sequences were found to form a stem loop structure and serve as "packaging signals" which can be recognized by the nucleocapsid protein. To date, little is known about the details of this interaction between the IBV genome and the IBV N protein and how it relates to virus assembly.

In addition, the IBV N protein self-associates to form multimers *in vitro* in the absence of RNA. How the N proteins self-associate and how self-association is coupled with RNA interaction and virus assembly is elusive.

Our research focuses on functional and structural studies of the IBV N protein in an attempt to define how the viral genome is incorporated into newly formed viral particles and how this process is coupled with nucleocapsid assembly.

CHAPTER 2. MATERIALS AND METHODS.

2.1 Cloning and expression of the IBV N protein

The full length N protein of IBV (strain Beaudette) consists of 409 amino acids with a molecular weight of 45.03 kDa. The gene (1227 nucleotides) encoding the IBV-N protein was amplified by PCR using the Pfu polymerase (Stratagene) with the forward (5'-ATT ATT CAT ATG GCA AGC GGT AAA GCA GC-3') and reverse primer (5'-ATTATT CTC GAG TCA AAG TTC ATT CTC TCC TA-3') and cloned into the pET 29b (Novagen) vector using T4 ligase (Research Biolabs). The underlined sequences correspond to NdeI and XhoI sites, respectively. Proteins (lacking the His$_6$ tag due to the insertion of a stop codon in the reverse primer) were expressed in *E. coli* BL21(DE3). The cells were grown at 37 ºC in Luria-Bertani medium containing 100 mg/mL Kanamycin until the culture reached an OD600 of 0.7. Protein expression was induced by the addition of 1 mM isopropyl- β -D-thiogalactopyranoside (IPTG) for 3 hr at 30 ºC. Cells harvested and resuspended at 4 ºC in a buffer containing 20 mM Na$_3$PO4 (pH 7.8) were lysed by sonication and the remaining insoluble material was removed by centrifugation at 20,000 rpm for 20 min at 4 ºC.

2.2 Purification of the IBV N protein

The IBV-N protein precipitated with ammonium sulfate at 30% saturation, was centrifuged and resuspended in PBS, dialyzed against buffer A (20 mM HEPES [pH 6.8], 1 mM EDTA, 1 mM DTT), and loaded onto a cation exchange chromatography column (Mono S HR 5/5; GE Biosciences) preequilibrated with buffer A. Elution was carried out using an NaCl gradient of buffer B (20 mM HEPES [pH 6.8], 1 mM EDTA, 1 mM DTT, 1 M NaCl). Fractions containing the protein—as shown by SDS-PAGE—were pooled and concentrated to 10–15 mg/mL by ultrafiltration using a Centriprep device (Millipore) with a molecular weight cutoff of 10 kDa. Size exclusion chromatography (Superdex 200; Amersham) was carried out in a buffer containing 20 mM Tris-HCl (pH 8.0), 150 mM NaCl, 1 mM DTT, 0.1 % NaN_3. The protein was concentrated to 10 mg/mL as determined by the Bradford assay (Bio-Rad), using BSA as a standard.

2.3 Dynamic light scattering

The IBV N protein was concentrated to 10 mg/mL after purification in buffer: 20 mM Tris pH 8.0, 150 mM NaCl, 1mM DTT. The N protein aggregation status and homogeneity was investigated by Dynamic Light Scattering (DLS) to determine the size distribution of protein particles in solution.

DLS uses a coherent and monochromatic light source from a laser to observe time-dependent fluctuations in the scattered intensity. These fluctuations arise from the fact that the particles are small enough to undergo random thermal (Brownian) motion and the distance between them is therefore constantly varying. Constructive and destructive interference of light scattered by neighbouring particles within the illuminated zone gives rise to the intensity fluctuation at the detector plane which, as it arises from particle motion, contains information about this motion. Analysis of the time dependence of the intensity fluctuation can therefore yield the diffusion coefficient of the particles from which, via the Stokes-Einstein equation ($D = kT / 6\pi\eta R_h$), knowing the viscosity (η) of the medium at a temperature (T), the hydrodynamic radius (R_h) or diameter of the particles can be calculated from Stokes-Einstein equation. This hydrodynamic size of protein particles depends on both mass and shape (conformation). According to the average size of the protein particle calculated from DLS, we know whether the protein aggregates or not.

2.4 Negative staining electron microscopy of the IBV N protein

2 μL of purified IBV N protein with a concentration 0.5 mg/mL was placed on a pre-washed, glow-discharged copper grid. Following 2 min of sample absorption and washing of water, 2 μL of 1% uranyl acetate stain was applied. After 4 min staining, the grid was dried using Whatman filter paper. Samples were then viewed with an electron microscope (JEOLJEM 1010, operating at 100 kV).

2.5 Analysis of the IBV N protein and its two proteolytically stable fragments (N-terminal domain IBV-N29-160 and C-terminal domain IBV-N218-329)

The IBV N protein with a concentration of 10 mg/ml in buffer: 20 mM Tris pH 8.0, 100 NaCl, 1mM DTT was analyzed using a MALDI-TOF mass spectrometer (API 300 MS/MS; Applied Biosystems) to determine its molecular mass. Automated N-terminal amino acid sequence determination of the proteolytic fragments obtained by degradation of the IBV-N was performed using an Applied Biosystems Precise sequencer.

2.6 Cloning and expression of IBV-N29-160 and IBV-N218-329

N- and C-terminal fragments of the IBV-N gene coding for residues 29–160 and 218–329, respectively, were cloned into pET-16b using the following primers: 5'-AATA CATATG TCT TCT GGA AAT GCA TCT TG-3'; 50-AATA CTC GAG TCA CAG GGG AAT GAA GTC CC-30 and 5'-A AATA CAT ATG AAG GCA GAT GAA ATG GC-3'; 5'-AA ATA CTC GAG TCA CGT TCC TAC ACC ATC GAC-3'. These two proteins (hereafter named IBV-N29-160 and IBV-N218-329, respectively) were expressed in *E. coli* BL21(DE3). The cells were grown at 37ºC in Luria-Bertani medium containing 100 mg/mL ampicillin until the culture reached an OD600 of 0.7. Protein expression was induced by the addition of 1 mM isopropyl-β-D-thiogalactopyranoside (IPTG) for 3 hr at 30 ºC. Cells were harvested and resuspended at 4 ºC in a buffer containing 20 mM Na$_3$PO4 (pH 7.8) were lysed by sonication and the remaining insoluble material was removed by centrifugation at 20,000 rpm (48384 g) for 20 min at 4 ºC. The his-tag fusion proteins were firstly purified by Ni-NTA column (Qiagen) followed by ion exchange (Mono S, Amersham) and size exclusion chromatography (Superdex 75, Amersham). The N-terminal His_{10} tag followed by a Factor Xa cleavage site was cleaved during purification. Expression of the selenomethionylated protein IBV-N29-160 was carried out as described in Doublie (1997).

2.7 Size exclusion chromatography of the IBV N protein, IBV N29-160 and IBV N218-329

100 μL of the purified IBV N protein with a concentration 10mg/mL was loaded onto a size exclusion chromatography column Superdex 200 (Amersham) and eluted in buffer: 0.2 M NaCl, 20 mM Tris pH 8.0, 1 mM DTT. The BSA dimer (M.W. 134 kDa) was used to calibrate the column and compared to the IBV N protein. A Superdex 75 10/300 GL size exclusion chromatographic column (Amersham) mounted on an AKTA FPLC (GE Biosciences) was used to analyze the homogeneity and apparent multimerization states of IBV-N29-160 and IBV-N218-329, respectively with the same protein concentration. The buffer was 10 mM Tris-HCl, 0.2 M NaCl, 3 mM β-mercapto-ethanol (pH 7.5) and the flow rate was 0.5 mL/min. Standard protein markers (Amersham) used for calibration were ribonuclease A, 13.7 kDa, elution 14.88 mL; chymotrypsinogen A, 25.0 kDa, 13.81 mL; ovalbumin, 43.0 kDa, 11.81 mL; BSA, 67.0 kDa, 10.89 mL. Apparent

size/molecular weights were deduced by plotting Kav versus log (MW) with Kav = (Ve - Vo)/(Vt -Vo), where Ve is the elution volume of the protein, Vt is the total column bed volume, and Vo is the void volume.

2.8 Crosslinking experiments

The purified recombinant proteins IBV-N, IBV-N29-160, and IBV N218-329 were incubated with either glutaraldehyde or suberic acid bis N-hydroxy-succinamide ester (SAB) (Sigma-Aldrich) for 2 hr at 20 ºC using a constant amount of protein (5 mg) with increasing amounts of the crosslinking agent. The samples were submitted to electrophoresis on an 8%– 15% SDS-PAGE gel and stained with Coomassie blue.

2.9 RNA binding assay

The full-length IBV-N protein and the IBV-N29-160 and IBV-N218-329 fragments were expressed in *E. coli* BL21(DE3) cells and purified as described above. The polyhistidine tags of the truncated proteins were removed by digestion with Factor Xa. The purified proteins were separated on 15% SDS-PAGE, transferred to Hybond C extra membrane (Amersham), and probed with digoxin-labeled RNA representing the negative sense of the IBV genome from nucleotides 25,873–27,608. The probe was made by in vitro transcription using SP6 polymerase in the presence of digoxin according to the manufacturer’s instructions (Roche).

2.10 Crystallization of IBV N29-160 and IBV N218-329

After purification, IBV N29-160 and IBV N218-329 were concentrated to 10 mg/mL in buffer: 20 mM Tris pH 8.0, 100 mM NaCl, 1 mM DTT. Crystals of the recombinant IBV-N29-160 and IBV N218-329 were grown at 18 ºC by vapor diffusion using the hanging drop method. The crystal screens for two proteins were carried out in the conditions provided by Hampton research protein crystal screen kit 1, 2 and PEG/Ion kit. Two microliters of the protein at a concentration of 10 mg/mL was mixed with an equal volume of the precipitating solution. Crystals of IBV N29-160 and selenomethionylated IBV N29-160 were obtained in condition: 0.1 M sodium sulfate, 20% PEG 3350, yielding plate-shaped crystals growing to maximum dimensions of about 0.3 × 0.3 × 0.05 mm^3 in about 2 weeks. IBV-N218-329 crystals grow in condition: 4.3 M NaCl, 0.1 M Tris pH 8.5 with dimensions 0.05×0.05×0.1 mm^3 in about 2 weeks.

2.11 Data Collection, Structure Determination, and Refinement

For data collection, IBV N29-160 crystals were soaked in a cryoprotecting solution (25% glycerol, 0.1 M sodium sulfate, 20% PEG 3350 [pH 6.5]) before being mounted and cooled to 100 K in a nitrogen gas stream (Oxford Cryosystems, Oxford, UK). Diffraction intensities at three wavelengths were recorded from a selenomethionine (SeMet)-derivatized IBV-N29-160 crystal on beamline NW12 at the Photon Factory (Tsukuba, Japan) on charge-coupled device (CCD) detector (ADSC Corporation) using an attenuated beam of dimensions 0.1×0.1 mm^2 .

Integration, scaling, and merging of the intensities were carried out using program MOSFLM and SCALA from the CCP4 (1994). The six selenium atoms present within the two molecules of the asymmetric unit were located using the program SOLVE (Terwilliger, 2003). An initial electron density map was calculated and modified using the program RESOLVE (Terwilliger, 2003), using these selenium atom positions to locate the noncrystallographic symmetry (ncs) axis relating the two molecules in the asymmetric unit, and model building was first carried out in this map using the program O (Jones *et al*., 1991). For subsequent cycles, electron density maps were calculated using partial model phases combined with experimental MAD phases with the program REFMAC5 from the CCP4 (1994), which was used for the initial refinement of the structure, that included ncs restraints. A few cycles of refinement using molecular dynamics with a slow cooling protocol using a maximum likelihood target incorporating phase probability distribution encoded in the form of Hendrickson Lattman coefficients were subsequently carried out using the program CNS (Brunger *et al*., 1998), with ncs restraints. A data set for the native IBV-N29-160 was collected on an R axis IV++ image plate detector using CuKa radiation from a Micromax-007 rotating anode (Rigaku/MSC) operating at 20 mA and 40 kV. The SeMet model was placed in the native crystal form and adjustments to the model were carried out using difference Fourier maps calculated with REFMAC5, which was used for refinement. Superposition of structures and rms deviation calculations were carried out using the program LSQKAB from the CCP4 (1994).

The native IBV N218-329 crystals were soaked in a cryoprotecting solution of 20% glycerol, 50 mM Tris pH 8.5, 3.7 M NaCl before being mounted and cooled to 100K in a nitrogen gas stream (Oxford cryosystems). Diffraction intensities were recorded on beamline ID14-4 at European Synchrotron Radiation Facility (Grenoble, France) on an ADSC charge-coupled device detector. Data integration, scaling, and merging of the intensities were done by program MOSFLM and SCALA from CCP4 suite (1994). The initial model was built by molecular replacement (MR) from IBV (strain Gray) C-terminal domain (residue 226-333) coordinates kindly provided by Professor Prasad, B. V. Venkataram at Baylor Medical College. Sequences difference between IBV Gray and Beaudette strain in N protein C-terminal domain are shown in alignment (Fig.3.22). The two C-terminal domains of the N protein share 93.5% similarity in sequence between the strains Beaudette and Gray and their structures can be superimposed with an r.m.s deviation of 0.57 Å. Manual model building was carried out by using COOT (Emsley *et al.*, 2004). Model refinement was performed using a combination of REFMAC5 and simulated annealing of CNS (Brunger *et al.*, 1998). The stereochemistry of the structures was checked with PROCHECK (Laskowski *et al.*, 1993). Electrostatic solvent-accessible surfaces were generated by PyMOL. Superposition of structures and rms deviation calculations were carried out using the program COOT (Emsley *et al.*, 2004). All the figures were produced with the program PyMOL (DeLano, 2002).

CHAPTER 3. RESULTS AND DISCUSSIONS.

3.1 The IBV N protein

3.1.1 Expression and purification

The IBV N protein (45 kDa) without any affinity tag at termini was expressed in *E. coli* and purified by 30% ammonium sulfate precipitation, ion exchange chromatography (Mono S column) and size exclusion chromatography (Superdex 200 column) (Fig. 3. 1). After purification, the protein purity was more than 95% and suitable for crystallization trials and other assays.

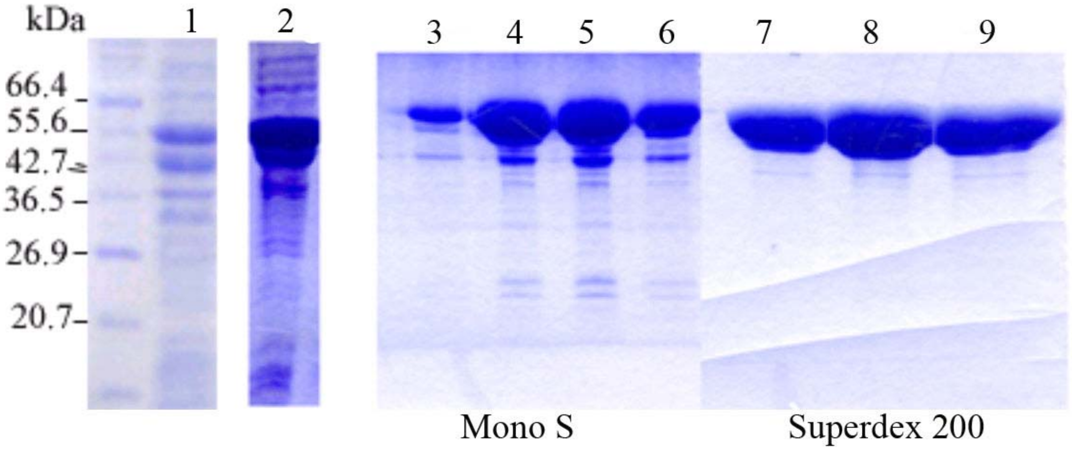


Fig. 3. 1. Expression and purification of the IBV N protein.
IBV N protein was expressed in *E.coli* and purified by 30% ammonium sulfate precipitation, Mono S and Superdex 200 column.
Lane 1: Total protein after 1 mM IPTG induction.
Lane 2: 30% ammonium sulfate precipitation.
Lanes 3-6: Mono S column purification
Lanes 7-9: Superdex 200 column purification.

3.1.2 Mass spectrometry

Mass spectrometry shows the molecular mass of IBV N protein is 45.040 kDa, which is consistent with the 45.032 kDa calculated from IBV N protein sequence (Fig. 3. 2). The difference between the measured mass weight and calculated mass weight is within the error range for the MALDI-TOF analysis.

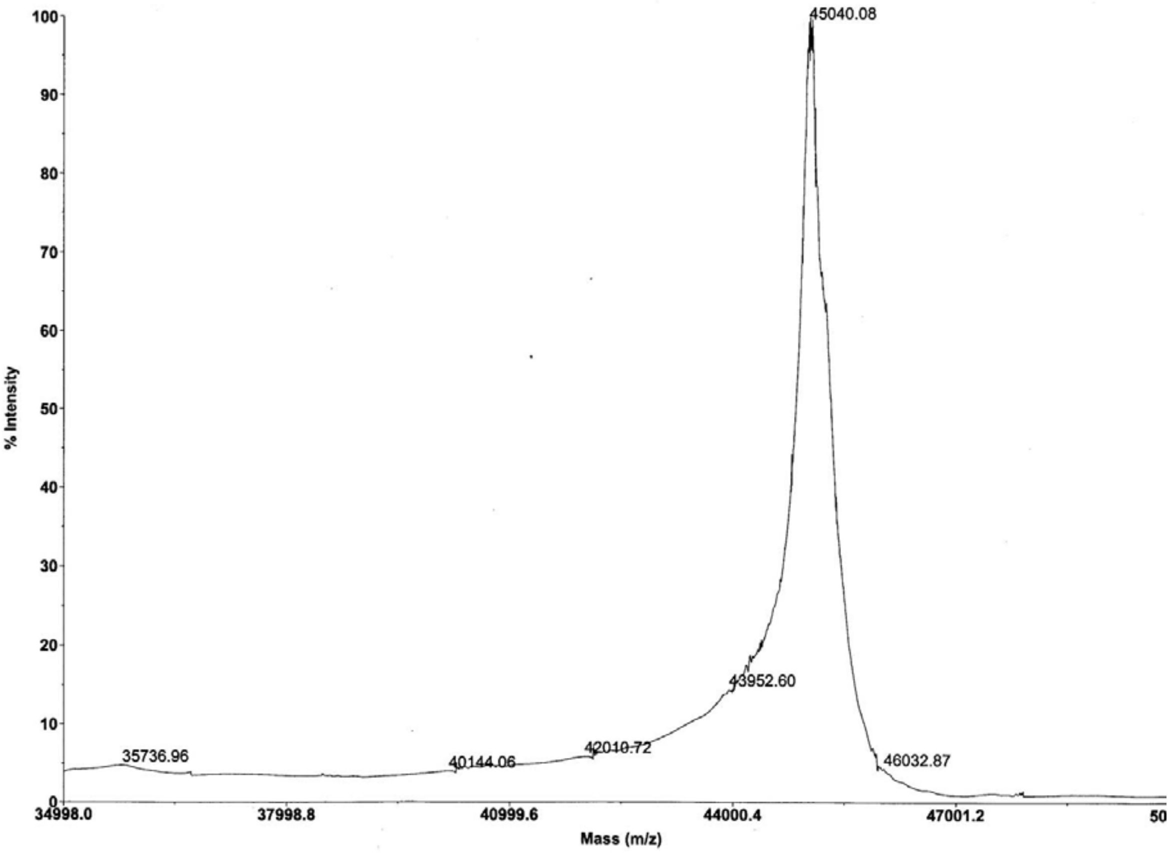


Fig. 3. 2. Mass spectrometry of the IBV N protein
The mass spectrometry shows the molecular mass of the IBV N protein (monomer) is 45.04 kDa.

3.1.3 Dynamic Light Scattering (DLS)

Biomolecular behaviour of proteins in solution could be measured using dynamic light scattering. Protein molecules move randomly in solution, thus causing different intensities of light to scatter from the moving molecule when exposed to light. This leads to time-dependent fluctuations in the intensity of the scattering light. The fluctuation correlates directly to the rate of diffusion of the molecule through the solvent. The fluctuation therefore can be analyzed to determine the hydrodynamic radius of the protein molecule in solution.

The average size of protein particle calculated from DLS is 16.6 nm in diameter (8.3 nm in radius) (Fig. 3. 3). The BSA (monomer: 67 kDa), a general standard used for DLS, only has a diameter of 5-6 nm. Therefore, it is not likely the 45 kDa N protein can form the 16.6 nm diameter particle without aggregation or multimerization.

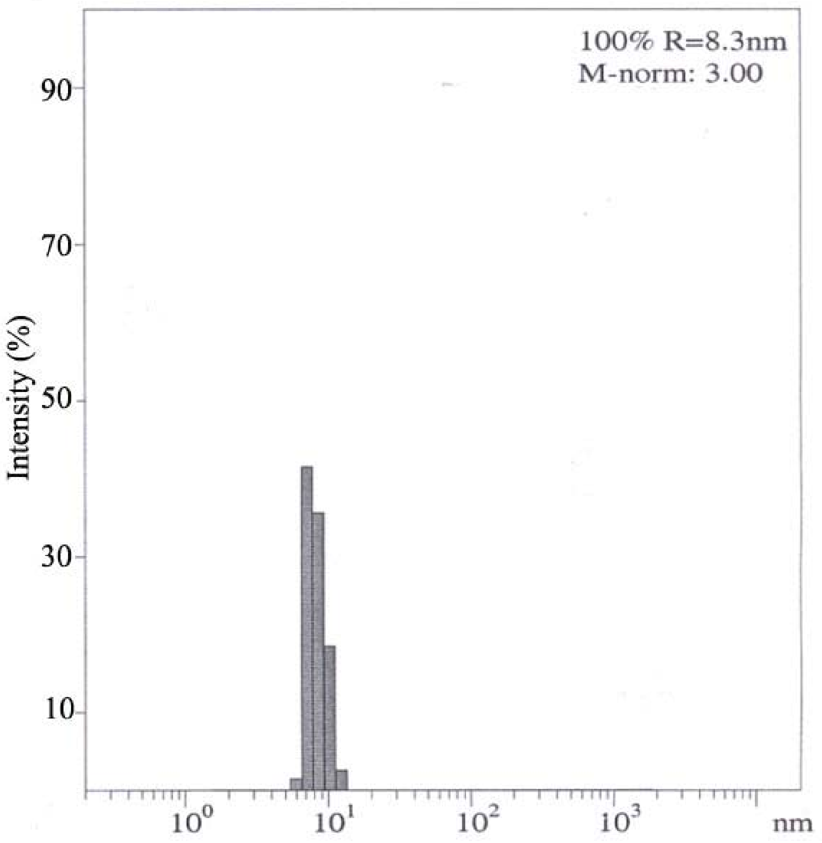


Fig. 3. 3. Dynamic light scattering of the IBV N protein
DLS shows the IBV N protein in solution has an average size of 16.6 nm in diameter (8.3 nm in radius).

3.1.4 Electron microscopy

Negative staining electron micrography shows that the N protein forms regular round particles with a diameter ranging from 8 nm to 16 nm (Fig. 3. 4). This is consistent with the observation by DLS that the N protein has a size around 16 nm in diameter. In addition, this result indicates that the N protein is likely to form multimers which are soluble fractions rather than aggregation which generally causes heterogeneous distribution of protein and therefore precipitation.

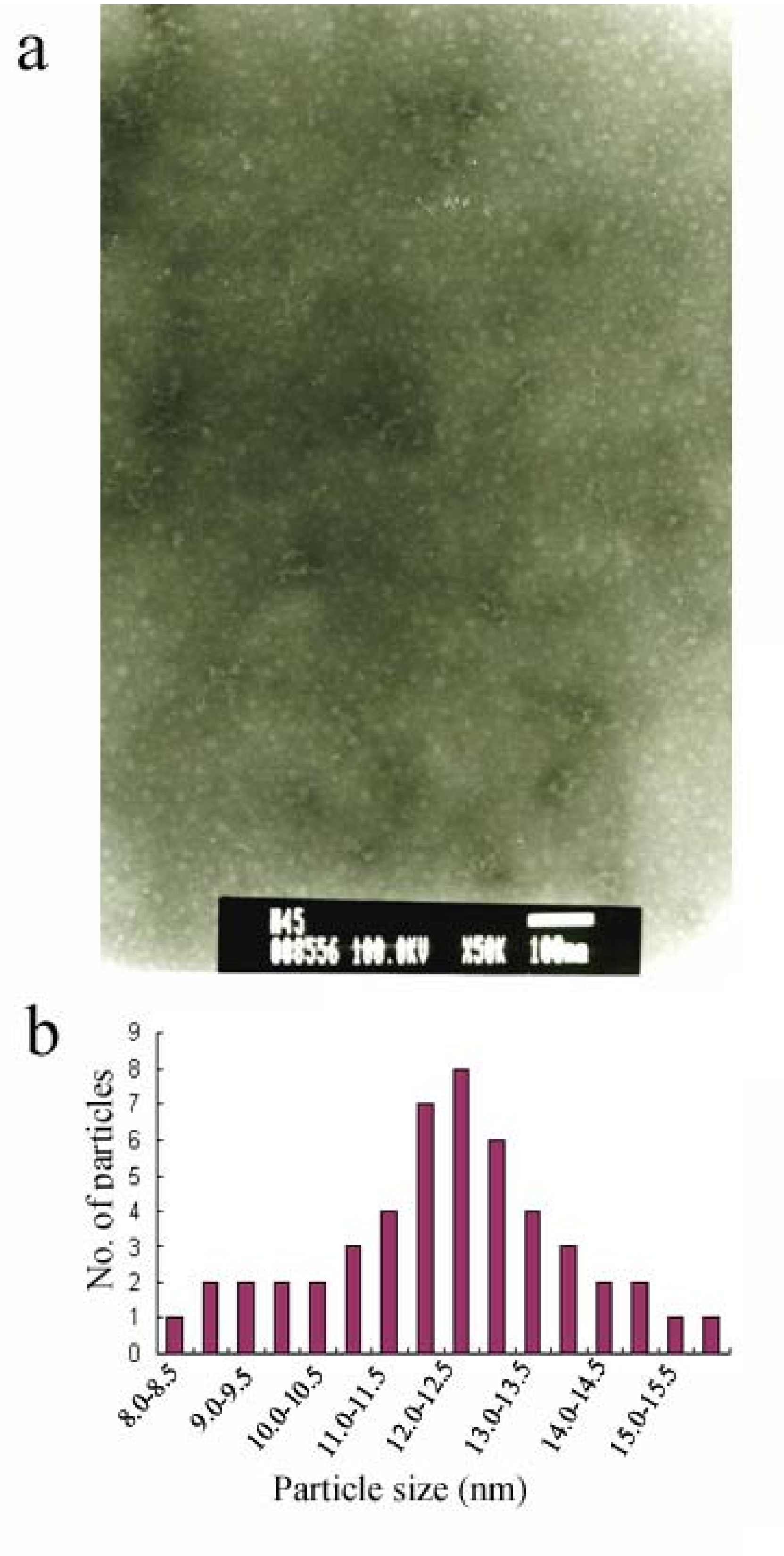


Fig. 3. 4. Electron microscopy of the IBV N protein.

The bar has a length of 100 nm in **(a)**. 50 Protein particles in the EM photography were randomly chosen and measured according to this calibration. The diameter of the N protein particles ranges from 8 nm to 16 nm in **(b)**.

3.1.5 Size exclusion chromatography

The IBV N protein homogeneity and apparent multimerization states were further analyzed by size exclusion chromatography (Superdex 200). Standard protein markers (Amersham) were used for calibration. Based on the elution profile (Fig. 3. 5 a), A linear standard curve (Fig. 3. 5 b) was deduced with lg M.W ($\log_{10}$ M.W) as Y axis and Kav as X axis.

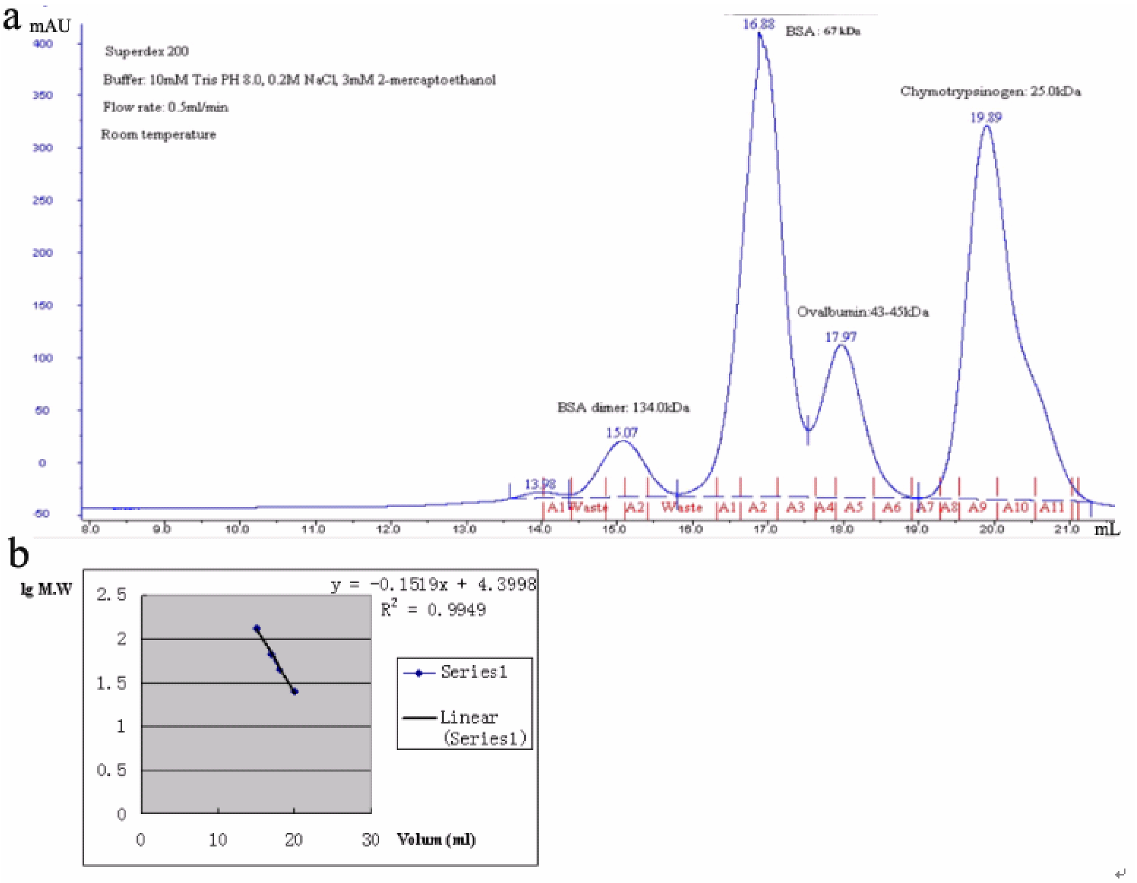


Fig. 3. 5. Standard protein marker for Superdex 200 column (a) and linear standard curve for calculating M.W. (b).

The marker used in (a) were BSA dimer (134 kDa), BSA monomer (67 kDa), ovalbumin (43 kDa), chymotrypsinogen A (25 kDa).

The IBV N protein elution profile (Fig. 3.6) shows that the N protein was eluted out much earlier than the BSA dimer (134 kDa) with a flow rate 0.5 mL/min. This suggests that the N proteins either form much larger multimer than a dimer or is a elongated molecule with a large Stoke's radius. The apparent M.W of the N protein calculated from linear standard curve showed in (Fig. 3.5b) was 235 kDa which already exceed the maximum M.W Superdex 200 column can measure (200 kDa maximum).

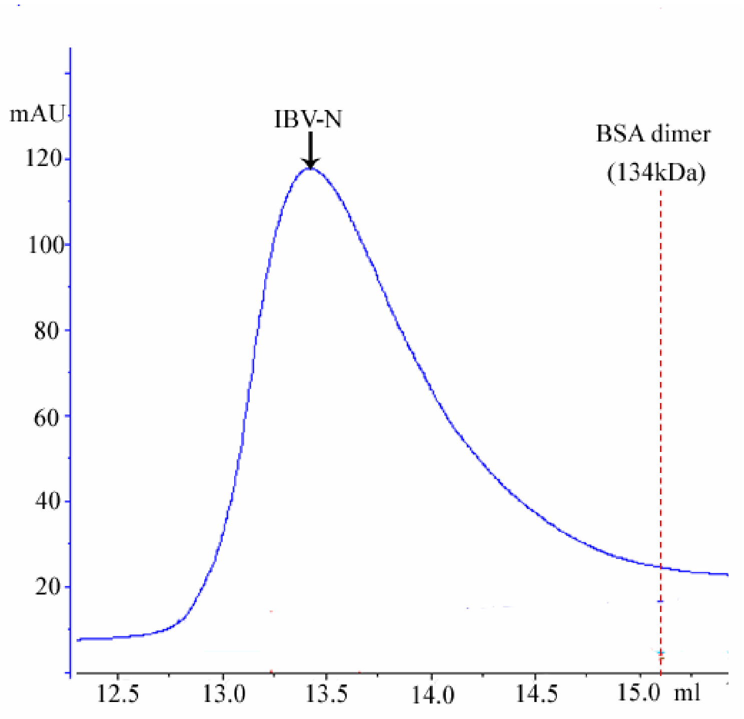


Fig. 3. 6. Size exclusion chromatography of the IBV N protein.
With a flow rate of 0.5 mL/min, the IBV N protein was eluted out much earlier than BSA dimer, which showed the IBV N has a much larger M.W than BSA dimer's M.W 134 kDa.

Even though an accurate apparent M.W of the N protein has not been measured, the profile provides some information to understand the N protein's multimerization state. By comparison with BSA marker (134 kDa), the N protein presumably forms a hexamer although we cannot exclude an octamer.

3.1.6 Crosslinking

Crosslinking experiments were performed using glutaraldehyde, a short self-polymerizing reagent mostly reacting with the amino and amine groups of lysine and histidine, respectively (Buehler *et al.*, 2005), and suberic acid bis N-hydroxy-succinamide ester (SAB) (Sigma-Aldirch), a reagent which only crosslinks lysine residues at larger distances. Concentrations of crosslinking agent higher than 0.1 mM led to the formation of dimers, tetramers (but not trimers), and larger oligomers of IBV-N, along with the disappearance of monomeric species (Fig. 3. 7).

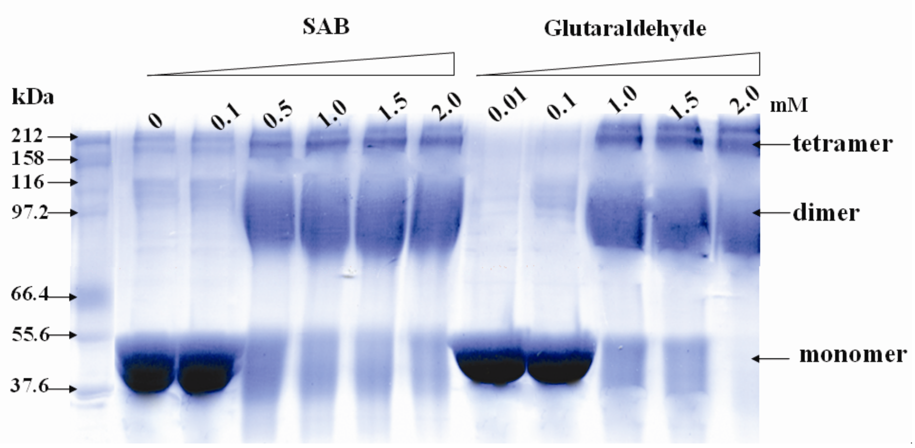


Fig. 3. 7. Crosslinking of the IBV N protein
Purified N proteins were incubated with increasing amount of either glutaraldehyde or SAB for 2 hours at 20 °C. Multimerization status of the N proteins was checked by SDS-PAGE gel stained by coomassie blue.

3.1.7 Discussion

The 45 kDa IBV N protein expressed in *E.coli* is soluble and forms multimers as shown by DLS and size exclusion chromatography. DLS shows the N protein form particles with average diameter of 16 nm which is consistent with the negative staining E.M results. Monomer of the N protein is not likely to form a particle of this size in solution because the protein standard BSA (monomer: 67 kDa) used for DLS only has a diameter of 5-6 nm. This average size of protein particles calculated by DLS depends on both the protein molecular weight and its shape (conformation). We can calculate the M.W of the N protein from DLS by comparing average size of the N protein to that of BSA standard (monomer 67 kDa) which normally has a diameter of 5-6 nm. However, this calculation is based on the assumption that both two proteins have the same shape. Moreover, the N protein's heterogeneity caused by different multimerization stages and rapid degradation will affect the accuracy of the M.W calculated from DLS. Thus, we can not precisely calculate an apparent molecular weight of the IBV N protein from DLS. DLS coupled with analytical ultracentrifugation can monitor the properties of macromolecules in solution and provide information about the oligomeric state of the protein.

The N protein seems to form a multimer higher than dimer with comparison to dimer BSA marker (134 kDa) because the N protein was eluted much faster than BSA dimer in size exclusion chromatography. Coupled with the crosslinking result that N protein can form dimer, tetramer and higher multimer, but not trimer, the N protein most likely forms tetramer, hexamer or octamer.

3.2 Structure determination of the IBV N protein

3.2.1 Structural organization of the IBV N protein

The IBV N protein consists of two globular domains: the N-terminal domain (residues 29-160) and the C-terminal domain (residues 218-329), which were determined by limited proteolysis by trace of proteases from *E.coli*. The protein N-terminal sequencing and mass spectrum allowed the identification of the two domain's boundaries. The structures of the two domains were solved by crystallography. There is a flexible loop (58 residues in length) rich in arginine, glycine and serine between the N- and C-terminal domains. At N-terminus and C-terminus of the N protein, there are two flexible tails: N-tail (residues 1-29) and C-tail (residues 329-409), respectively (Fig. 3. 8a). All these flexible domains might become ordered in presence of nucleic acid in the context of the nucleocapsid.

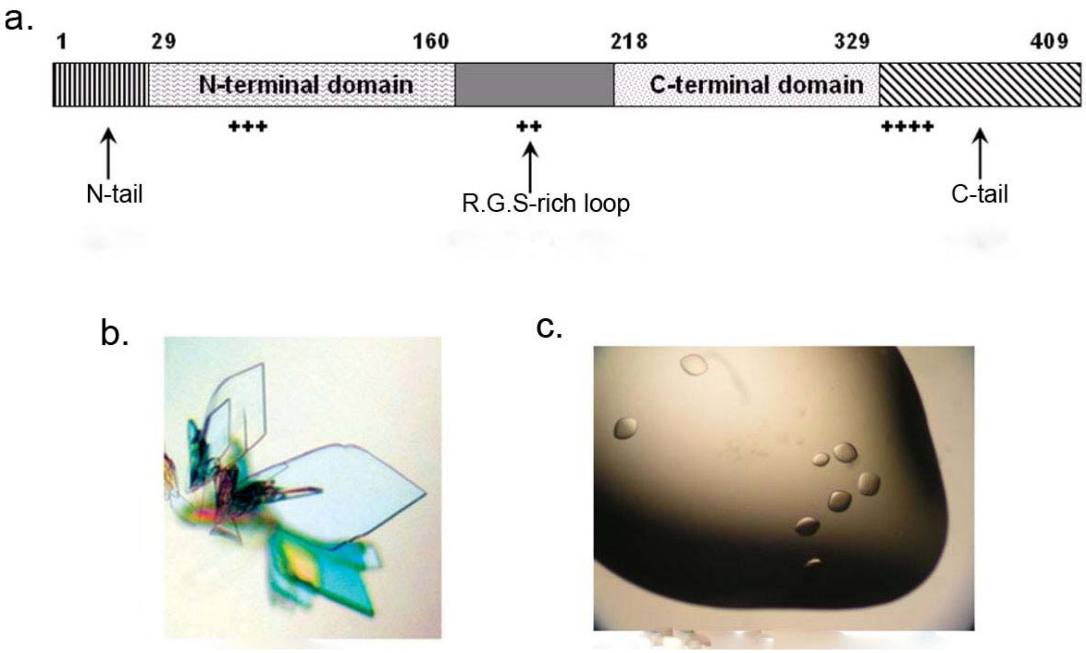


Fig. 3. 8. Structural organization of the IBV N protein (a), the N-terminal domain crystal (b) and the C-terminal domain crystal (c).

The N-terminal domain (residues 29-160) and C-terminal domain (residues 218-329) of the IBV N protein can form 2D plate crystals (b) and single lens-shaped crystals (c), respectively.

The two proteolytically stable domains can be crystallized in totally different conditions. The crystal morphology of the N-terminal domain is large 2D plates in contrast to the C-terminal domain's crystal which is single small lens-shaped crystals (Fig. 3. 8).

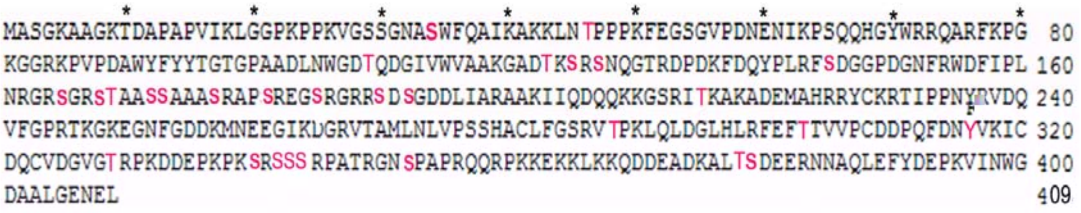
The presence of flexible regions in the IBV N protein's structure, including R.G..S loop and N or C tail (Fig. 3. 8a), are obstacles to crystallize the full length N protein. More importantly, IBV N protein has multimerization property as shown by DLS (Fig. 3. 3) and size exclusion chromatography (Fig. 3. 6). Different multimerization stages may lead to tetramer, hexamer or even higher multimer formation (Fig. 3. 7). Consequently, it seems not possible to crystallize the full length protein because of the heterogeneity caused by the N protein's multimerization behavior. .

The expression vector encoding the N-terminal region (residue 29-160) or C-terminal region (residue 218-329) was expressed in *E.coli* and purified by different chromatographic methods. The N- and C-terminal domains were named IBV N29-160 and IBV N218-329, respectively.

The structures of the N- and C-terminal domains were solved by different strategies. The N-terminal sequence is devoid of methionine and cysteine. In order to solve the phase problem, three residues (62 Ile, 104 Leu and 116 Val) were mutated to methionine in order to incorporate selenomethionine. The lysine 85 was mutated to cysteine to introduce a potential mercury binding site. The N-terminal domain with four site mutations can be crystallized in the same condition as the native protein. The structure of the N-terminal domain was finally solved using MAD data. The C-terminal domain was solved by molecular replacement. The initial searching model is the IBV (strain Gray) C-terminal domain.

The N tail, C tail and R.G.S rich loop make up 40% of the IBV N protein in length. The rest 60% of the N protein are composed of the N- and the C-terminal domains. The structures of the N- and C-terminal domains provide important information for the N protein's RNA-binding and self-association. Furthermore, the N tail, C tail and R.G.S rich loop may also have biological functions.

Interestingly, the R.G.S rich loop (resides 161-217) contains 9 serines which are potential sites for phosphorylation as predicted by program NetPhos (Blom et al, 1999). 19 serines, 9 threonines and 1 tyrosine in the N protein sequence are predicted to be phosphorylation sites (Fig. 3. 9).



```
         *         *         *         *         *         *         *         *
MASGKAAGKTDAPAPVIKLGGPKPPKVGSSGNASWFQAIKAKKLNTPPPKFEGSGVPDNENIKPSQQHGYWRRQARFKPG  80
KGGRKPVPDAWYFYYTGTGPAADLNWGDTQDGIVWVAAKGADTKSRSNQGTRDPDKFDQYPLRFSDGGPDGNFRWDFIPL 160
NRGRSGRSTAASSAAASRAPSREGSRGRRSDSGDDLIARAAKIIQDQQKKGSRITKAKADEMAHRRYCKRTIPPNYRVDQ 240
VFGPRTKGKEGNFGDDKMNEEGIKDGRVTAMLNLVPSSHACLFGSRVTPKLQLDGLHLRFEFTTVVPCDDPQFDNYVKIC 320
DQCVDGVGTRPKDDEPKPKSRSSSRPATRGNSPAPRQQRPKKEKKLKKQDDEADKALTSDEERNNAQLEFYDEPKVINWG 400
DAALGENEL                                                                        409
```

Fig. 3. 9. Phosphorylation site prediction of the IBV N protein

As shown by red color, 19 serines, 9 threonines and 1 tyrosine in the IBV N protein sequence are predicted to be phosphorylation sites. 9 serines in Arg/Gly/Ser/ (R.G.S) rich loop (residues 161-217) are predicted to be phosphorylated sites.

The IBV N protein is a phosphorylated protein. This can be demonstrated by treating the viral N protein with protein phosphatase (Fig. 3. 10). In SDS-PAGE gel, the N protein purified from the virus has a larger M.W compared to recombinant N protein which has no post-translational modification. Protein phosphatase treatment can reduce the viral N protein M.W as shown in lane 2 (Fig. 3. 10).

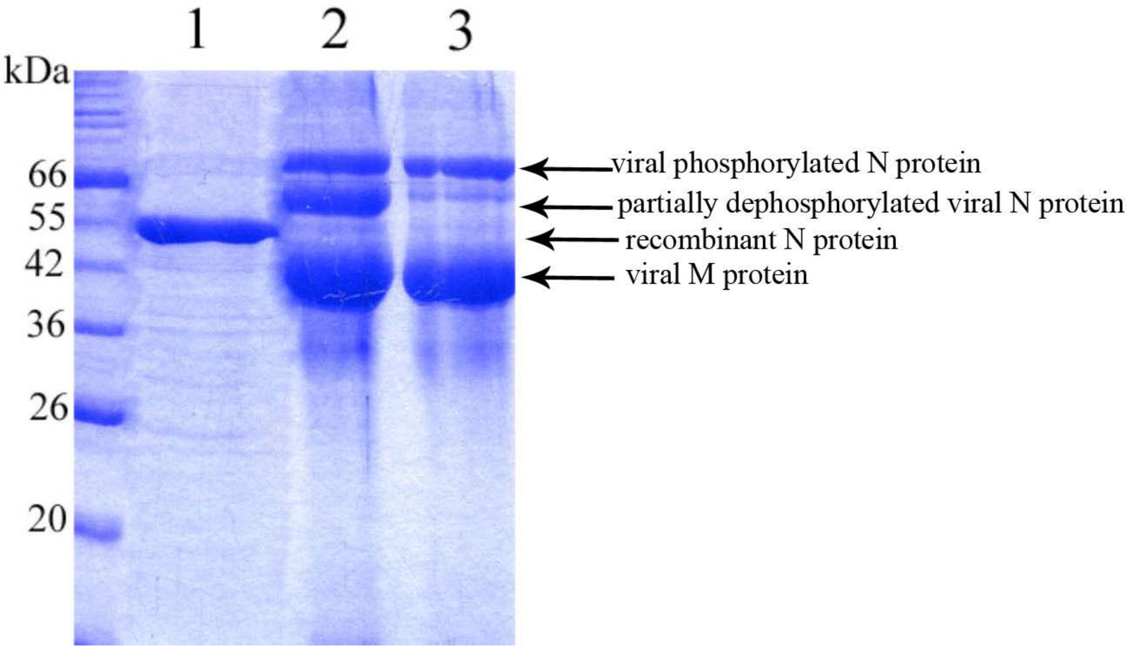

1
2
3
kDa
66
55
42
36
26
20
viral phosphorylated N protein
partially dephosphorylated viral N protein
recombinant N protein
viral M protein

Fig. 3. 10. Comparison of recombinant N protein to viral phosphorylated N protein and partially dephosphorylated viral N protein.

Lane 1: the recombinant 45 kDa N protein
Lane 2: the viral phosphorylated N protein and partially dephosphorylated viral N protein after protein phosphatase treatment.
Lane 3: the viral phosphorylated N protein.

3.2.2 Identification of proteolytically stable N-terminal domain

The full-length IBV-N protein comprising 409 residues was expressed in *E. coli* in a soluble form and purified as described in the Experimental Procedures.

The protease inhibitor cocktail (Roche) was added before the bacteria were lysed by sonication. During last step purification by gel filtration chromatography, we collected the fractions only from the sharpest peak of gel filtration elution. The fractions from the smooth peak of elution were not included because the degraded proteins or/and different multimers of the IBV N protein might be included in those fractions.

We added protease inhibitor cocktail (Roche) to the protein solution before setting up crystal trials. Limited proteolysis can cleave the flexible region of the protein. Even though there are some traces of proteases from the *E. coli* in the protein, the proteases can cleave the flexible region of full length N protein and therefore help crystallize partial fragments of the protein.

The crystal screen was set up by vapor diffusion, in which the protein and precipitate solution are allowed to equilibrate in a closed container with a larger aqueous reservoir. Because the precipitate concentration is the major solute present, vapor diffusion in the close system results in net transfer of water from the protein solution to the reservoir, until the precipitate concentration is the same in both solutions. During this process, the protein may form crystals because the water in the protein solution is evaporating.

Crystallization trials with the full-length IBV N protein produced crystals that grew from a precipitate (PEG 4000) after about 3 months. Analysis of dissolved crystals using SDS-PAGE reveals that they contain a fragment of the full-length protein of about 14.7 kDa (Fig. 3. 11). A domain of similar size could be obtained by incubating the IBV-N protein at room temperature for the same period (Fig. 3. 11). Thus, this polypeptide fragment presumably derives from slow proteolysis of IBV-N by traces of *E. coli* proteases present in the crystallization solution. In order to identify its nature, this proteolytically stable fragment was subjected to mass spectrometry, which revealed a mass of 14,692 Da (Fig. 3. 12). N-terminal amino acid sequencing identified residues Ser-Ser-Gly-Asn-Ala-Ser-Trp, which are located at positions 29–35 of the IBV-N amino acid sequence. Given that Ser-29 is the first amino acid of the fragment, the closest mapping onto the sequence gives Leu-160 as the C-terminal residue (calculated mass 14,691 Da). The IBV-N29-160 protein shares 37% amino acid sequence identity with the N-terminal RNA binding domain of a comparable domain from the SARS-CoV N protein (SARS-N45-181).

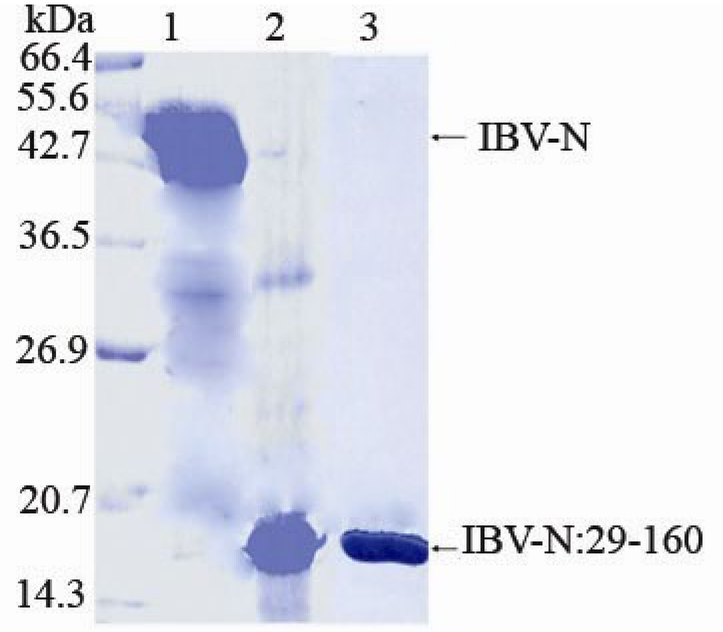


Fig. 3. 11. SDS-PAGE analysis of the full-length recombinant IBV-N protein and 14.7 kDa proteolytically stable fragment which can be crystallized.

Lane 1: the full length IBV N protein (45.0 kDa).

Lane 2 shows that dissolved crystals analyzed by SDS-PAGE gel and stained by coomassie blue is partial fragment about 14.7 kDa of the full length IBV-N protein.

Lane 3 shows that a similar size fragment of protein can be obtained by incubating the full length IBV-N protein at room temperature for the same period.

Full length 45.0 kDa N protein and the N-terminal proteolytically stable fragment of 14.7 kDa spanning residues 29–160 of the sequence which was crystallized are shown in lane 1 and lane 2, respectively. A similar size fragment of protein can be obtained by incubating the full length IBV-N protein at room temperature for the same period (lane 3). The SDS-PAGE gel was stained by coomassie blue.

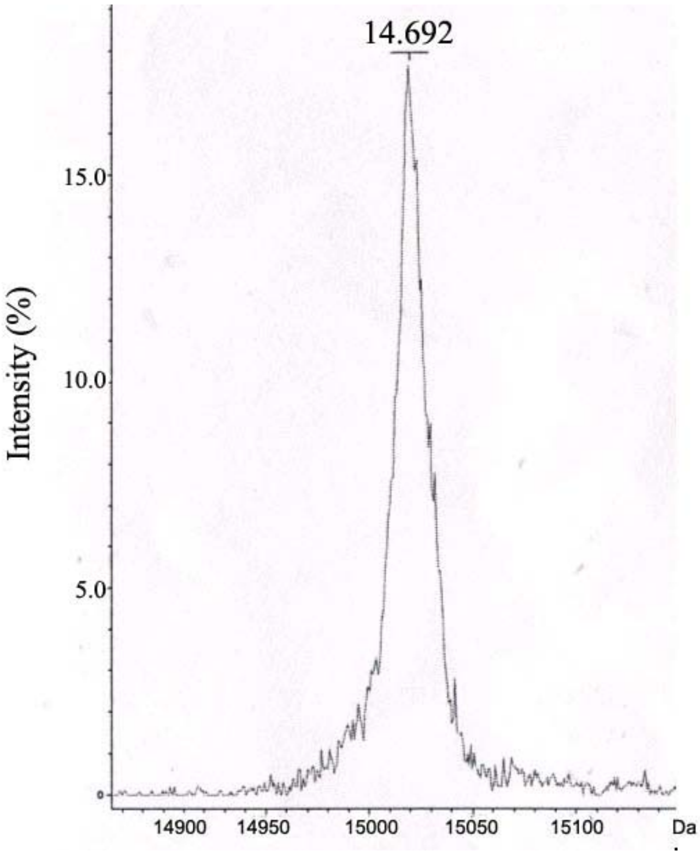


Fig. 3. 12. Mass spectrometry of 14.7 kDa proteolytically stable fragment.

3.2.3 Structure determination of the N-terminal domain and quality of the model

The recombinant IBV-terminal domain (IBV-N29-160) can be expressed in *E. coli* BL21(DE3). Purified recombinant IBV-N29-160 can be crystallized in condition: 0.1 M sodium sulfate, 20% PEG 3350 by hanging drop at 18 °C. The IBV-N29-160 crystals diffract beyond 2.0 Å. Attempts to solve the structure by molecular replacement using the averaged NMR structure of a SARS-CoV nucleocapsid N-terminal domain deposited in the Protein Data Bank (PDB ID: 1SSK) (Huang *et* *al*., 2004) were unsuccessful, even though the two structures turned out to adopt a related fold (Fig. 3. 15). The IBV N29-160 protein is devoid of methionine and cysteine residues. Thus, in order to assist structure determination using the multiwavelength anomalous dispersion (MAD) method, Ile-62, Leu-104, and Val-116 were mutated to methionine (Fig. 3. 13). These hydrophobic amino acid residues have been shown to introduce little perturbation in the native protein structure when substituted by methionine residues (Gassner & Matthews, 1999). In addition, the presumably exposed residue Lys-85 (as suggested by an amino acid sequence alignment with the SARS-CoV N protein) was mutated to Cys in order to introduce a potential binding site for mercury compounds (Fig. 3. 13).

The four residues chosen for mutation are not close to the N- or C-terminus of the IBV N N-terminal domain sequence. The distances between the mutated residues are separated by 12-20 amino acids. Based on the N protein secondary structure prediction, all the four residues chosen for mutation are located in the flexible loop (Fig. 3. 13). Previous research showed that the side chains of isoleucine, leucine and valine are most similar to the side chain of methionine in structure among all the amino acids (Gassner & Matthews, 1999). The hydrophilic lysine 85 was mutated to cysteine because the mutated cysteine should be exposed at the surface of the protein structure accessible for binding by mercury atom. After each mutation, we tested whether the mutated protein could be crystallized or not. Fortunately, the protein with 4 residue mutations could be crystallized in the same condition as the native protein.

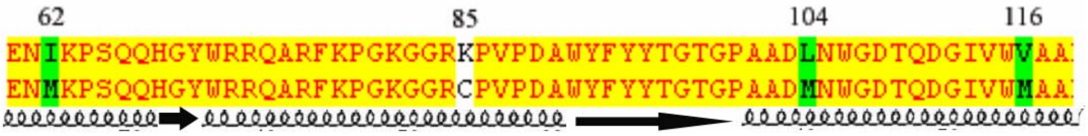

62
85
104
116
ENIKPSQQHGYWRRQARFKPGKGGRKPVPDAWYFYYTGTGPAADLNWGDTQDGIVWVAA
ENMKPSQQHGYWRRQARFKPGKGGRCPVPDAWYFYYTGTGPAADMNWGDTQDGIVWMAA

Fig. 3. 13. Alignment of native IBV N with mutated strand.
Ile-62, Leu-104, and Val-116 were mutated to methionine and Lys-85 was mutated to cysteine in order to introduce selenium-methionine substitution site or mercury binding site. The predicted secondary structure elements are labeled below the sequence for the IBV N protein.

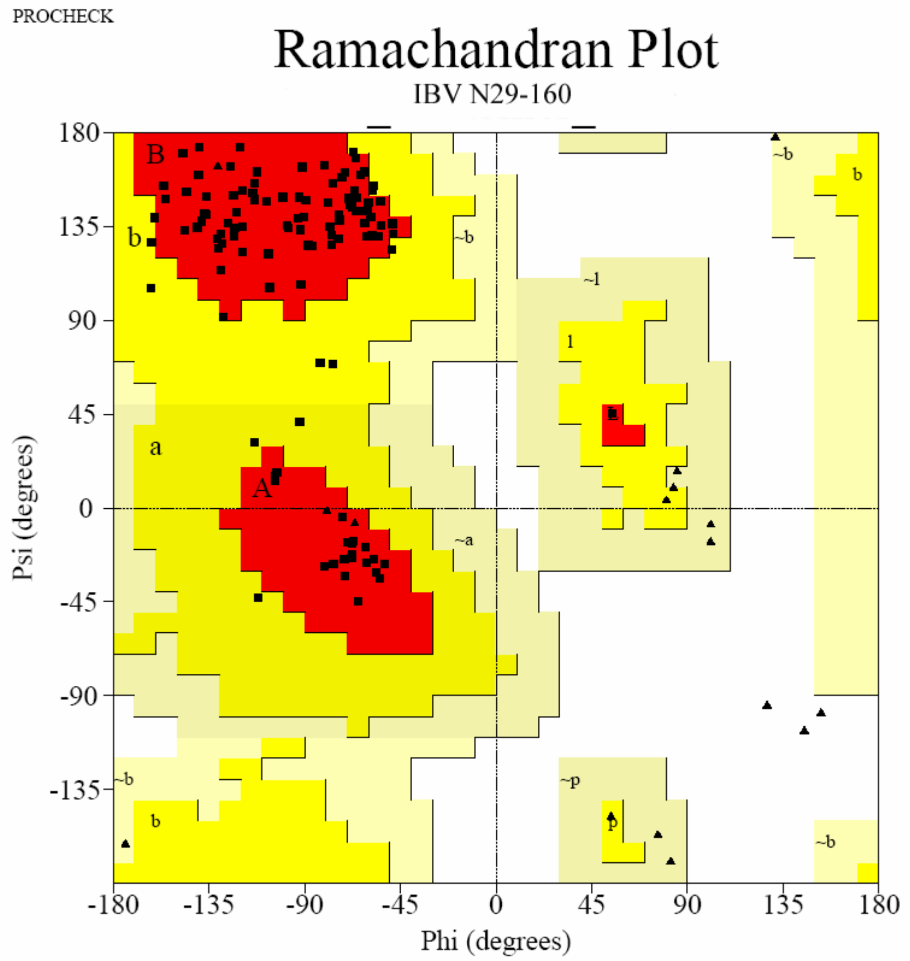
This mutated fragment of IBV-N29-160 was used for structure determination by using the MAD method with crystals containing the selenomethionyl protein. Data collection, phasing, and refinement statistics are summarized in Tables 3 and 4 for the selenomethionine-derivatized crystal (SeMet) and for the native protein crystal.

The native IBV-N29-160 structure was determined and refined by using the MAD data coordinate. Overall, the path of the main chain is unambiguously defined in clear electron density for the two IBV-N29-160 molecules present in the asymmetric unit in each crystal form. A total of 134 protein residues per molecule (two extra residues at the N terminus derive from the cloning procedure) were included in the final models, which have excellent stereochemical parameters as well as 182 and 188 well defined water molecules in native and SeMet asymmetric unit, respectively (Fig. 3. 14 and Table 3). Electron density is absent for the Lys-81 side chain which is exposed to the solvent.

Table 4. The IBV N N-terminal domain data collection and phasing statistics

Data Set	Native IBV-N:29-160	SeMet IBV-N: 29-160 (3 residues mutated to Met)		
X-ray source	Rotating anode	Synchrotron (Photon Factory NW-12)		
Detector	R-axis IV++ image plate	ADSC charged coupled device (CCD)		
		Peak	Inflection	Remote
Wavelength (Å)	1.5418	0.97943	0.97956	0.98729
	a=35.48 Å		*a*=34.77 Å	
Cell parameters	*b*=35.72 Å		*b*=35.37 Å	
	c=56.11 Å		*c*=55.95 Å	
	α=99.05°		α=100.51°	
P1	β=93.93°		β=95.48 °	
	γ=109.53°		γ=110.16°	
Resolution (Å)	20-2.0 (2.1-2.0)		20-1.95 (2.05-1.95)	
Total number of reflections	75,720	76,265	64,999	72,832
Number of unique reflection	19,908	20,083	17,032	19,684
Completeness (%)	92.6 (89.8)	96.6 (95.0)	96.5 (95.2)	95.6 (87.8)
Multiplicity	3.8 (3.7)	3.8 (3.7)	3.8 (3.6)	3.7 (3.5)
R_{sym}[c]	0.061 (0.497)	0.05 (0.118)	0.05 (0.131)	0.06(0.177)
I/σ(I)	7.6 (1.8)	8.6 (3.7)	8.3 (4.4)	9.1 (6.0)
Solvent content (%)	43.3		40.6	
No of Se sites	-		6	
Phasing power[e]	-	0.7/0.6	0.6/0.4	0.2/1.1
f ' / f '' [f]	-	-8.1 / 5.7	-10.5/3.3	-4.3/0.5
Figure of merit [g] 20-2.5 Å	-		0.61/0.793	

[a]The numbers in parentheses refers to the last (highest) resolution shell. [b] For the SeMet crystal, Friedel pairs are treated as different reflections. [c]$R_{sym} = \Sigma_h \Sigma_i |I_{hi} - \langle I_h \rangle| / \Sigma_{h,i} I_{hi}$, where I_{hi} is the *i*th observation of the reflection h, while $\langle I_h \rangle$ is its mean intensity [e] Anomalous phasing power / Dispersive phasing power where anomalous phasing power is $|^{\lambda i}F_h| - |^{\lambda i}F_{-h}|$/anomalous lack-of-closure and dispersive phasing power is $|^{\lambda i}F_h| - |^{\lambda j}F_h|$/dispersive lack-of-closure [f]Values of f' and f'' where estimated from a scan of the absorption edge using program CHOOCH (Evans & Pettifer, 2001). Figure of merit are given before and after real space density modification respectively



Plot statistics

Residues in most favoured regions [A,B,L]	92	88.8%
Residues in additional allowed regions [a,b,l,p]	11	10.2%
Residues in generously allowed regions [~a,~b,~l,~p]	0	0.0%
Residues in disallowed regions	0	0.0%
	----	------
Number of non-glycine and non-proline residues	103	100.0%
Number of end-residues (excl. Gly and Pro)	2	
Number of glycine residues (shown as triangles)	16	
Number of proline residues	13	

Total number of residues	134	

Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms
and R-factor no greater than 20%, a good quality model would be expected
to have over 90% in the most favoured regions.

Fig. 3. 14. Stereochemistry analysis of native IBV N protein N-terminal domain by Procheck

Table 5. The IBV N N-terminal domain refinement statistics

	Native	SeMet
Resolution range (Å)	19.92 – 1.85	20.0 – 1.95
Intensity cutoff (F/σ(F))	none	none
No of reflections: completeness (%)	100.	96.1
Used for refinement	18,921	16,077
Used for Rfree calculation	1,026	881
No of non hydrogen atoms		
Protein	2130	2128
Water molecules	188	182
Rfactor[§] (%)	22.96	22.73
Rfree# (%)	27.03	27.59
Rms deviations from ideality		
Bond lengths (Å)	0.007	0.008
Bond angles (°)	1.05	1.14
Ramanchandran plot		
Residues in most favoured regions (%)	88.8	90.3
Residues in additional allowed regions (%)	10.2	9.2
Residues in generously allowed regions (%)	1.0	0.5
Overall G factor*	0.10	0.04
PDB accession code	2BXX	2BTL

[§] Rfactor =Σ $||F_{obs}| - |F_{calc}|| / \Sigma|F_{obs}|$.

[#] Rfree was calculated with 5% of reflections excluded from the whole refinement procedure.

* G factor is the overall measure of structure quality from PROCHECK (Laskowski *et al.*, 1993).

3.2.4 Overall Structure of the N-terminal domain

The two monomers present in the asymmetric unit can be superimposed with an r.m.s. deviation of 0.5 Å for their main chain atoms. The IBV-N29-160 monomer has approximate overall dimensions of 35 Å x 35 Å x 30 Å and consists of a core formed by a five-stranded antiparallel β-sheet with the topology β_4-β_2-β_3-β_1-β_5 which faces a smaller antiparallel sheet composed of only two strands $\beta_{1'}$-$\beta_{4'}$ which are absent in the SARS-N protein N-terminal domain (Fig. 3. 15a). A long flexible hairpin loop $\beta_{2'}$-$\beta_{3'}$, which is inserted between the β_2 and β_3 strands protrudes largely from the protein core. This extension is mobile as shown by higher than average temperature factors and contains several basic residues which are conserved across various coronavirus N proteins sequences (Fig. 3. 16). The temperature factor is a measure of how much the atom oscillates around its average position specified in the model. From the temperature factors computed from refinement, we learn which atoms in the molecule are most mobile, and gain some insight into the dynamics of the static model. The $\beta_{2'}$-$\beta_{3'}$ hairpin loop, as shown by the high temperature factors, are the most flexible region in protein structure. In addition, the $\beta_{2'}$-$\beta_{3'}$ hairpin loop is made of basic residues, including R76, K78, K81 and R84. The positive charged residues in $\beta_{2'}$-$\beta_{3'}$ hairpin loop may interact with negative charged phosphate groups of RNA. The Extended loops spanning up to 30 residues connect the various secondary structure elements, presumably introducing flexibility to the overall architecture. This potential adaptability to various structural contexts might be important for assembly and disassembly of the nucleocapsid during the virus life cycle. The overall fold is similar to the SARS-N45-181 protein (Fig. 3. 15) with a few structural differences, such as the presence of a short 3_{10} helix connecting strands $\beta_{1'}$ and β_2.

The three dimensional structure of the SARS-CoV N protein's N-terminal domain was solved by nuclear magnetic resonance spectroscopy. The protein consists of a five-stranded β sheet with a folding topology distinct from any other RNA-binding protein. A flexible β hairpin (β2'-β3') extends beyond the core of the protein which consists of five antiparallelel β sheets with the topology β4-β2-β3-β1-β5 (Fig. 3. 15). The residues in the extended β hairpin loop are predominantly basic with 5 of 15 residues being arginines or lysines. The long flexible β hairpin with its positive charged surface may grasp RNA against the β sheet like U1A RNP RNA-binding protein, where a highly positively charged loop between β2 and β3 and face of β sheet is involved in RNA binding (Huang *et al.*, 2004).

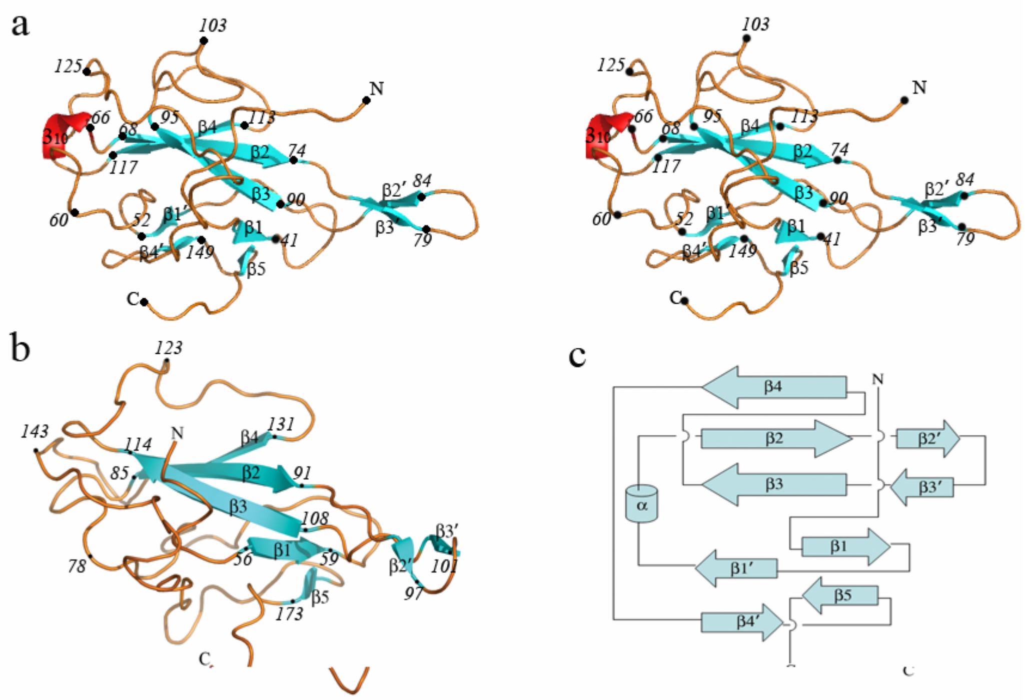


Fig. 3. 15. Overall Fold of the IBV-N N-terminal domain.

(a and b) Comparison of the folds adopted by IBV-N29-160 ([a]; shown as a stereoview, top) and the N-terminal domain of the SARS-CoV nucleocapsid protein (b) (Huang *et al.*, 2004). The two proteins are displayed in the same orientation. Secondary structure elements and some residue numbers are indicated.

(c) Topology diagram of the IBV-N29-160 protein. Its N- and C-terminal ends are labeled.

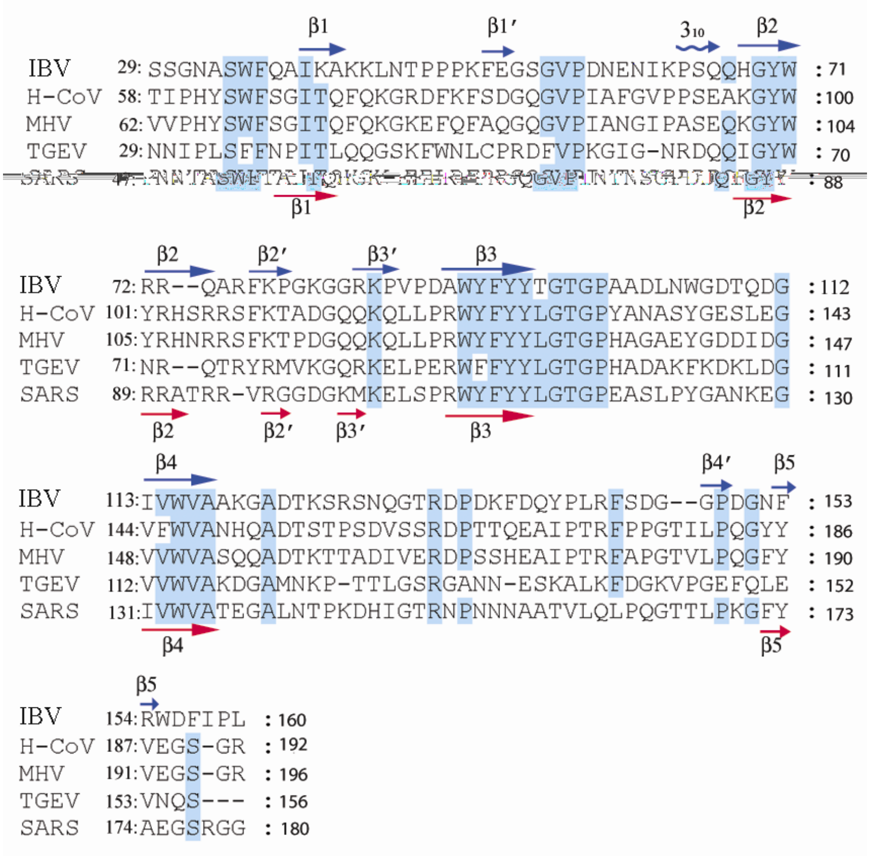

β1
β1′
3₁₀
β2
IBV 29: SSGNASWFQAIKAKKLNTPPPKFEGSGVPDNENIKPSQQHGYW : 71
H-CoV 58: TIPHYSWFSGITQFQKGRDFKFSDGQGVPIAFGVPPSEAKGYW : 100
MHV 62: VVPHYSWFSGITQFQKGKEFQFAQGQGVPIANGIPASEQKGYW : 104
TGEV 29: NNIPLSFFNPITLQQGSKFWNLCPRDFVPKGIG-NRDQQIGYW : 70
SARS 47: : 88
β1
β2
β2
β2′
β3′
β3
IBV 72: RR--QARFKPGKGGRKPVPDAWYFYYTGTGPAADLNWGDTQDG : 112
H-CoV 101: YRHSRRSFKTADGQQKQLLPRWYFYYLGTGPYANASYGESLEG : 143
MHV 105: YRHNRRSFKTPDGQQKQLLPRWYFYYLGTGPHAGAEYGDDIDG : 147
TGEV 71: NR--QTRYRMVKGQRKELPERWFFYYLGTGPHADAKFKDKLDG : 111
SARS 89: RRATRR-VRGGDGKMKELSPRWYFYYLGTGPEASLPYGANKEG : 130
β2
β2′
β3′
β3
β4
β4′
β5
IBV 113: IVWVAAKGADTKSRSNQGTRDPDKFDQYPLRFSDG--GPDGNF : 153
H-CoV 144: VFWVANHQADTSTPSDVSSRDPTTQEAIPTRFPPGTILPQGYY : 186
MHV 148: VVWVASQQADTKTTADIVERDPSSHEAIPTRFAPGTVLPQGFY : 190
TGEV 112: VVWVAKDGAMNKP-TTLGSRGANN-ESKALKFDGKVPGEFQLE : 152
SARS 131: IVWVATEGALNTPKDHIGTRNPNNNAATVLQLPQGTTLPKGFY : 173
β4
β5
β5
IBV 154: RWDFIPL : 160
H-CoV 187: VEGS-GR : 192
MHV 191: VEGS-GR : 196
TGEV 153: VNQS--- : 156
SARS 174: AEGSRGG : 180

Fig. 3. 16. Structure based alignment of coronavirus nucleocapsid amino-acid sequences corresponding to the proteolytically stable N-terminal fragment.
Secondary structure elements are labeled above the sequence for IBV-N29-160 and below for the SARS-CoV N-terminal fragment (Huang *et al.*, 2004). Sequences of IBV (Infectious Bronchitis Virus, strain Beaudette, NP_040838); H-CoV, (Human coronavirus, strain HKU1, YP_173242); MHV (Murine hepatitis virus strain 1, AAA46439); TGEV (porcine transmissible gastroenteritis virus, strain RM4, AAG30228); SARS (SARS CoV, 1SSK_A) were obtained from GenBank. Conserved residues are shown in a cyan background.

Overall, a 3D structural alignment between the SARS-CoV and IBV nucleocapsid N-terminal domains using the program DALI (Holm & Sander, 1993) shows that a total of 124 equivalent C atoms can be superimposed with an r.m.s deviation of 3.0 Å (Fig. 3. 17). The Z score is 10.4 confirming the global similarity of the two folds. The rather large difference between the SARS-CoV and IBV N protein N-terminal domains structures accounts for the failure of molecular replacement procedures to solve the latter structure using the former as a model.

By using the SARS-CoV N protein as a model, we calculated the phases and tried to solve the structure of the N-terminal domain of the IBV N using Molrep from CCP4 (1994). After a few refinement cycles, the rather high Rfactor (0.342) did not improve and the correlation factor are only up to 40%.

The important structural differences we observe between the SARS-CoV and IBV N protein N-terminal domain structures may stem from an inherent mobility of the coronavirus nucleocapsid structure or from a large uncertainty of the atomic positions determined by NMR or both. Unlike protein models defined by crystallography, the models defined by NMR can have several possibilities for one atomic position, which may lead to the failure to solve the IBV nucleocapsid N-terminal domain using the NMR model. A search through the PDB did not return any other protein with a statistically significant Z-score emphasizing the uniqueness of this fold as noted by Huang *et al.* (2004).

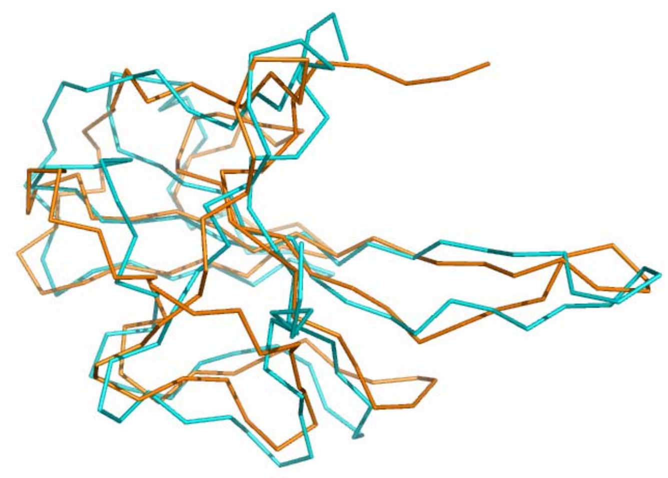


Fig. 3. 17. Superposition of the IBV-N N-terminal (yellow) domain and the SARS-N N-terminal domain (blue)

The IBV N-terminal domain can be superimposed with the SARS N-terminal domain with r.m.s.d 3.0 Å.

3.2.5 Non-crystallography dimerization of the IBV N-terminal domain.

In our crystal structure, the two monomers assemble into a butterfly-shaped dimer related by a 180º rotation, burying in this interaction an accessible surface area of 560 Å^2 (Fig. 3. 18). The transformation is not a pure rotation since a residual translation is needed to bring the two monomers into coincidence.

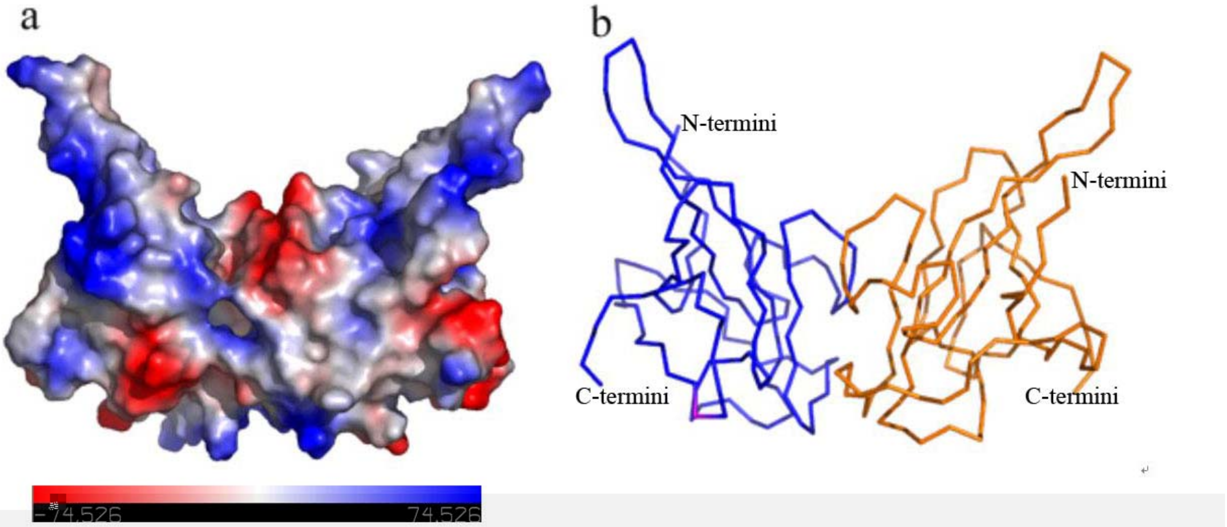


Fig. 3. 18. Dimer of the N-terminal domain at asymmetric unit.

Electrostatic surface of dimer was shown in (a) and two monomers are represented by ribbon in different colors (b).

The relatively small surface area suggests a rather weak binding affinity, an observation in agreement with the fact that, using size exclusion chromatography, the recombinant IBV-N29-160 protein predominantly elutes as a monomer (Fig. 3. 20).

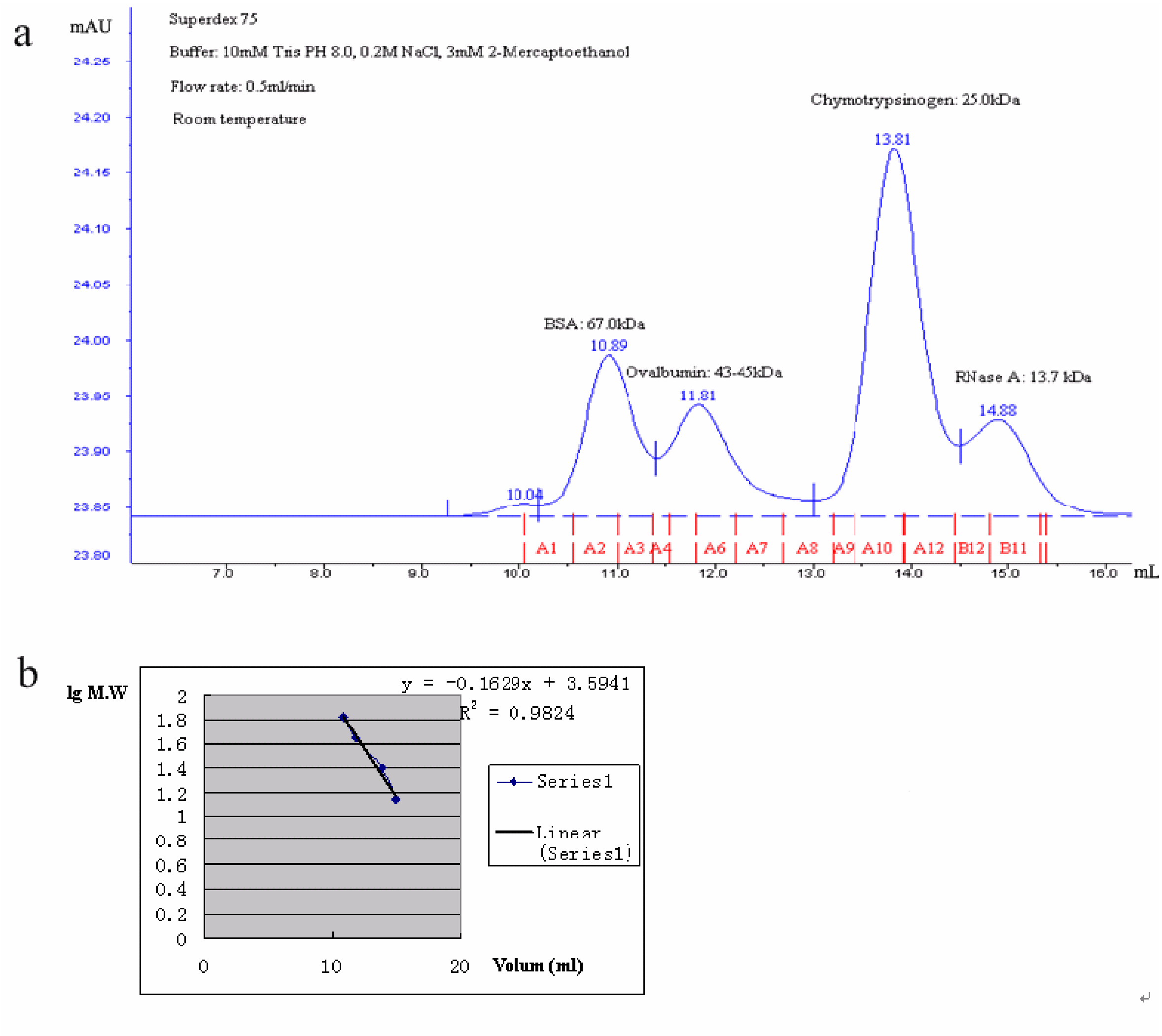


Fig. 3. 19. Standard protein marker for Superdex 200 column (a) and linear standard curve for calculating M.W. (b).

The marker used in (a) were BSA monomer (67 kDa), Ovalbumin (43 kDa), Chymotrypsinogen A (25 kDa) and Ribonuclease A (13.7 kDa).

The purified IBV N29-160 was loaded onto a size exclusion column (Superdex 75) which was pre-calibrated with gel filtration markers (Fig. 3. 19). The linear standard curve was used to estimate the apparent M.W of IBV N29-160 in accordance with the protein elution volume from size exclusion chromatography (Fig. 3. 20).

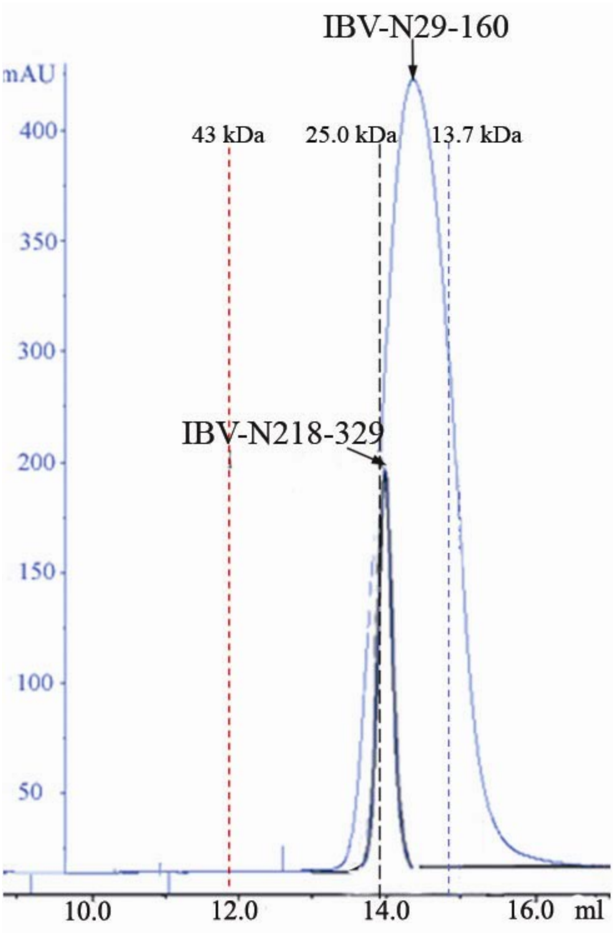


Fig. 3. 20. Size exclusion chromatography elution profiles of the IBV-N N-terminal domain (IBV N29-160) and the C-terminal domain (IBV-N 218-329).

The vertical axis shows absorbance at 280nm. The horizontal axis indicates the elution volume in ml. Three thin vertical lines indicate the positions of molecular weight of protein standards. (from left to right: Ovalbumin: 43 kDa, Chymotrypsinogen A: 25.0 kDa and Ribonuclease A: 13.7 kDa). The large difference in absorbance between IBV-N29-160 and IBV-N218-329 stems from the different individual molar absorbance coefficients at 280 nm of IBV-N29-160 (40540 $M^{-1}cm^{-1}$) and IBV-N218-329 (4080 $M^{-1}cm^{-1}$) respectively.

This is also consistent with our findings of a different dimeric interface adopted by the same recombinant IBV-N29-160 protein in a non related crystal form (with space group C2) that diffracts to 3.0 Å. Two molecules form a dimer in the asymmetric unit. The structure of this crystal form was refined to an Rfactor 0.26 and an Rfree 0.328.

3.2.6 Identification of a proteolytically stable C-terminal domain

The C-terminal domain that ranges from residue 218 to 329 was named IBV N218-329. The C-terminal domain was derived from slow degradation by incubating the protein at 18 °C for 1 month. The sequence of this domain was determined by mass spectrometry and N-terminal protein sequencing. Mass spectrometry showed that this proteolytically stable fragment which can be crystallized has a molecular mass: 12.668 kDa. The sequencing result identified the N-terminal residues: Lys-Ala-Asp-Glu-Met-Ala.

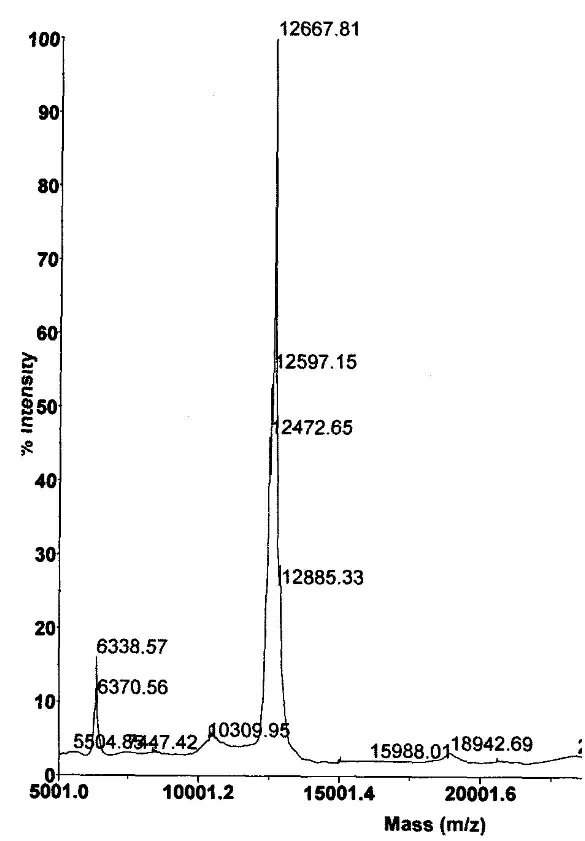


Fig. 3. 21. Mass spectrometry of the IBV N C-terminal domain.
The proteolytically stable C-terminal domain has a molecular mass: 12.668 kDa.

3.2.7 Structure determination of the C-terminal domain of the IBV N and quality of the model.

The IBV N protein C-terminal domain (IBV-N218-329) was expressed in *E.coli* BL21(DE3). The purified IBV-N218-329 can be crystallized in hanging drops in the condition: 4.3 M NaCl, 0.1 M Tris-Cl, pH 8.5. Lens-shaped single crystals diffract to about 2.6 Å (Fig. 3. 8). The structure of IBV-N218-329 was solved by molecular replacement using the IBV (Gray strain) C-terminal domain structure (PDB accession: 2GE7). Data collection, phasing, and refinement statistics are summarized in Table 4 and 5.

Radiation damage is a serious problem at synchrotron. It is best to get a complete dataset first, and then collect additional images to increase the multiplicity. The option STRATEGY in MOSFLM was used to determine the smallest angle for data collection that will give the highest possible completeness. In case of the IBV N protein C-terminal domain crystal data collection, data collection by 70º (1º oscillation) gave more than 90% completeness with an overall redundancy 1.65. The data collection range of 70º and start/end angles were determined by STRATEGY based on the crystal symmetry and orientation

Table 6. The C-terminal domain crystallization and data collection

Crystallization condition	4.3 M NaCl, 0.1 M Tris-Cl pH 8.5
X-ray source	ID14-4,ESRF (Grenoble)
Wavelength	0.97626 Å (70º, 1º oscillation)
Cell parameters	$P4_3$
	a= *b*=61.59 Å *c*=91.88 Å
	α=β=γ=90.00°
Resolution (Å)	20-2.6 (2.7-2.6)
Total number of reflections	31,078
No. of unique reflection	18,835
Number of molecules in ASU	2
Completeness (%)	90.6 (90.5)
Redundancy	1.65 (1.66)
$\boldsymbol{R_{sym}}$[b]	0.079 (0.417)
I/σ(I)	13.59 (3.1)

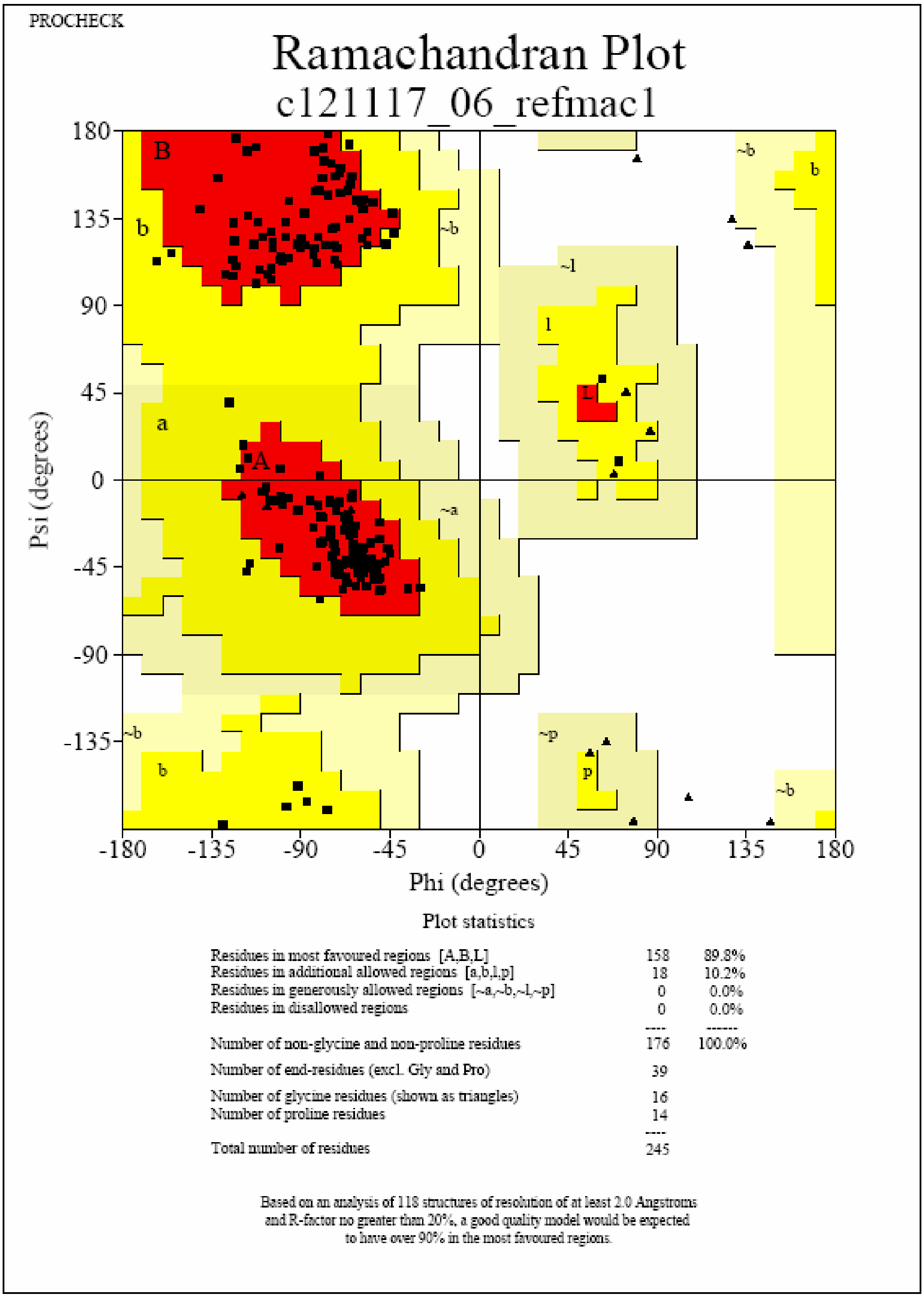
[a]The numbers in parentheses refers to the last (highest) resolution shell.

[b]$R_{sym} = \Sigma_h \Sigma_i |I_{hi} - \langle I_h \rangle| / \Sigma_{h,i} I_{hi}$, where I_{hi} is the *i*th observation of the reflection h, while $\langle I_h \rangle$ is its mean intensity

In 4_3 and 4_1screw axis, symmetry axis is present along a certain direction in real space along c. Because the symmetry axis contains a translational component, there will be a repeat within the projected density between 0 and 1. It will be quadruple repeat for a 4_1 and 4_3. For the 4_1 and 4_3 screw axis, only the reflections of 00l: l=4n will be non-zero. For the 4_2 screw axis, only the reflections of 001: l=2n will be non-zero.

If the axis is a pure rotational operator (for example P4), the reflections on the corresponding reciprocal space axis c^*, or the 00l will have an arbitrary, nonrepetitive distribution along its entire length from 0 to 1.

The space group of the IBV N protein C-terminal domain was identified as $P4_3$ instead of $P4_1$ with the translation function (program Molrep from CCP4i) by using IBV (Gray strain) N protein C-terminal domain coordinate as a model. Translation functions show that space group $P4_3$ gives higher correlation coefficiences between F_{obs} and F_{calc} than the space group $P4_{1.}$



Plot statistics

```
Residues in most favoured regions  [A,B,L]                  158     89.8%
Residues in additional allowed regions  [a,b,l,p]            18     10.2%
Residues in generously allowed regions  [~a,~b,~l,~p]         0      0.0%
Residues in disallowed regions                                0      0.0%
                                                           ----    ------
Number of non-glycine and non-proline residues              176    100.0%

Number of end-residues (excl. Gly and Pro)                   39

Number of glycine residues (shown as triangles)              16
Number of proline residues                                   14
                                                           ----
Total number of residues                                    245
```

Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms
and R-factor no greater than 20%, a good quality model would be expected
to have over 90% in the most favoured regions.

Fig. 3. 22. Stereochemistry analysis of the C-terminal domain structure by procheck

Table 7. Refinement statistics of the C-terminal domain

Resolution Range (Å)	20-2.6 (2.7-2.6)
Number of reflections	8,941
Rfactor[§]	0.204
Rfree[#]	0.256
Mean Bond length deviation	0.009 Å
Mean Bond angle deviation	1.176°
Ramachandran statistics	
Residues in most favored regions (%)	89.8
Residues in additional allowed regions (%)	10.2
Residues in generously allowed regions (%)	-
Residues in disallowed (poor density) regions (%)	-
Overall G factor*	0.05

Rfactor = Σ ||F_{obs}| - |F_{calc}|| / Σ |F_{obs}|.

[#] Rfree[#] was calculated with 5% of reflections excluded from the whole refinement procedure.

*G factor is the overall measure of structure quality from PROCHECK (Laskowski *et al.*, 1993).

3.2.8 Overall structure of the C-terminal domain of the IBV N

The C-terminal domain exists as tightly intertwined two-fold symmetric dimer (Fig. 3. 23) with two β strands and one helix from one monomer making extensive contacts with the other and burying a total surface of approximately 5,000 Å^2 in their interaction. This is consistent with results reported by Surjit *et al.* (2004) who show that SARS-CoV nucleocapsid protein exists as a dimer through the C-terminal dimerization domain by using the yeast-two hybrid system and points to conserved assembly properties between the SARS-CoV and IBV, in spite of significant amino-acid differences between their two nucleocapsid proteins.

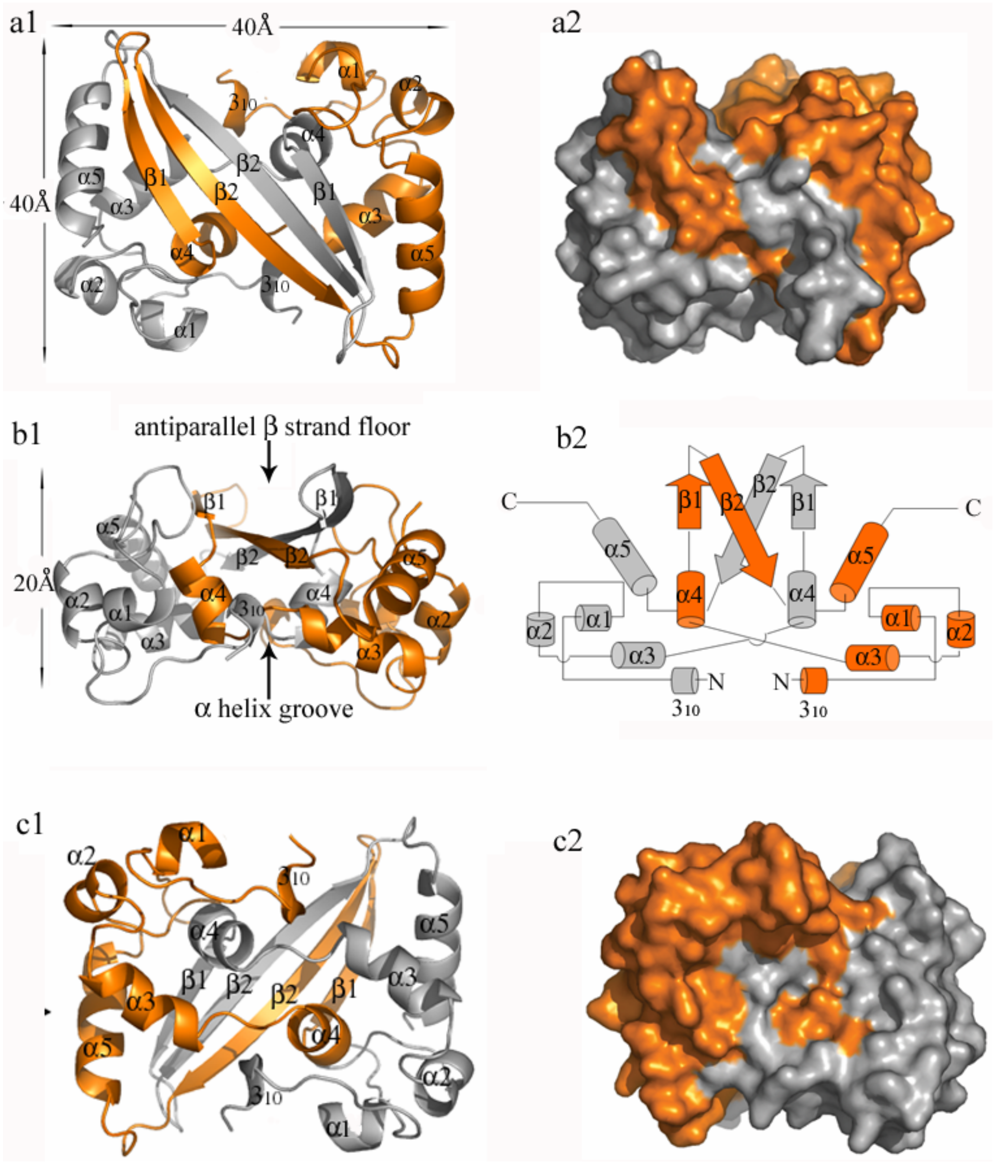


Fig. 3. 23. Secondary structural elements and solvent accessible surface of the C-terminal domain.

Two monomers are colored in white and gold, respectively. Tightly intertwined dimer-association is shown by secondary structure or solvent accessible surface.

- **(a)** Four antiparallel β strands (β1 and β2 from two monomers) form an antiparallel β strand floor flanked by two α helix α 5. Solvent accessible surface of antiparallel β strand floor is shown in a2.
- **(b)** Top view of antiparallel β strand floor and α helix groove (as shown by arrow). Topology of dimer C-terminal domain is shown on b2.
- **(c)** On the opposite side of antiparallel β strand floor, α 1, α 2, α 3, α 4 from two monomers form α helix groove. Solvent accessible surface of α helix groove is shown in c2.

The elements of secondary structure of the C-terminal domain are shown in Fig. 3. 24. The C-terminal domain has a rectangular shape delimited by edges formed by the C-terminal α helix α5 with an approximate overall dimension of 40Å×40Å×20Å (Fig. 3. 23).

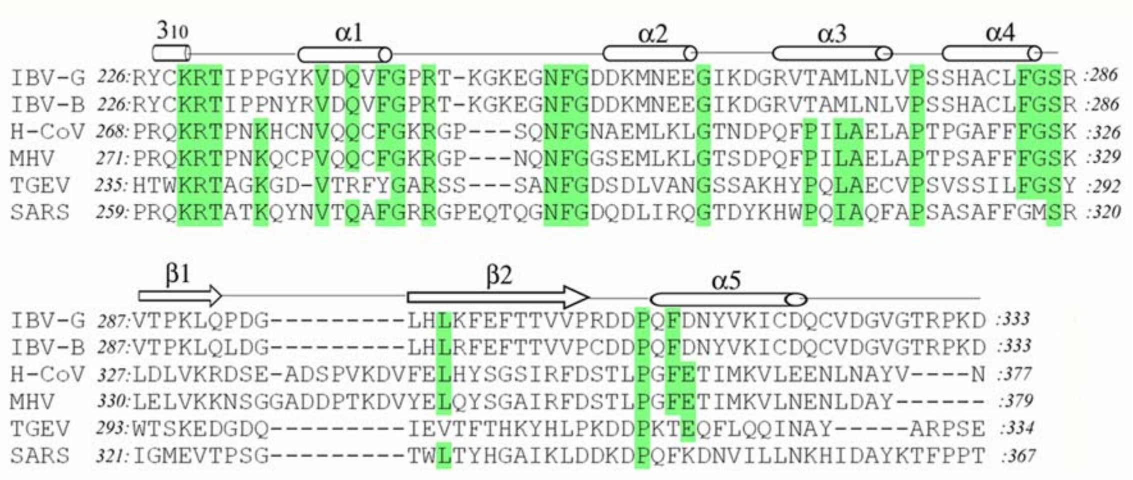

3₁₀
α1
α2
α3
α4
IBV-G 226:RYCKRTIPPGYKVDQVFGPRT-KGKEGNFGDDKMNEEGIKDGRVTAMLNLVPSSHACLFGSR :286
IBV-B 226:RYCKRTIPPNYRVDQVFGPRT-KGKEGNFGDDKMNEEGIKDGRVTAMLNLVPSSHACLFGSR :286
H-CoV 268:PRQKRTPNKHCNVQQCFGKRGP---SQNFGNAEMLKLGTNDPQFPILAELAPTPGAFFFGSK :326
MHV 271:PRQKRTPNKQCPVQQCFGKRGP---NQNFGGSEMLKLGTSDPQFPILAELAPTPSAFFFGSK :329
TGEV 235:HTWKRTAGKGD-VTRFYGARSS---SANFGDSDLVANGSSAKHYPQLAECVPSVSSILFGSY :292
SARS 259:PRQKRTATKQYNVTQAFGRRGPEQTQGNFGDQDLIRQGTDYKHWPQIAQFAPSASAFFGMSR :320
β1
β2
α5
IBV-G 287:VTPKLQPDG---------LHLKFEFTTVVPRDDPQFDNYVKICDQCVDGVGTRPKD :333
IBV-B 287:VTPKLQLDG---------LHLRFEFTTVVPCDDPQFDNYVKICDQCVDGVGTRPKD :333
H-CoV 327:LDLVKRDSE-ADSPVKDVFELHYSGSIRFDSTLPGFETIMKVLEENLNAYV----N :377
MHV 330:LELVKKNSGGADDPTKDVYELQYSGAIRFDSTLPGFETIMKVLNENLDAY------ :379
TGEV 293:WTSKEDGDQ---------IEVTFTHKYHLPKDDPKTEQFLQQINAY-----ARPSE :334
SARS 321:IGMEVTPSG---------TWLTYHGAIKLDDKDPQFKDNVILLNKHIDAYKTFPPT :367

Fig. 3. 24. Structure-based alignment of coronavirus nucleocapsid amino acid sequences corresponding to the C-terminal dimerization domain.

Secondary structure elements are labeled above the sequence for the IBV C-terminal dimerization domain. Sequences for the IBV N proteins were obtained from Swiss-Prot (IBV-G [Gray strain], P32923; IBV-B [Beaudette strain], P69596). Sequences for human coronavirus (H-CoV; strain HKU1, YP_173242); MHV (strain 1, AAA46439); SARS (SARS-CoV, NCAP_CVHSA) and porcine transmissible gastroenteritis virus (TGEV; strain RM4, AAG30228) were obtained from GenBank. Conserved residues are shaded in green.

The dimer structure of the C-terminal domain features two concave floors: an antiparallel β-strand floor (Fig. 3. 23a) and an α-helix groove (Fig. 3. 23c) which are mainly composed of β strands and α helices, respectively. Recent biochemical and mass spectrometric studies on the IBV N protein (Beaudette strain) have suggested the possibility of a disulfide bridge in the C-terminal domain. However, no intermolecular disulfide bond is seen in the dimer structure. The present structure of the C-terminal domain is consistent with the previous size exclusion chromatography observation that the C-terminal domain is a dimer in solution (Fig. 3. 20). In addition, crosslinking experiments show that C-terminal domain (IBV N218-329) forms dimers, trimers, tetramer, and higher oligomers for concentrations of crosslinking agent higher than 1 mM with a concomitant decrease in monomer species, thus confirming the contribution of the C-terminal domain to the IBV N protein multimerization (Fig. 3. 25).

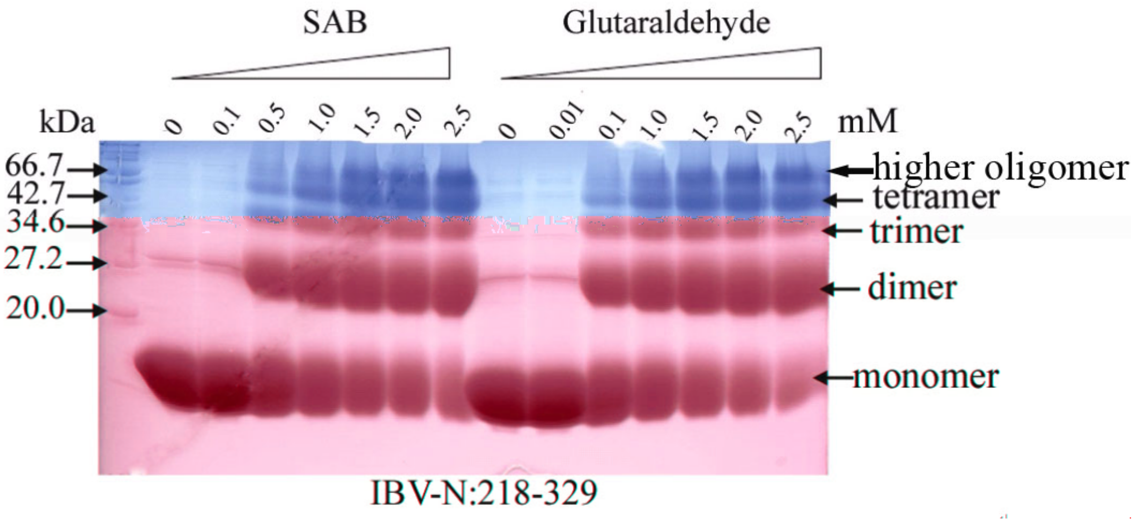


Fig. 3. 25. Crosslinking of the C-terminal domain
Purified IBV N218-329 was incubated with increasing amount of either glutaraldehyde or SAB for 2 hours at 20 °C. Multimerization status of N proteins was checked by SDS-PAGE gel stained by coomassie blue.

In the dimer structure, α 4-β1-β2 makes extensive contact with the other monomer (Fig. 3. 26). Analysis of interchain contacts in the dimer shows that two monomers are associated mostly by hydrogen bonds. 279H, 283Y in α4, 285S, 286R 287V in β1 , 296L, 298L, 302F,306V in β2 play important roles in inter-chain interaction by forming hydrogen bonds with the other monomer. The dimer structure of the C-terminal domain provides information for drug design. The residues on α4−β1-β2 which are important for dimer formation can be targeted, especially the 279H, 283Y of α4 that anchors one monomer to the other (Fig. 3. 26b).

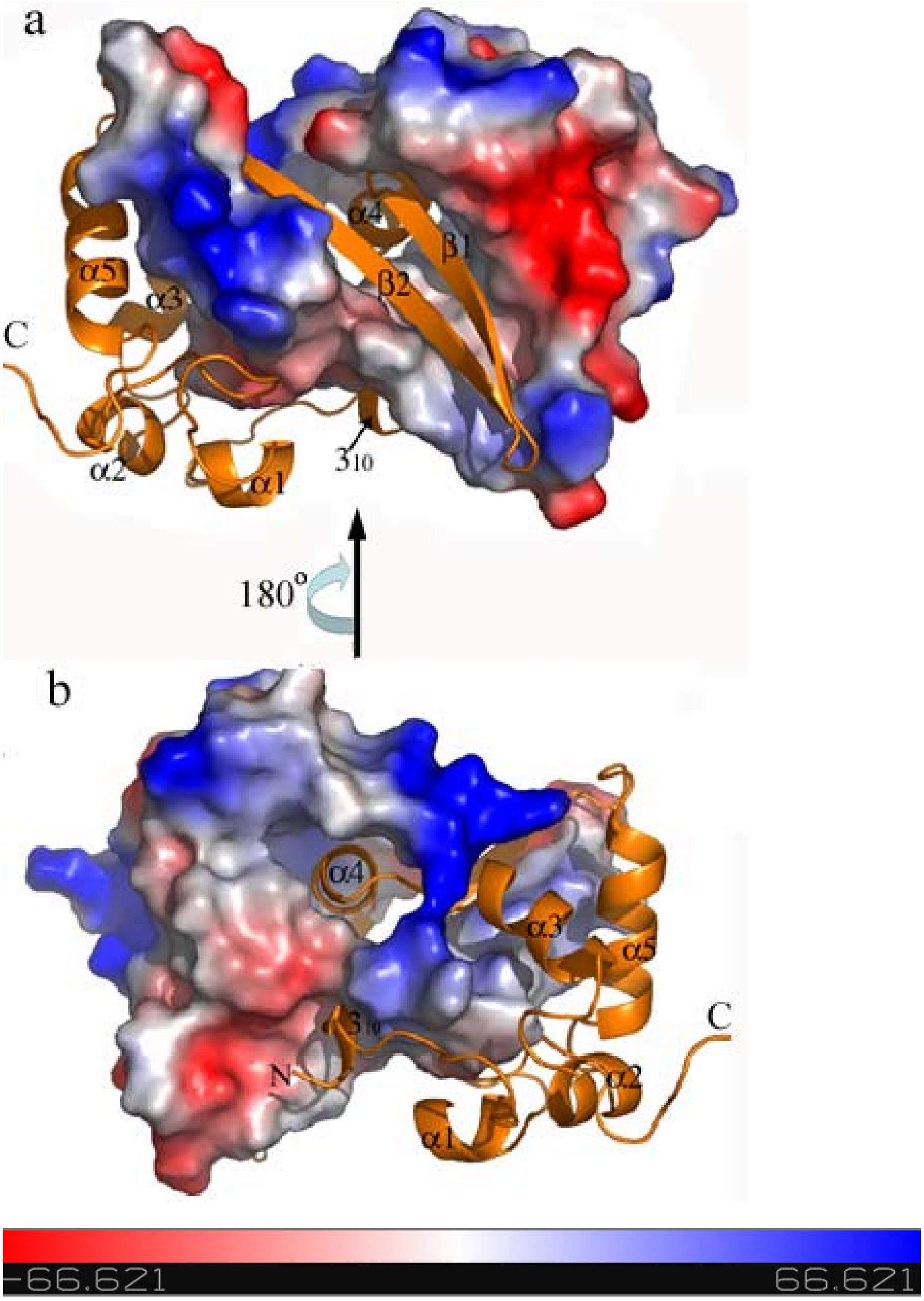


Fig. 3. 26. The important role of α 4-β1-β2 in the C-terminal domain dimer formation.

In the dimer structure, one monomer is represented as a cartoon and the other one is represented by its electrostatic surface. **(a)** β1-β2 makes contact with the other monomer by inserting into the groove formed by β2, α3 and α5. **(b)** α4 is anchored in a hole formed by the other monomer helices 3_{10}-α1-α2-α3.

3.2.9 Super helical packing of the C-terminal domain of the IBV N

In the tetragonal crystal form, two monomers form a tightly associated dimer in the asymmetric unit. The C-terminal domain dimers related by 4_3 screw axis display two kinds of inter-dimer contacts in the crystal. In interface I, 230R of one dimer makes contact with 263I, 264K and 265D of another dimer by hydrogen bonds and two dimers are related by 4_3 screw axis (Fig. 3. 27). In interface II, 313F, 317V, 291L and 293L of one dimer make contact with the other dimer by hydrophobic interactions (Fig. 3. 27). No interdimer disulfide bond was found between 308C and other dimer 308C (Fig. 3. 27). Two dimers connected by interface II are also related by 4_3 screw axis.

These two interfaces create two kinds of super helical structures with different organizations (Fig. 3. 28). The helical structure created by interface I is colored in blue (Fig. 3. 28a). Four dimers (1～4) related by 4_3 screw axis through interface I have dimensions of 54Å×54Å square (see top view, Fig. 3. 28b and c). Through interface II interaction, 4 dimers (1～4) colored in red form a 63Å×63Å square (see top view, Fig. 3. 28b). Two super helical structures created by two type interactions are not separated but intertwined. The dimer 2 in Fig. 3. 28c is shared by both 63Å×63Å and 54Å×54Å super helical structures.

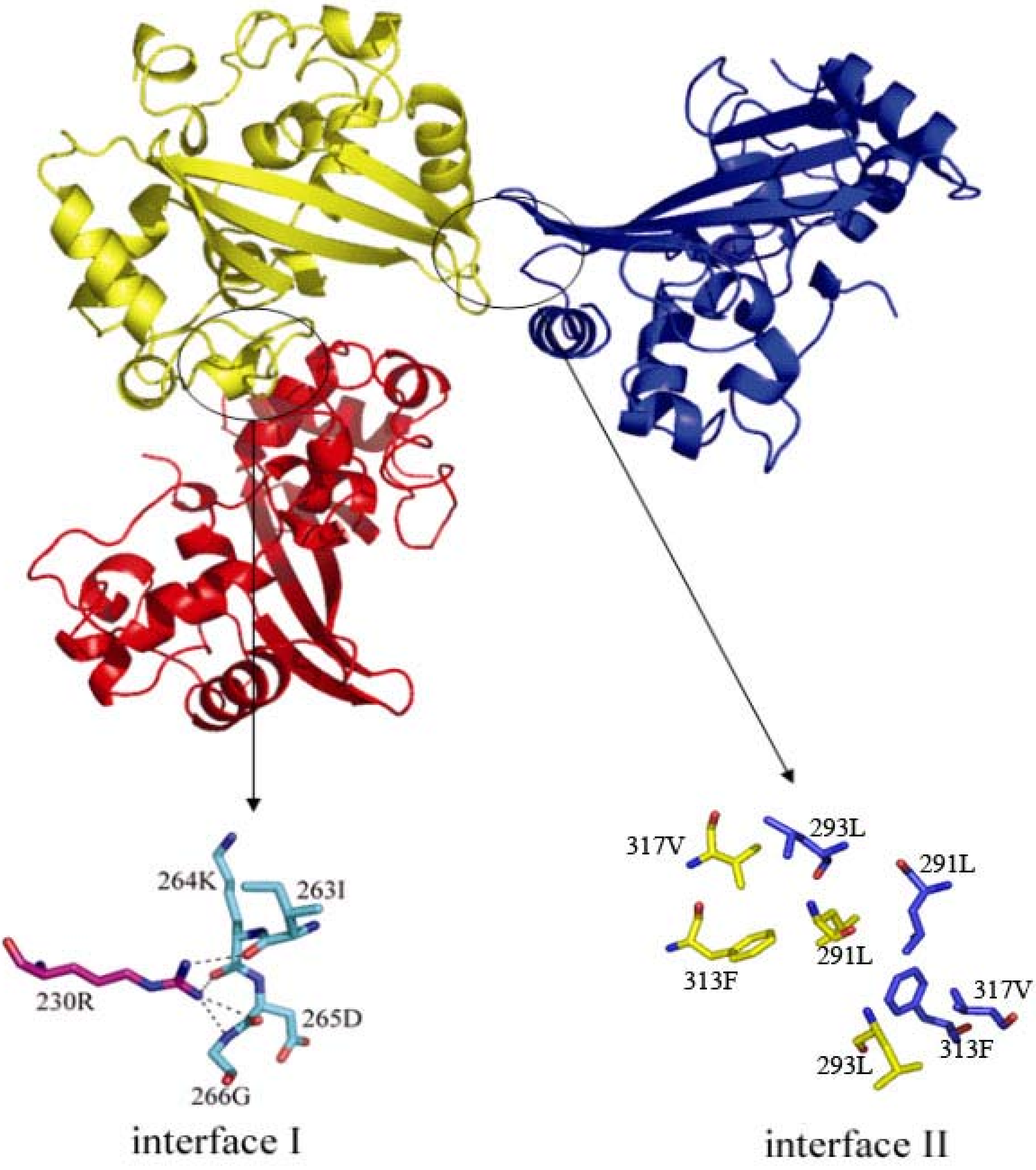


Fig. 3. 27. Interface I and interface II in the C-terminal domain P4$_3$ crystal packing
In the tetragonal crystal form, every dimer is related to other dimers by two kinds of interface: I and II. The residues involved in interface I and II interaction are labeled and indicated by sticks.

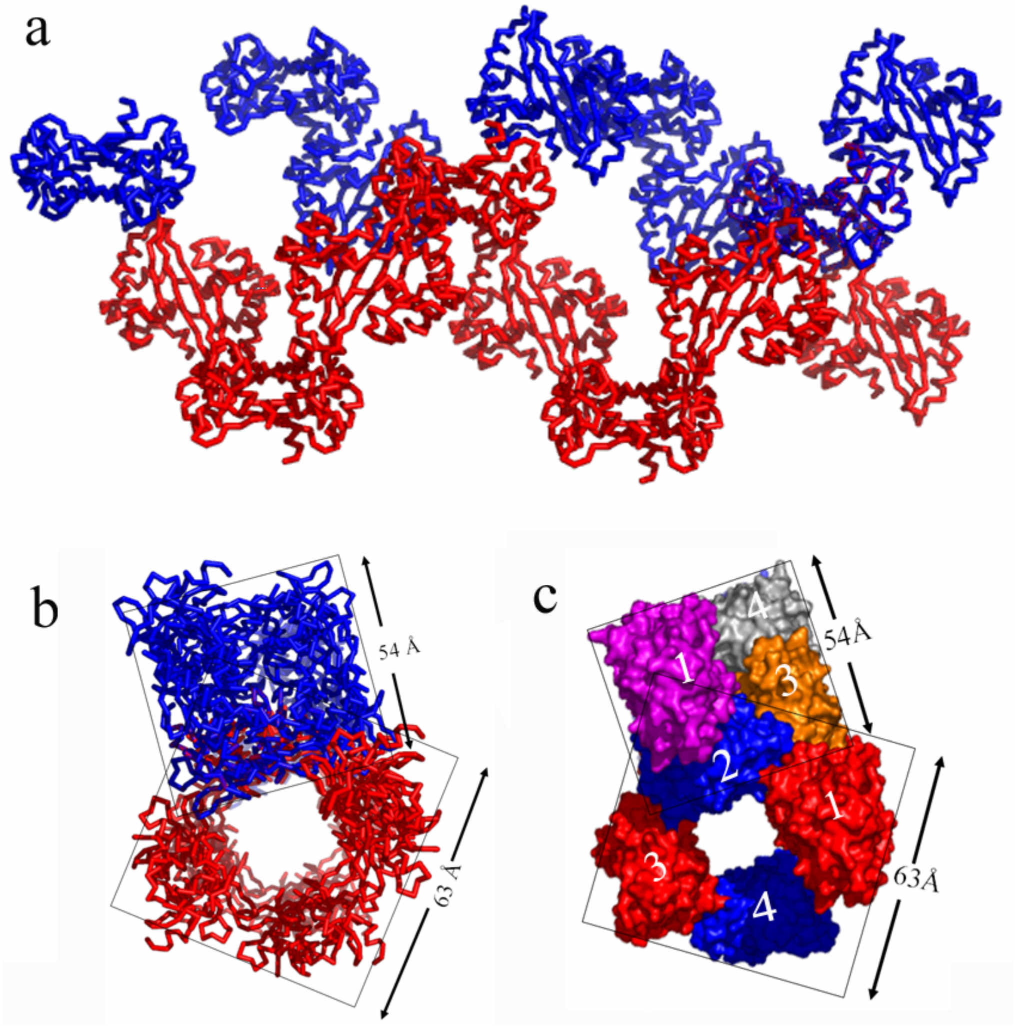


Fig. 3. 28. The C-terminal domain packing in the tetragonal crystal forms two different interfaces.

The C-terminal domains are represented by ribbon in **(a)** and **(b)**. Two super helical structures colored in blue and red are created by interface I and II interaction, respectively in **(a)** and **(b)**. Two super helical structures (blue and red) intertwine on side view **(a)** and form 54Å×54Å square and 63Å×63Å square which are composed of four dimers on top view [**(b)** and **(c)**]. Four dimers numbered from 1 to 4 in 54Å×54Å and 63Å×63Å super helical structures are colored differently in **(c)**. Dimer 1～4 are related by 4_3 screw axis. Two helical structures intertwine by sharing dimer No.2 **(c)**.

The formation of the coronavirus nucleocapsid involves self-association of the N protein and interaction with RNA resulting in a structure which is RNase resistant. Our analysis of the C-terminal domain crystal structure reveals a tight dimer suggesting that the full length protein functions as at least a dimer or a high multimer such as tetramer or hexamer. The dimerization of the C-terminal domain can provide a structural scaffold during nucleocapsid formation. Zhou *et al* (2000) demonstrated that the IBV N protein interacts specifically with RNA sequences located at the 3' noncoding region of the viral genome. Both N- and C-terminal domains of the IBV N protein bind to 3' noncoding region of the viral genome (Zhou & Collisson, 2000). Thus, the C-terminal domain of the IBV N protein may serve a dual purpose of mediating the self-association during nucleocapsid formation and providing a complementary surface for the interaction with viral RNA as well as the N-terminal domain of the IBV N protein.

The relative orientation of the N-terminal domain with respect to the C-terminal domain remains unknown because these two domains are connected by a 58-residues protease-sensitive loop, rich in serine, glycine and alanine residues, which is presumably mobile. The RNA binding regions of these two domains could face each other, engulfing the RNA between them, thus conferring resistance to RNases.

In the tetragonal crystal form, the C-terminal domain can self-associate to form two helical structures by two kinds of inter-dimer interactions with buried interfaces of more than 1,000 Å^2. Propagation of any single type of interaction would lead to a rigid helical nucleocapsid. The C-terminal domain can serve as a building block of a helical nucleocapsid through the C-terminal domain strong dimerization capacity and inter-dimer stacking.

The C-terminal domain sequence of the IBV N protein (Beaudette strain) shares 94.9% identity with that of the IBV N protein (Gray strain). A superposition of these two dimers shows that a total of 198 equivalent C_α atoms can be superimposed with rms deviation of 0.5 Å.

Interestingly, electron microscope studies on MHV showed that the virus core shell contains a filamentous (probably helical) structure consisting of the viral RNA and N proteins (Risco *et al.*, 1996). The super helical structures formed by the IBV N protein C-terminal domain in the tetragonal crystal form may be relevant to the organization of the N protein in nucleocapsid formation like porcine reproductive and respiratory syndrome virus (PRRSV), a member of the arterviruses. Coronaviruses are related to arteriviruses by their similar genome organizations and viral replication mechanism (Doan & Dokland, 2003). *Coronaviridae* and *Arteriviridae* are grouped in the same order *Nidovirales*. The PRRSV N protein forms helical structure by inter-dimer packing in 3-fold screw axis and also have a filamentous nucleocapsid structure revealed by electron microscopy (Doan & Dokland, 2003). However, the PRRSV dimer packing is not inconsistent with the formation of the viral nucleocapsid. More studies on nucleocapsid formation and nucleocapsid protein self-association are required to get insight into the formation of the coronavirus nucleocapsid.

3.3 RNA binding of the IBV N protein

Coronaviruses have been classified into four groups, with SARS-CoV being the founding member of an independent group. The N protein sequences are more similar within each group (～40% identity) than across the groups (20 to 30%). Although the IBV N protein interacts with the 3' non-coding region of its genomic RNA, computer analyses of the N protein could not identify any known motif which is specific for RNA binding (Zhou ＆ Collisson, 2000). The amino terminus of the IBV N protein is highly basic while the carboxyl terminus is relatively acidic. Gel shift assays showed that the amino and carboxyl regions of the IBV N protein, but not the middle region, can interact with the RNA from the 3' non-coding region of the IBV genome (Zhou ＆ Collisson, 2000).

A stem-loop structure at the MHV 3' non-coding region is identified to be essential for RNA replication. This result is in agreement with studies by Huang *et al* (2003) who used N.M.R to demonstrate that SARS-CoV N45-181 could bind a 32-mer oligoribonucleotide located at the 3' end of the SARS-CoV genome (5-'CGAGGCCACGCGGAGUACGAUCGAGGGUACAG-3'). Interestingly, this oligoribonucleotide has a highly conserved sequence across various coronaviruses including IBV, and adopts a unique tertiary structure (Robertson *et al*., 2005).It is evident that the 3' noncoding region plays very important roles in viral RNA replication and transcription (Zhou ＆ Collisson, 2000).

In order to package the viral genome of 27.6 kb, the IBV-N protein must provide extended surfaces to bind the viral RNA genome both specifically and non-specifically (without a requirement for a special base sequence). The N- and C-terminal regions of IBV-N encompassing residues 1 to 171 and 268 to 407 respectively interact with non-coding regions of the viral genomic RNA located at its 3' end (Zhou *et al.*, 2000). As the fragment 1-91 does not bind RNA, residues between 91 and 171 were proposed to either make direct contacts with RNA or be necessary for the integrity of the protein structure (Zhou *et al.*, 2000). Since the segment 92-95 includes strictly conserved hydrophobic residues which are buried in the protein core in our structure, we propose that the fragment 1-91 studied by Zhou *et al* (2000) was probably poorly folded and thus non active. We tested nucleic acid binding by IBVN29-160 and found that the recombinant fragment was able to bind an oligoribonucleotide from the 3' end of the viral genome [nucleotides 26539 to 27609 from IBV genome (Accession code: NC_001451)] (Fig.3.28).

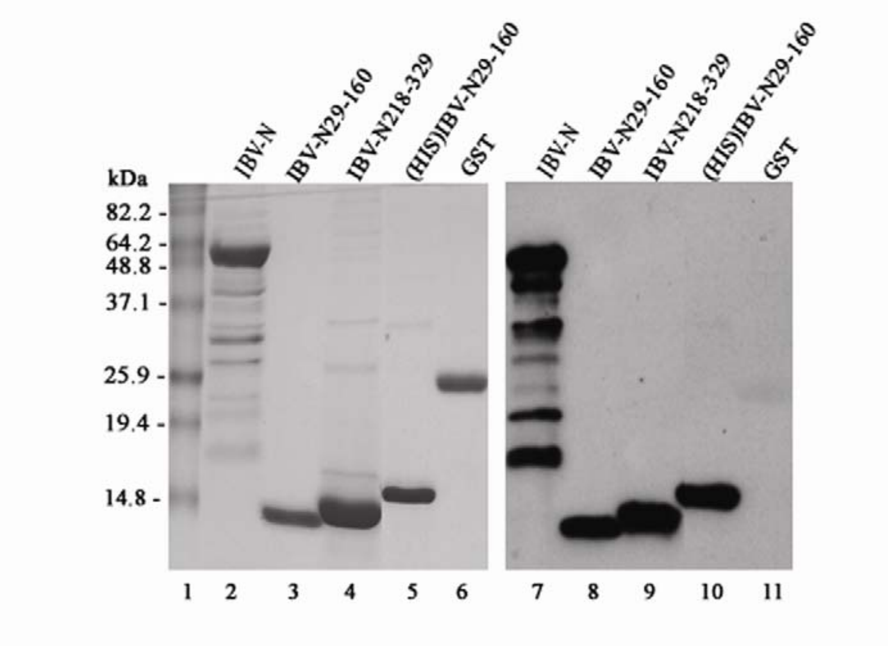


Fig. 3. 29. Analysis of the RNA-binding activity of the full-length IBV-N protein, the N-terminal domain (IBV-N29-160) and the C-terminal domain (IBV-N218-329) fragments.

The purified IBV-N (lanes 2 and 7), IBV-N29-160 (lanes 3 and 8), IBV-N218-329 (lanes 4 and 9), His-tagged IBV-N29-160 (lanes 5 and 10) and GST (negative control, lanes 6 and 11) were separated on a 15% SDS-PAGE gel. The proteins were either visualized by coomassie brilliant blue staining (lanes 1-6), or transferred to Hybond C-extra membrane (Amersham) and detected by Northwestern blot with a Digoxin-labeled RNA probe corresponding to the IBV genome sequence from nucleotides 26539 to 27608 (lanes 7-11). Molecular masses of standard proteins are indicated.

A surface representation of electrostatic charges of the IBV-N29-160 protein shown in Fig. 3. 30 reveals a striking segregation in the charges distribution on the protein surface. The $\beta_{2'}$-$\beta_{3'}$ hairpin forms a basic patch at the thumb whereas the base is acidic (Fig. 3. 30). These two charged patches are separated by a neutral and rather hydrophobic platform contributed by residues projecting from strands β_4-β_2-β_3 that form a palm like structure. An alignment of nucleocapsid proteins amino-acids sequences from various coronaviruses highlights the conservation of several residues exposed at the protein surface suggesting that some might play a role in nucleic acid recognition (Fig. 3. 30). The topology of the protein and its charge distribution suggest a mode of RNA binding in which its phosphate groups would project towards the basic $\beta_{2'}$-$\beta_{3'}$ hairpin possibly making electrostatic interactions with the conserved positively charged Arg-76 and Lys-78 residues, while the sugar and base moieties would contact the hydrophobic platform. In this model, the exposed hydrophobic residues Tyr-92 and Tyr-94 (strand β3) could form stacking interactions with the bases as was observed for instance in complexes between the vaccinia virus protein VP39 and mRNA (Hu *et al*., 1999) or between the matrix protein VP40 from Ebola virus and a triribonucleotide (Gomis-Ruth *et al*., 2003). As suggested by Huang *et al.* (2004), additional favorable interactions might be formed upon closure of the flexible $\beta_{2'}$-$\beta_{3'}$ hairpin onto the incoming RNA ligand.

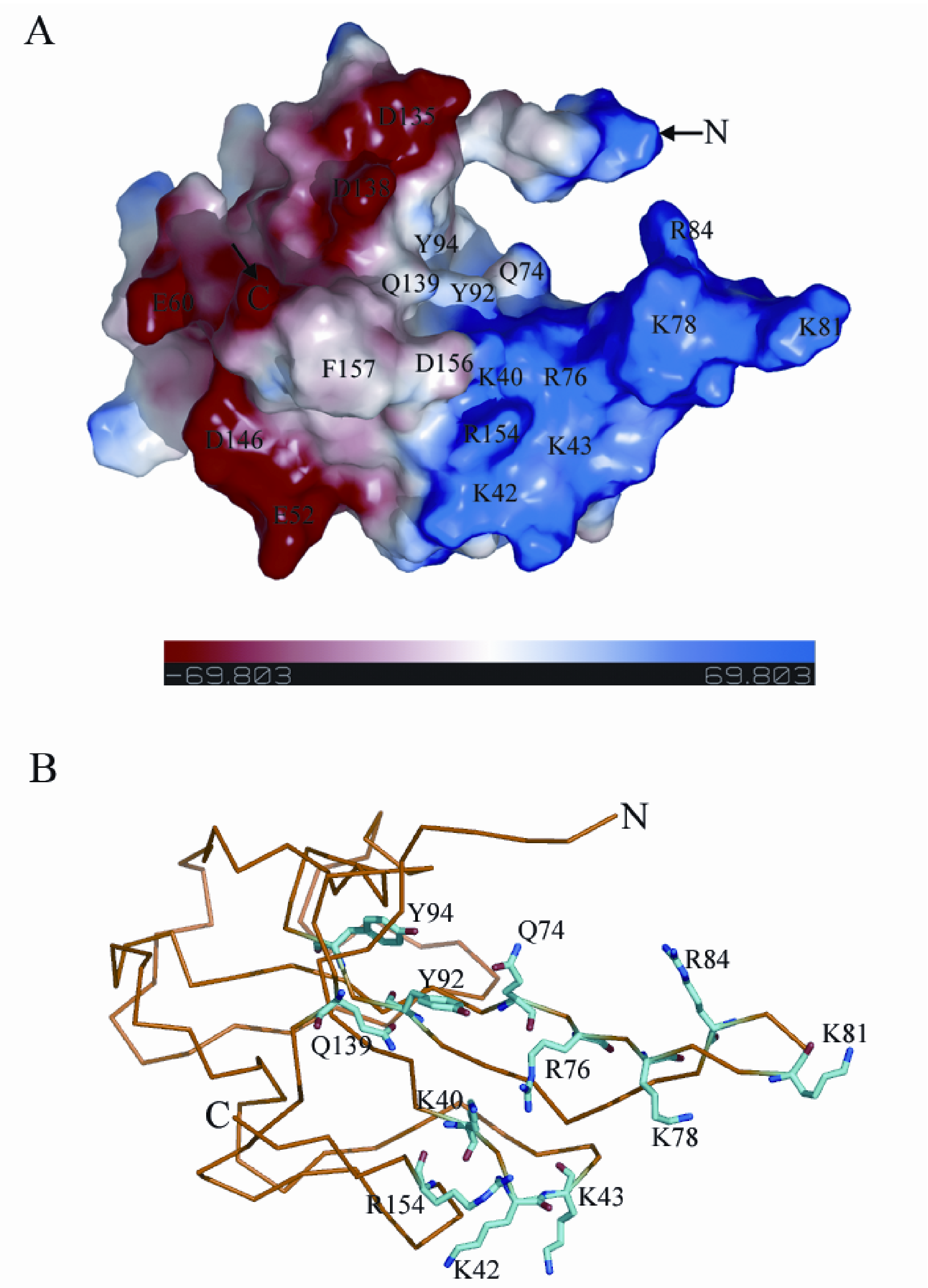


Fig. 3. 30. Electrostatic surface of the N-terminal domain and proposed RNA binding residues. **(A)** Surface representation of the IBV-N29-160 fragment with electrostatic potentials colored in blue (positive) and red (negative) respectively. Residues which are suggested to participate in RNA binding are labeled. The N and C-terminal ends of the polypeptide chains are indicated. **(B)** Close-up view of the proposed RNA binding site of the IBV N29-160 fragment. The Cα trace of IBV N29-160 is displayed. Side chains which are likely to participate in nucleic acid binding are shown as sticks.

As well as the N-terminal domain, the IBV N C-terminal was found to bind an oligoribonucleotide from the 3' end of the viral genome (Fig. 3. 29). As shown in Fig. 3. 23, the C-terminal domains form a tightly associated dimer with an anti-parallel β-strand side and a helix groove side. Analyzing the residues on the surface of the two sides will help understand how the C-terminal domain interacts with RNA. The residues on the β-strand floor composed of 4 anti-parallel β-strands (β1-β2-β'1-β'2), especially positive charged amino acids 247K, 249K, 286R, 290R and 299R, form an "X" shape positive distribution which provides a large area which are potential RNA binding sites (Fig. 3. 31a).

In the middle of the helix groove, there is a hydrophilic centre comprising of 277S, 278S. Surrounding that, there are positive charged resides 229K and 279H and 226R and 264K on the outer layer. The phosphate groups of RNA may interact with positive residues, while the sugar and base may contact the hydrophobic resides like 272L which are around the hydrophilic centre (Fig. 3. 31b).

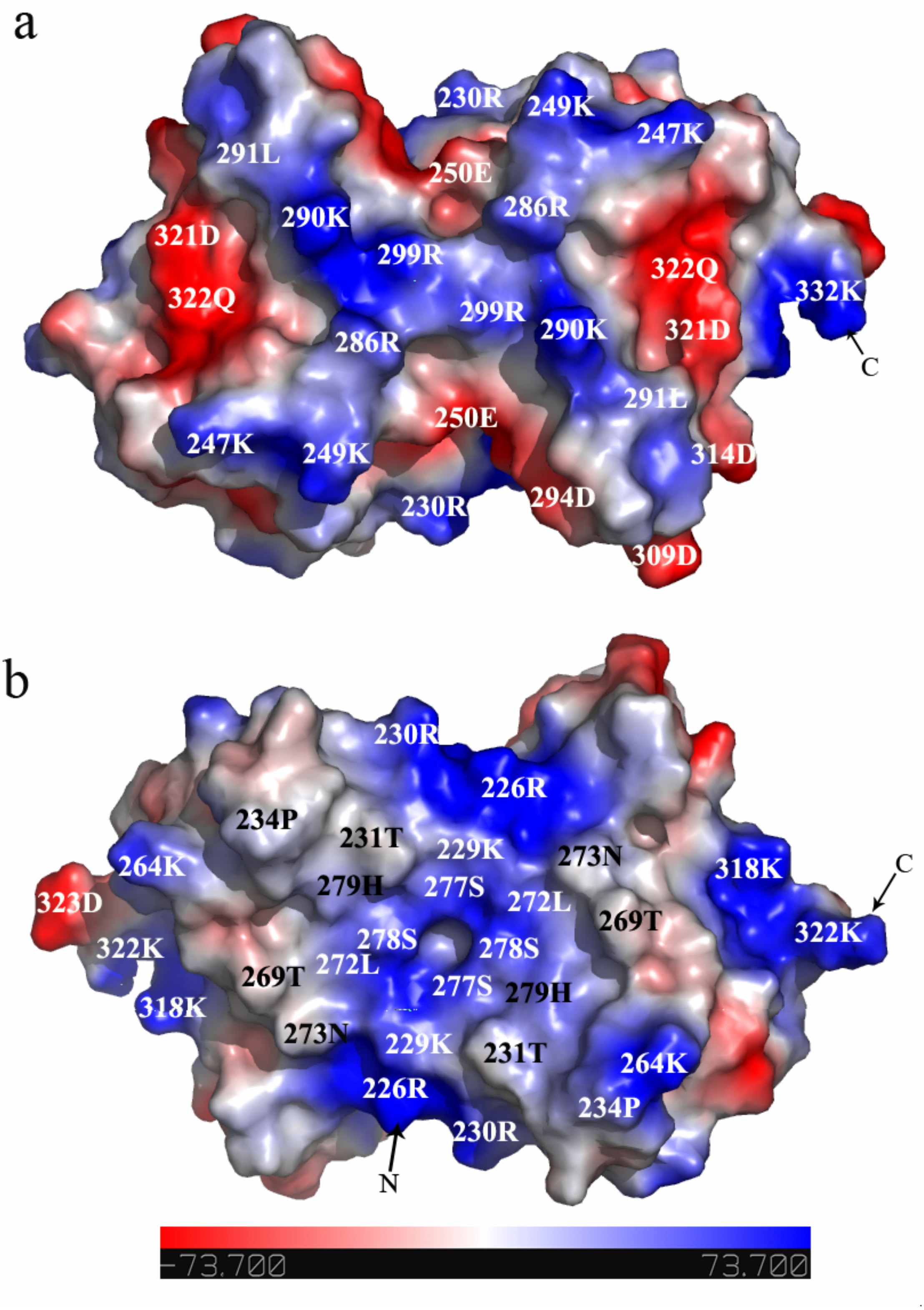


Fig. 3. 31. Electrostatic surface of the C-terminal domain.

Exposed residues on anti-parallel β strand floor **(a)** and helix groove **(b)** are labeled.

RNA viruses have high rates of sequence divergence and genome recombination. Thus, it is a great challenge to study the evolutionary relationships among viruses. Structural studies have provided another important tool to reveal distant relationships among viruses. Among positive-strand ssRNA viruses, high resolution structures of the nucleocapsid or capsid proteins are currently reported from different virus families. Although these nucleocapsids have the equivalent function to package the viral genome into virions, the size and structure of the nucleocapsid proteins are remarkably diverse among virus families. Sindbis virus and Semliki Forest virus in the *Togaviridae* family are small single-stranded RNA viruses. The C-terminal domains of nucleoproteins from Sindbis virus and Semliki Forest virus adopt a chymotrypsin-like β-barrel fold (Choi *et al.*, 1997; Choi *et al.*, 1991). The core proteins from the West Nile and dengue virus in the *Flaviviridae* are dimers composed entirely of α-helical bundles (Dokland *et al.*, 2004). As indicated by a systematic structural homology search (Holm *et al.*, 1998), the coronavirus N protein C-terminal domain closely resembles the nucleocapsid protein of PRRSV in the *Arteriviridae* (Fig 3.33). Coronaviruses are related to arteriviruses by their similar genome organization and viral replication mechanisms (Doan & Dokland, 2003). *Coronaviridae* and *Arteriviridae* form the order *Nidovirales*. The PRRSV capsid-forming domain (C-terminal 73 to 123 amino acid residues) has a similar dimeric structure (Doan & Dokland, 2003) and exhibits self-association mediated by a salt bridge as seen in the IBV N C-terminal domain (Gray strain) (Fig. 3. 32).

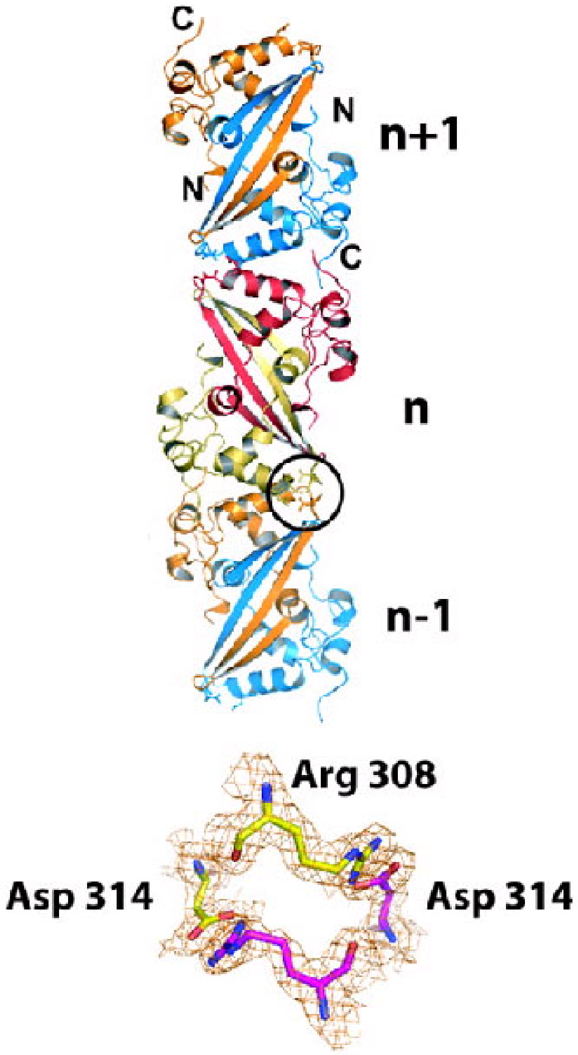


Fig. 3. 32. Crystal-packing interactions between the C-terminal dimers of the IBV (Gray strain) N protein.

In crystal-packing interactions of space group P2(1)2(1)2(1), one dimer is in the asymmetric unit (ASU). Three consecutive dimers from the neighboring ASU (numbered n, n -1, and n+1) related by one of the three orthogonal 2_1 screw axes are shown. Each monomer has been given a different color. The N- and C-terminal ends for the n +1 are indicated. The salt bridge interaction between dimers n and n-1 seen in the interface is circled. A closeup view of the salt bridge interaction with an electron density map is shown in the inset below.

Thus, the observed structural similarity between the N proteins of the IBV and PRRSV suggests that the members of the *Coronaviridae* and *Arteriviridae* families (order *Nidovirales*) share a possible common origin.

Interestingly, coronaviruses, which are clearly related to arteriviruses based on their similar genomic organization, are structurally quite different and have a much larger capsid protein possibly forming a helical structure. Unlike the C-terminal domain fold which is shared with the arterivirus PRRSV, the fold of the N-terminal domain is observed only in the coronavirus N proteins. The corresponding basic N-terminal RNA binding domain in the much shorter PPRSV N protein appears to be largely disordered (Doan & Dokland, 2003).

MS2 is a single-stranded RNA bacteriophage of 3569 nucleotides that is able to infect *E. coli* bacteria (Nagai *et al*., 1996). The single-stranded RNA genome is packaged in a protein shell that consists of 180 copies of the coat protein arranged with a T=3 icosahedral symmetry (Nagai, 1996). The binding of a coat protein dimer (MS2) to a stem loop structure of 19-nucleotide is a convenient model system for RNA-protein interaction (Valegard *et al*., 1994). By binding to a stem loop structure of 19-nucleotide containing the initiation codon of the replicase gene, the MS2 coat protein shuts off the synthesis of replicase, switching the viral replication cycle to virus assembly. The observed protein-RNA interaction arises primarily from the conserved residues within the β strands of the coat protein dimer. The conserved residues involving in RNA interaction were identified by mutational studies (Valegard *et al*., 1994).

The U1A protein is a component of U1 snRNP, a large RNA-protein complex involved in pre-mRNA splicing; it binds to hairpin II within U1 snRNA (Scherly *et al*., 1989; Nagai *et al*., 1996). In RNA-binding proteins, such as MS2 bacteriophage coat protein, U1A protein and aspartyl-tRNA synthetase (Nagai, 1996), RNA binds to the surface of a β strand. Bases in the single-stranded RNA have a strong tendency to stack on either adjacent bases or aromatic side chains. Presumably the β strand provides a large surface on which RNA bases can be splayed out (Nagai, 1996). The IBV N C-terminal domain antiparallel β strand floor consisting of β1-β2-β1-β2 is likely to be the nucleic acid binding site. This arrangement is similar to MS2 bacteriophage coat protein floor (Valegard *et al*., 1994). Superposition of the IBV N C-terminal domain β1-β2-α3 and MS2 coat protein floor is shown in Fig. 3. 33e. 286 Arg, 288 Thr, 299 Arg and 301 Glu in IBV C-terminal domain correspond to residues in MS2 coat protein responsible for RNA binding. Presumably, these residues in IBV N C-terminal domain antiparallel β strand are also responsible for RNA binding.

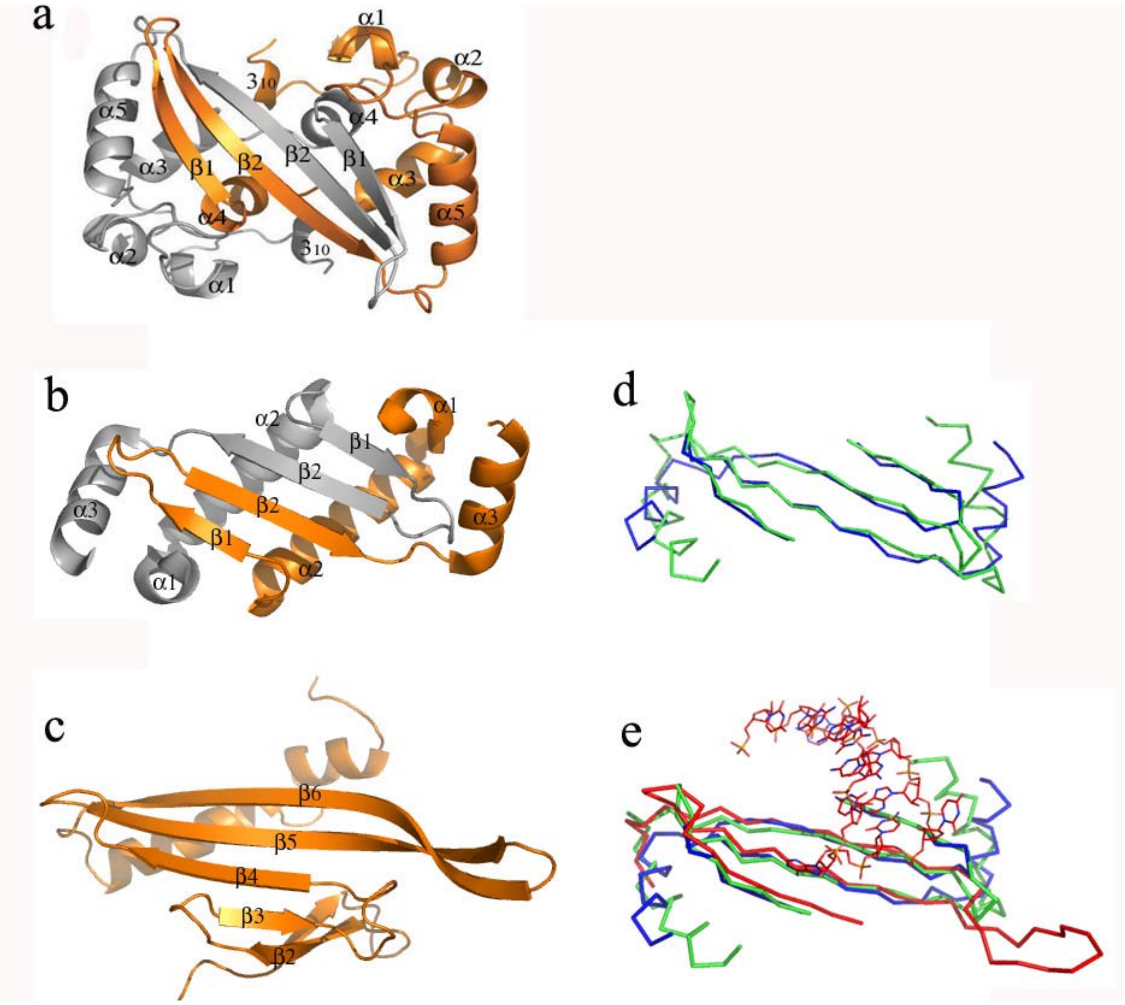


Fig. 3. 33. Comparison of the C-terminal domain to PRRSV N protein and MS2 coat protein.

IBV N C-terminal dimerization domain, PRRSV N protein and MS2 coat protein are shown in **(a)**, **(b)** and **(c)**, respectively. The three structural alignment between IBV N C-terminal's β1-β2-α5 (blue) and PRRSV N protein's β1-β2-α3 (green) shows that a total of 27 equivalent Cα atoms can be superimposed with rms deviation 3Å **(d)**. β1-β2-α5 of the IBV N C-terminal domain and PRRSV N protein is similar to MS2 coat protein's binding sites as shown by superimposing of β1-β2-α5 of the IBV N C-terminal domain and PRRSV N protein with MS2 coat protein's β4-β5-β6 (red) **(e)**. The 19 nucleotides RNA which binds to the MS2 is represented by stick.

3.4 possible model for nucleocapsid formation

A flexible filamentous nucleocapsid formed by the close association of the N proteins with viral genomic RNA is a common feature in many enveloped ssRNA viruses including coronaviruses. Cryoelectron microscope studies on transmissible gastroenteritis virus (TGEV) show that virus core contains the filamentous (possibly helical) nucleocapsid, a structure consisting of the viral RNA and N protein (Risco *et al.*, 1996).

The formation of the coronavirus nucleocapsid involves self-association of the N protein and interaction with RNA resulting in a structure which is RNase resistant. Our crystal structure analysis of the IBV N C-terminal domain reveals a tightly associated dimer, suggesting that the full-length N protein is likely to function as a dimer, with the C-terminal domain providing a structural scaffold and both the N- and C-terminal domains serving as a module for RNA interaction (Fig. 3. 34). The relative orientation of the N-terminal domain with respect to the C-terminal domain in the N protein remains unknown, because in the full-length protein these two domains are connected by a 57-residue protease-sensitive loop (R.G.S rich loop), rich in serine and glycine residues, which is presumably mobile. In the filamentous or helical ribonucleocapsid, the RNA binding regions of these two domains could face each other, engulfing the RNA between them, thus conferring resistance to RNases.

Spontaneous “self-assembly” of tobacco mosaic virus (TMV) has been extensively studied *in vitro*. Native TMV coat protein expressed in and purified from *E.coli* form stacked aggregates. TMV assembly is initiated by a specific interaction between coat protein aggregate and an RNA containing the origin-of-assembly sequence (Huang *et al.*, 1994).

The IBV N protein was observed to form multimers by electron microscopy, size exclusion chromatography and dynamic light scattering. Multimerization of the IBV N proteins are capable of increasing the protein surface area accessible for binding the viral genomic RNA thus providing the elementary building block for nucleocapsid assembly. Indeed, several crystal structures of capsid proteins have revealed the presence of multimers that present continuous patches of basic residues at their surface: the capsid proteins of the West Nile Virus and Borna disease virus form tetrameric assemblies (Dokland *et al.*, 2004; Rudolph *et al.*, 2003) and the nucleocapsid protein of porcine respiratory syndrome virus (PRRSV), an arterivirus, form dimers (Doan & Dokland, 2003). Unfortunately, since these structures were determined in the absence of an RNA ligand, it is difficult to evaluate to which extent multimer formation is coupled with nucleic acid recognition. In TMV, native coat protein expressed in and purified from *E.coli* form nonhelical, stacked aggregates. However, *in vivo* coexpression of coat protein and single-stranded RNA containing the TMV origin-of-assembly sequence give high yields of helical pseudovirus particles. The IBV N protein aggregation *in vitro* may result from lack of virus RNA which may play important role in nucleocapsid formation (Huang *et al.*, 1994).

Further complexity for the study of coronavirus nucleocapsid assembly stems from its interaction with the M protein endodomain (Kou & Masters, 2002; Narayanan *et al.*, 2000, 2003) and from the fact that several coronavirus proteins can interact with single strand RNA, including the nsp9 replicase protein from the SARS-CoV (Egloff *et al*., 2004; Sutton *et al*., 2004). In the absence of a nucleic acid ligand, the N protein is composed of two main globular domains loosely connected by an Arg/Gly/Ser (R.G.S) rich loop that is highly sensitive to proteolysis. These connecting regions may undergo modifications (eg: phosphorylation) that could influence the multimerization state of the protein and control its interaction with RNA. Sumoylation of Lys-62 of SARS-CoV N protein expressed in mammalian cells was proposed to promote dimerization of the protein (Li *et al*., 2005). It is not known whether similar modifications of the IBV N protein occur in virus infected cells.

Nevertheless, a possible model for coronavirus nucleocapsid formation can be proposed (Fig. 3. 34). The C-terminal dimerization domains are likely to form a scaffold of helical nucleocapsid through dimerization and inter-dimer stacking (Fig. 3. 34a). The amino-terminus of the C-terminal domain is pointing out from the helix structure as shown in gold color in Fig. 3. 34b, from where the N-terminal domain of the IBV N protein was connected to helical structure by the Arg/Gly/Ser (R.G.S) rich loop (residues 161-217) (Fig. 3. 8) located between the N- and C-terminal domains of the IBV N protein (Fig. 3. 34c).

In the model, the N terminal domains on outer shell are mainly responsible for RNA binding. The flexible R.G.S rich loop connected the N-terminal domain to helical structure may allow the N-terminal domain to adjust to grab RNA. The C-terminal domains of the IBV N protein are mainly responsible for forming the rigid helical structure by dimerization and interdimer stacking. The carboxyl terminus of the M protein was reported to interact with N protein carboxyl terminus (Kuo & Masters., 2002; Verma *et al*., 2006). The N protein carboxyl terminus is located in the interior shell of the helical structure in our model. The M protein carboxyl terminus may bind to the interior shell of the helical nucleocapsid and stabilize the helical structure. The virus RNA may induce helical nucleocapsid formation and is essential part of helical nucleocapsid as reported in TMV. In addition, the nucleocapsid may protect the RNA from RNase digestion.

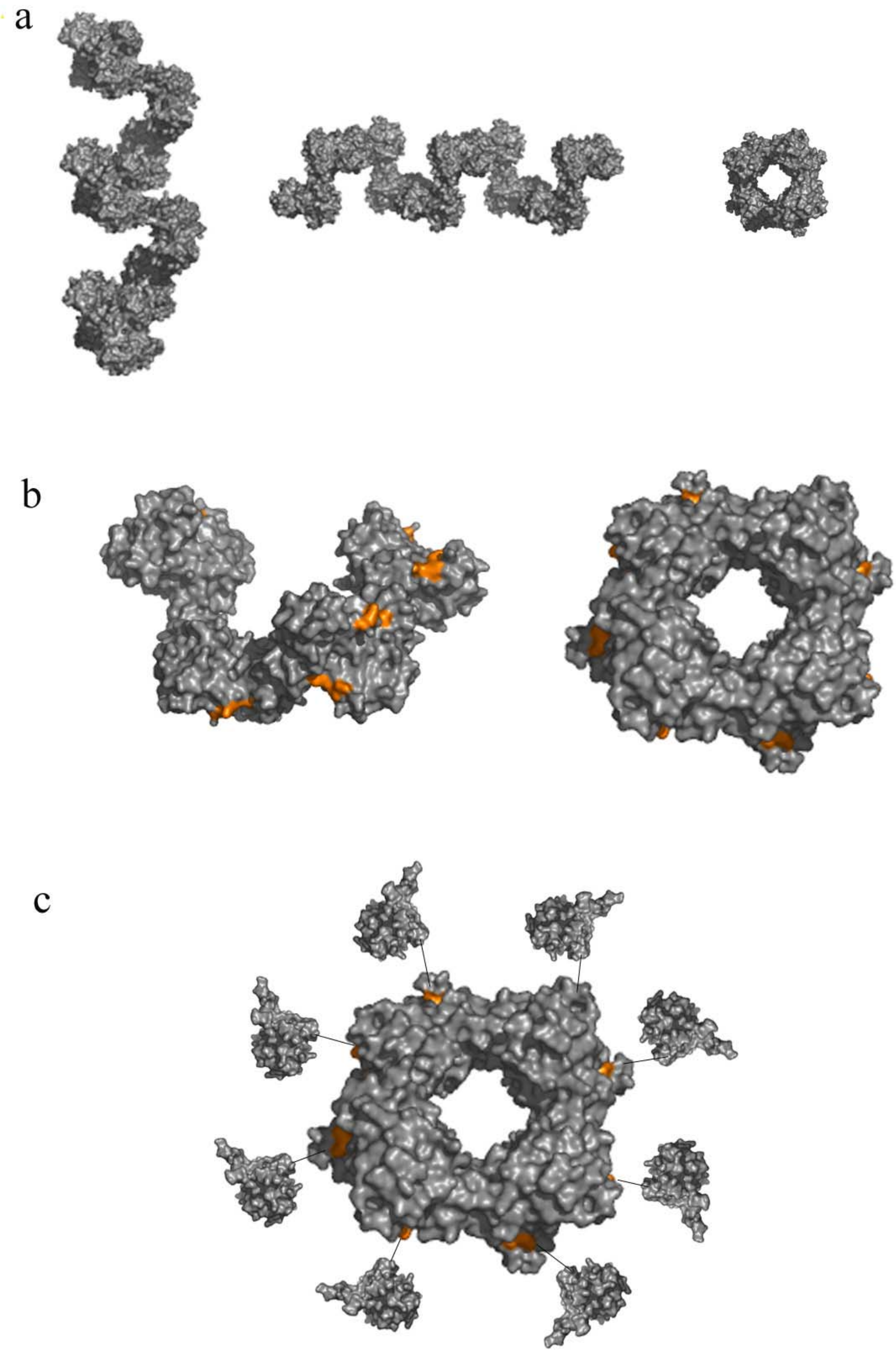


Fig. 3. 34. Possible model for the IBV helical nucleocapsid formation.
(A). Surface representation of helical structure formed by the N protein C-terminal domains dimerization and inter-dimer stacking through 4_3 screw axis. (b) Amino terminuses of the N protein C-terminal domains are colored in gold. (c) The N protein N-terminal domains are on the outer shell connecting to amino terminuses of the C-terminal domains by R.G.S rich loops which are symbolized by straight line.

CHAPTER 4. DISCUSSIONS AND FUTURE WORK

We described the structures of the two proteolytically resistant domains of the N protein from the IBV. These two domains are located at its N- and C-terminal ends, respectively. The structural details of these two domains provide a module for the N protein specific interactions with RNA and nucleocapsid formation.

However, some questions, as we describe below, remain unanswered and need to be addressed in future work.

4.1 The interaction between the N protein and RNA

The assembly of a virus particle is a relatively complex process that involves a large number of protein subunits interacting with each other and with the viral nucleic acid (Liljias, 1999). The viral nucleic acid is specifically recognized in a way that avoids the packaging of unrelated nucleic acids. The packaging signals were seen to form a stem-loops structure that is recognized specifically by a nucleocapsid protein. The best characterized protein-nucleic acid recognition is the binding of coat protein of RNA bacteriophage MS2 to a stem-loop in the single-stranded viral RNA (Valegard *et al.*, 1994). The interaction of the MS2 capsid protein dimer and the RNA stem-loop has been studies extensively. The protein-RNA contacts arise primarily from conserved residues within the β-sheets of the capsid protein dimer (Valegard *et al.*, 1994). The basic residues interact with the phosphate groups of the 19-nucleotide stem-loop RNA mostly by hydrogen bonds.

Our studies indicate that both N and C-terminal domains of IBV N protein are capable of binding an oligoribonucleotide from the 3’ end of viral genome which contains a highly conserved 32-nucleotide sequence across various coronaviruses. Interestingly, the conserved 32-nucleotide RNA forms a stem-loop structure (Robertson, *et al.*, 2005). The packaging signals have been identified at 3’ end of the viral genome of MHV (Fosmire *et al.*, 1992), BCV (Cologna & Hogue, 2000) and at the 5’ end of the TGEV (Escors *et al.*, 2003), but not unambiguously for the IBV genome. It is unknown whether the conserved 32-nucleotide or other sequences in IBV genome is the packaging signal.

Our structural studies show that the basic residues in the flexible hairpin-loop (β_2’-β_3’) of the IBV N N-terminal domain are potential RNA-binding sites. In addition, the C-terminal domain of the IBV N protein with an antiparallel β-sheet floor flanked by several helices has similar secondary structure elements as other RNA-binding proteins such as MS2 coat protein and PRRSV N protein. Antiparallel β-sheet is very common secondary structure element found in ribonucleoproteins (RNPs) for RNA binding (Mattaj, 1993; Nagai, 1996). Presumably, the sheets provide a large surface on which RNA can be splayed out (Nagai, 1996). However, we lack atomic details regarding the interaction between the N protein and RNA.

Identifying the IBV packaging signal will be crucial for crystallizing the N protein (N or C-terminal domain) complexed with RNA. The structure of the protein-RNA complex will indicate which amino acids contact with RNA (phosphates groups or bases). Based on the complex structure, it could be possible to design short peptides or compounds which can potentially inhibit the association of the N protein and RNA.

4.2 The nucleocapsid formation

The coat proteins of many simple viruses can form virus-like particle when they are expressed in a suitable expression system. These protein components seem to have all the properties that are needed for the formation of large particles (Liljias, 1999). The IBV N protein, as shown by our results, is a multimerization protein. The IBV N C-terminal domain seems to have the property to form a super helical structure through dimerization and inter-dimer stacking. Both the IBV N protein N- and C-terminal domains are likely to interact with the RNA genome, presumably the stem loop, to form the viral nucleocapsid.

We cannot classify the IBV capsid (helix, icosahedron or others) from the negative staining electron microscopy (EM) due to its low resolution. Compared to the negative staining EM, the cryo-EM can reveal higher resolution data with less distortions of the structure caused by drying, flattening, non-uniform staining and radiation damage (Baker *et al*., 1999). Cryo-EM in combination with image reconstruction has long been a primary tool for classifying viruses and exploring their structure. Additionally, the N-terminal and C-terminal domains of the IBV N, which were determined by x-ray crystallography, can be fitted into the cryo-EM reconstruction of the viral capsid. The fit of crystallographic model into cryo-EM reconstruction was used by Lescar *et al* (2001) to solve the fusion glycoprotein shell of Semliki Forest Virus, an alphavirus.

The encapsidation of the viral genome in small icosahedral viruses seems to occur by the condensation of the coat protein around the nucleic acid molecule. Self-assembly of the viral capsid protein also depends on the pH and ionic strength (Lepault *et al.*, 2001). More studies by using cryo-EM and image reconstruction are required to know whether the IBV N protein is able to self-assemble in the presence and absence of RNA, and under different conditions of ATP, ADP, phosphate, cations and pH that may affect the dynamics of the N protein self-assembly.

4.3 The interaction between the N and M protein

In addition to its interaction with RNA, the N protein also interacts with the M protein embedded in the viral membrane. The M protein is thought to possess three transmembrane segments and a large C-terminal endodomain that interacts with the nucleocapsid and possibly also with the RNA genome (Sturman *et al.*, 1980; Kou & Masters, 2002; Narayanan *et al.*, 2003). Based on reverse genetic complementation assays, the interaction region between these two proteins has been mapped to their C termini in mouse hepatitis virus (MHV) (Kuo & Masters, 2002). The C terminus of the M protein is significantly basic, and recent mutational studies on the M protein have demonstrated that its interaction with the N protein is predominantly electrostatic in nature (Luo *et al.*, 2006). Cryo-EM studies on transmissible epidemic diarrhea virus (TGEV) core structure also show that the M protein is the main structural component of the core shell (Risco *et al.*, 1996). To date, the interaction between the M and N protein in IBV is not determined. Due to low sequence identity of the N proteins across coronaviruses, we can not assume that the IBV N protein C-terminus interacts with the M protein like TGEV and MHV. Identification of the M-N protein interaction in IBV will contribute to our understanding of the virus assembly and core structure

4.4 The phosphorylation of the N protein

Several coronavirus N proteins are phosphorylated, including IBV, MHV and TGEV (Chen *et al*., 2005). The role of phosphorylation in the virus life cycle is unknown, although the phosphorylated IBV N proteins were reported to have a higher binding affinity with viral RNA than non-viral RNA (Chen *et al*., 2005). Our result shows that the Arg/Gly/Ser (R.G.S)-rich loop (residues 161-217) between the two domains of the N protein may be phosphorylated post-translationally. 9 serines in R.G.S rich loop are predicted to be phosphorylated, which is consistent with the observation that the IBV N protein Ser190, Ser192, Thr378 and Ser378 are phosphorylated (Chen *et al*., 2005). The Ser190 and Ser192, and Thr378 and Ser378 are located in the N protein R.G.S rich loop and C-tail (Fig. 3. 8), respectively. The involvement of the phosphorylated R.G.S rich loop and C-tail in RNA binding remains an unanswered question. Further investigations are needed to identify the precise role of the IBV N protein phosphorylation.

Reference

(1994). "The CCP4 suite: programs for protein crystallography." Acta Crystallogr D Biol Crystallogr 50(Pt 5): 760-3.

Anand, K., J. Ziebuhr, P. Wadhwani, J. R. Mesters and R. Hilgenfeld (2003). "Coronavirus main proteinase (3CLpro) structure: basis for design of anti-SARS drugs." Science 300(5626): 1763-7.

Anderson, R. and F. Wong (1993). "Membrane and phospholipid binding by murine coronaviral nucleocapsid N protein." Virology 194(1): 224-32.

Baker, S. C., C. K. Shieh, L. H. Soe, M. F. Chang, D. M. Vannier and M. M. Lai (1989). "Identification of a domain required for autoproteolytic cleavage of murine coronavirus gene A polyprotein." J Virol 63(9): 3693-9.

Baker TS, Olson NH, Fuller SD (1999) Adding the third dimension to virus life cycles: three-dimensional reconstruction of icosahedral viruses from cryo-electron micrographs. Microbiol Mol Biol Rev 63:862-922, table of contents

Barretto, N., Jukneliene, D., Ratia, K., Chen, Z., Mesecar, A.D., Baker, S.C. (2006) Deubiquitinating activity of the SARS-CoV papain-like protease. Adv Exp Med Biol 581:37-41

Baric, R. S., G. W. Nelson, J. O. Fleming, R. J. Deans, J. G. Keck, N. Casteel and S. A. Stohlman (1988). "Interactions between coronavirus nucleocapsid protein and viral RNAs: implications for viral transcription." J Virol 62(11): 4280-7.

Bartlam, M., H. Yang and Z. Rao (2005). "Structural insights into SARS coronavirus proteins." Curr Opin Struct Biol 15(6): 664-72.

Bhella, D., Ralph, A., Lindsay, B., Murphy, B. and Yeo, R.P. (2002) Significant differences in nucleocapsid morphology within the Paramyxoviridea. J of General Viro. 83, 1831-1839.

Bisht, H., A. Roberts, L. Vogel, A. Bukreyev, P. L. Collins, B. R. Murphy, K. Subbarao and B. Moss (2004). "Severe acute respiratory syndrome coronavirus spike protein expressed by attenuated vaccinia virus protectively immunizes mice." Proc Natl Acad Sci U S A 101(17): 6641-6.

Blom, N., Gammeltoft, S., and Brunak, S. (1999). “Sequence- and structure-based prediction of eukaryotic protein phosphorylation sites”. Journal of Molecular Biology 294(5): 1351-1362.

Bosch, B. J., B. E. Martina, R. Van Der Zee, J. Lepault, B. J. Haijema, C. Versluis, A. J. Heck, R. De Groot, A. D. Osterhaus and P. J. Rottier (2004). "Severe acute respiratory syndrome coronavirus (SARS-CoV) infection inhibition using spike protein heptad repeat-derived peptides." Proc Natl Acad Sci U S A 101(22): 8455-60.

Braaten D, Franke EK, Luban J (1996) Cyclophilin A is required for the replication of group M human immunodeficiency virus type 1 (HIV-1) and simian immunodeficiency virus SIV(CPZ)GAB but not group O HIV-1 or other primate immunodeficiency viruses. J Virol 70:4220-7

Brierley, I., M. E. Boursnell, M. M. Binns, B. Bilimoria, V. C. Blok, T. D. Brown and S. C. Inglis (1987). "An efficient ribosomal frame-shifting signal in the polymerase-encoding region of the coronavirus IBV." Embo J 6(12): 3779-85.

Brierley, I., P. Digard and S. C. Inglis (1989). "Characterization of an efficient coronavirus ribosomal frameshifting signal: requirement for an RNA pseudoknot." Cell 57(4): 537-47.

Brockway, S. M., C. T. Clay, X. T. Lu and M. R. Denison (2003). "Characterization of the expression, intracellular localization, and replication complex association of the putative mouse hepatitis virus RNA-dependent RNA polymerase." J Virol 77(19): 10515-27.

Brunger, A. T., P. D. Adams, G. M. Clore, W. L. DeLano, P. Gros, R. W. Grosse-Kunstleve, J. S. Jiang, J. Kuszewski, M. Nilges, N. S. Pannu, R. J. Read, L. M. Rice, T. Simonson and G. L. Warren (1998). "Crystallography & NMR system: A new software suite for macromolecular structure determination." Acta Crystallogr D Biol Crystallogr 54(Pt 5): 905-21.

Buehler, P. W., R. A. Boykins, Y. Jia, S. Norris, D. I. Freedberg and A. I. Alayash (2005). "Structural and functional characterization of glutaraldehyde-polymerized bovine hemoglobin and its isolated fractions." Anal Chem 77(11): 3466-78.

Burd CG, Dreyfuss G (1994) Conserved structures and diversity of functions of RNA-binding proteins. Science 265:615-21

Cann, A.J (1997). Principles of Molecular Virology. Academic Press. 2nd Edition

Caspar D.L.D. and Klug A.,(1962). Physical principles in the construction of regular viruses. Cold Spring Harbor Symp. Quant. Biol. 27,1-24.

Cavanagh, D. (1983). "Coronavirus IBV: structural characterization of the spike protein." J Gen Virol 64 (Pt 12): 2577-83.

Chastain, M., Tinoco, I. (1991) Structural elements in RNA. Nucleic Acid Res Mol Biol. 40:131-177.

Chen, H., A. Gill, B. K. Dove, S. R. Emmett, C. F. Kemp, M. A. Ritchie, M. Dee and J. A. Hiscox (2005). "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(2): 1164-79.

Choi, H. K., L. Tong, W. Minor, P. Dumas, U. Boege, M. G. Rossmann and G. Wengler (1991). "Structure of Sindbis virus core protein reveals a chymotrypsin-like serine proteinase and the organization of the virion." Nature 354(6348): 37-43.

Choi, H. K., G. Lu, S. Lee, G. Wengler and M. G. Rossmann (1997). "Structure of Semliki Forest virus core protein." Proteins 27(3): 345-59.

Choi, K. S., P. Huang and M. M. Lai (2002). "Polypyrimidine-tract-binding protein affects transcription but not translation of mouse hepatitis virus RNA." Virology 303(1): 58-68.

Cologna, R. and B. G. Hogue (2000). "Identification of a bovine coronavirus packaging signal." J Virol 74(1): 580-3.

Cologna, R., J. F. Spagnolo and B. G. Hogue (2000). "Identification of nucleocapsid binding sites within coronavirus-defective genomes." Virology 277(2): 235-49.

Corse, E. and C. E. Machamer (2000). "Infectious bronchitis virus E protein is targeted to the Golgi complex and directs release of virus-like particles." J Virol 74(9): 4319-26.

Corse, E. and C. E. Machamer (2002). "The cytoplasmic tail of infectious bronchitis virus E protein directs Golgi targeting." J Virol 76(3): 1273-84.

De Guzman, R. N., Z. R. Wu, C. C. Stalling, L. Pappalardo, P. N. Borer and M. F. Summers (1998). "Structure of the HIV-1 nucleocapsid protein bound to the SL3 psi-RNA recognition element." Science 279(5349): 384-8.

de Haan, C. A., M. Smeets, F. Vernooij, H. Vennema and P. J. Rottier (1999). "Mapping of the coronavirus membrane protein domains involved in interaction with the spike protein." J Virol 73(9): 7441-52.

de Haan, C. A., P. S. Masters, X. Shen, S. Weiss and P. J. Rottier (2002). "The group-specific murine coronavirus genes are not essential, but their deletion, by reverse genetics, is attenuating in the natural host." Virology 296(1): 177-89.

de Haan, C. A., K. Stadler, G. J. Godeke, B. J. Bosch and P. J. Rottier (2004). "Cleavage inhibition of the murine coronavirus spike protein by a furin-like enzyme affects cell-cell but not virus-cell fusion." J Virol 78(11): 6048-54.

de Haan, C. A., L. Kuo, P. S. Masters, H. Vennema and P. J. Rottier (1998). "Coronavirus particle assembly: primary structure requirements of the membrane protein." J Virol 72(8): 6838-50.

de Haan, C. A. and P. J. Rottier (2005). "Molecular interactions in the assembly of coronaviruses." Adv Virus Res 64: 165-230.

DeLano, W.L. The PyMOL User's Manual. (2002). DeLano Scientific, San Carlos, CA, USA.

Delmas, B., J. Gelfi, R. L'Haridon, L. K. Vogel, H. Sjostrom, O. Noren and H. Laude (1992). "Aminopeptidase N is a major receptor for the entero-pathogenic coronavirus TGEV." Nature 357(6377): 417-20.

Doan, D. N. and T. Dokland (2003). "Structure of the nucleocapsid protein of porcine reproductive and respiratory syndrome virus." Structure 11(11): 1445-51.
Dokland, T., M. Walsh, J. M. Mackenzie, A. A. Khromykh, K. H. Ee and S. Wang (2004). "West Nile virus core protein; tetramer structure and ribbon formation." Structure 12(7): 1157-63.

Dokland T, Walsh M, Mackenzie JM, Khromykh AA, Ee KH, Wang S (2004) West Nile virus core protein; tetramer structure and ribbon formation. Structure 12:1157-63

Doublie, S. (1997) Preparation of selenomethionyl proteins for phase determination. Methods Enzymol 276:523-30

Dreyfuss G, Matunis MJ, Pinol-Roma S, Burd CG (1993) hnRNP proteins and the biogenesis of mRNA. Annu Rev Biochem 62:289-321

Duarte, M., K. Tobler, A. Bridgen, D. Rasschaert, M. Ackermann and H. Laude (1994). "Sequence analysis of the porcine epidemic diarrhea virus genome between the nucleocapsid and spike protein genes reveals a polymorphic ORF." Virology 198(2): 466-76.

Dveksler, G. S., C. W. Dieffenbach, C. B. Cardellichio, K. McCuaig, M. N. Pensiero,

G. S. Jiang, N. Beauchemin and K. V. Holmes (1993). "Several members of the mouse carcinoembryonic antigen-related glycoprotein family are functional receptors for the coronavirus mouse hepatitis virus-A59." J Virol 67(1): 1-8.

Egelman, E. H., Wu, S. S., Amrein, M., Portner, A. & Murti, G. (1989). The Sendai virus nucleocapsid exists in at least four different helical states. J Virol **63**, 2233-2243.

Egloff, M. P., F. Ferron, V. Campanacci, S. Longhi, C. Rancurel, H. Dutartre, E. J. Snijder, A. E. Gorbalenya, C. Cambillau and B. Canard (2004). "The severe acute respiratory syndrome-coronavirus replicative protein nsp9 is a single-stranded RNA-binding subunit unique in the RNA virus world." Proc Natl Acad Sci U S A 101(11): 3792-6.

Emsley, P. and K. Cowtan (2004). "Coot: model-building tools for molecular graphics." Acta Crystallogr D Biol Crystallogr 60(Pt 12 Pt 1): 2126-32.

Escors, D., A. Izeta, C. Capiscol and L. Enjuanes (2003). "Transmissible gastroenteritis coronavirus packaging signal is located at the 5' end of the virus genome." J Virol 77(14): 7890-902.

Escors, D., E. Camafeita, J. Ortego, H. Laude and L. Enjuanes (2001). "Organization of two transmissible gastroenteritis coronavirus membrane protein topologies within the virion and core." J Virol 75(24): 12228-40.

Evans, G., and Pettifer, R.F., (2001). CHOOCH: a program for deriving anomalous-scattering factors from X-ray fluorescence spectra. J. Appl. Crystallogr. 34, 82-86

Fosmire, J. A., K. Hwang and S. Makino (1992). "Identification and characterization of a coronavirus packaging signal." J Virol 66(6): 3522-30.

Gassner, N. C. and B. W. Matthews (1999). "Use of differentially substituted selenomethionine proteins in X-ray structure determination." Acta Crystallogr D Biol Crystallogr 55(Pt 12): 1967-70.

Ghisolfi L, Kharrat A, Joseph G, Amalric F, Erard M (1992) Concerted activities of the RNA recognition and the glycine-rich C-terminal domains of nucleolin are required for efficient complex formation with pre-ribosomal RNA. Eur J Biochem 209:541-8

Gibson TJ, Thompson JD, Heringa J (1993) The KH domain occurs in a diverse set of RNA-binding proteins that include the antiterminator NusA and is probably involved in binding to nucleic acid. FEBS Lett 324:361-6

Godet, M., J. Grosclaude, B. Delmas and H. Laude (1994). "Major receptor-binding and neutralization determinants are located within the same domain of the transmissible gastroenteritis virus (coronavirus) spike protein." J Virol 68(12): 8008-16.

Gomis-Ruth, F. X., A. Dessen, J. Timmins, A. Bracher, L. Kolesnikowa, S. Becker, H. D. Klenk and W. Weissenhorn (2003). "The matrix protein VP40 from Ebola virus octamerizes into pore-like structures with specific RNA binding properties." Structure 11(4): 423-33.

Gorlach M, Wittekind M, Beckman RA, Mueller L, Dreyfuss G (1992) Interaction of the RNA-binding domain of the hnRNP C proteins with RNA. Embo J 11:3289-95

Green SR, Mathews MB (1992) Two RNA-binding motifs in the double-stranded RNA-activated protein kinase, DAI. Genes Dev 6:2478-90

Harcourt, B. H., D. Jukneliene, A. Kanjanahaluethai, J. Bechill, K. M. Severson, C. M. Smith, P. A. Rota and S. C. Baker (2004). "Identification of severe acute respiratory syndrome coronavirus replicase products and characterization of papain-like protease activity." J Virol 78(24): 13600-12.

Hiscox, J. A., T. Wurm, L. Wilson, P. Britton, D. Cavanagh and G. Brooks (2001). "The coronavirus infectious bronchitis virus nucleoprotein localizes to the nucleolus." J Virol 75(1): 506-12.

Hoffman DW, Query CC, Golden BL, White SW, Keene JD (1991) RNA-binding domain of the A protein component of the U1 small nuclear ribonucleoprotein analyzed by NMR spectroscopy is structurally similar to ribosomal proteins. Proc Natl Acad Sci U S A 88:2495-9

Hogue, B. G., T. E. Kienzle and D. A. Brian (1989). "Synthesis and processing of the bovine enteric coronavirus haemagglutinin protein." J Gen Virol 70 (Pt 2): 345-52.

Holm, L. and C. Sander (1993). "Protein structure comparison by alignment of distance matrices." J Mol Biol 233(1): 123-38.

Holm, L. and C. Sander (1998). "Touring protein fold space with Dali/FSSP." Nucleic Acids Res 26(1): 316-9.

Hu, G., P. D. Gershon, A. E. Hodel and F. A. Quiocho (1999). "mRNA cap recognition: dominant role of enhanced stacking interactions between methylated bases and protein aromatic side chains." Proc Natl Acad Sci U S A 96(13): 7149-54.

Huang, Q., L. Yu, A. M. Petros, A. Gunasekera, Z. Liu, N. Xu, P. Hajduk, J. Mack, S. W. Fesik and E. T. Olejniczak (2004). "Structure of the N-terminal RNA-binding domain of the SARS CoV nucleocapsid protein." Biochemistry 43(20): 6059-63.

Hurst, K. R., L. Kuo, C. A. Koetzner, R. Ye, B. Hsue and P. S. Masters (2005). "A major determinant for membrane protein interaction localizes to the carboxy-terminal domain of the mouse coronavirus nucleocapsid protein." J Virol 79(21): 13285-97.

Hwang, D. J., I. M. Roberts and T. M. Wilson (1994). "Expression of tobacco mosaic virus coat protein and assembly of pseudovirus particles in Escherichia coli." Proc Natl Acad Sci U S A 91(19): 9067-71.

Imbert, I., Guillemot, J.C., Bourhis, J.M., Bussetta, C., Coutard, B., Egloff, M.P., Ferron, F., Gorbalenya, A.E., Canard, B. (2006) A second, non-canonical RNA-dependent RNA polymerase in SARS coronavirus. Embo J 25:4933-42

Ivanov, K. A., V. Thiel, J. C. Dobbe, Y. van der Meer, E. J. Snijder and J. Ziebuhr (2004). "Multiple enzymatic activities associated with severe acute respiratory syndrome coronavirus helicase." J Virol 78(11): 5619-32.

Ivanov, K. A. and J. Ziebuhr (2004). "Human coronavirus 229E nonstructural protein 13: characterization of duplex-unwinding, nucleoside triphosphatase, and RNA 5'-triphosphatase activities." J Virol 78(14): 7833-8.

Johnson MA, Cann AJ (1992) Molecular determination of cell tropism of human immunodeficiency virus. Clin Infect Dis 14:747-55

Jonassen, C. M., T. O. Jonassen and B. Grinde (1998). "A common RNA motif in the 3' end of the genomes of astroviruses, avian infectious bronchitis virus and an equine rhinovirus." J Gen Virol 79 (Pt 4): 715-8.

Jones, T. A., J. Y. Zou, S. W. Cowan and Kjeldgaard (1991). "Improved methods for building protein models in electron density maps and the location of errors in these models." Acta Crystallogr A 47 (Pt 2): 110-9.

Kanjanahaluethai, A. and S. C. Baker (2000). "Identification of mouse hepatitis virus papain-like proteinase 2 activity." J Virol 74(17): 7911-21.

Karn J, Graeble MA (1992) New insights into the mechanism of HIV-1 trans-activation. Trends Genet 8:365-8

Kenan DJ, Query CC, Keene JD (1991) RNA recognition: towards identifying determinants of specificity. Trends Biochem Sci 16:214-20

Kienzle, T. E., S. Abraham, B. G. Hogue and D. A. Brian (1990). "Structure and orientation of expressed bovine coronavirus hemagglutinin-esterase protein." J Virol 64(4): 1834-8.

Kiledjian M, Dreyfuss G (1992) Primary structure and binding activity of the hnRNP U protein: binding RNA through RGG box. Embo J 11:2655-64

Klausegger, A., B. Strobl, G. Regl, A. Kaser, W. Luytjes and R. Vlasak (1999). "Identification of a coronavirus hemagglutinin-esterase with a substrate specificity different from those of influenza C virus and bovine coronavirus." J Virol 73(5): 3737-43.

Kong, X.P., Onrust, R., O'Donnell, M. and Kuriyan, J (1992). Three-dimensional structure of the beta subunit of *E. coli* DNA polymerase III holoenzyme: a sliding DNA clamp. Cell 69, 425–437.

Krishna, T.S., Kong, X.P., Gary, S., Burgers, P.M. and Kuriyan, J (1993). Crystal structure of the eukaryotic DNA polymerase processivity factor PCNA. Cell 79, 1233–1243

Kuo, L. and P. S. Masters (2002). "Genetic evidence for a structural interaction between the carboxy termini of the membrane and nucleocapsid proteins of mouse hepatitis virus." J Virol 76(10): 4987-99.

Lai, M. M. and D. Cavanagh (1997). "The molecular biology of coronaviruses." Adv Virus Res 48: 1-100.

Lai, M. M. C. a. H., K.V. (2001). Coronaviridae: the virus and their replication in fundamental virology., Lippincott Raven.

Laskowski, R. A., D. S. Moss and J. M. Thornton (1993). "Main-chain bond lengths and bond angles in protein structures." J Mol Biol 231(4): 1049-67.

Lepault J, Petitpas I, Erk I, Navaza J, Bigot D, Dona M, Vachette P, Cohen J, Rey FA (2001) Structural polymorphism of the major capsid protein of rotavirus. Embo J 20:1498-507

Lescar, J., A. Roussel, M. W. Wien, J. Navaza, S. D. Fuller, G. Wengler and F. A. Rey (2001). "The Fusion glycoprotein shell of Semliki Forest virus: an icosahedral assembly primed for fusogenic activation at endosomal pH." Cell 105(1): 137-48.

Li, F., W. Li, M. Farzan and S. C. Harrison (2005a). "Structure of SARS coronavirus spike receptor-binding domain complexed with receptor." Science 309(5742): 1864-8.

Li, F. Q., H. Xiao, J. P. Tam and D. X. Liu (2005). "Sumoylation of the nucleocapsid protein of severe acute respiratory syndrome coronavirus." FEBS Lett 579(11): 2387-96.

Liao, Y., J. Lescar, J. P. Tam and D. X. Liu (2004). "Expression of SARS-coronavirus envelope protein in Escherichia coli cells alters membrane permeability." Biochem Biophys Res Commun 325(1): 374-80.

Liao, Y., Q. Yuan, J. Torres, J. P. Tam and D. X. Liu (2006). "Biochemical and functional characterization of the membrane association and membrane permeabilizing activity of the severe acute respiratory syndrome coronavirus envelope protein." Virology.

Liljas, L. (1999). "Virus assembly." Curr Opin Struct Biol 9(1): 129-34.

Lim, K. P., L. F. Ng and D. X. Liu (2000). "Identification of a novel cleavage activity of the first papain-like proteinase domain encoded by open reading frame 1a of the coronavirus Avian infectious bronchitis virus and characterization of the cleavage products." J Virol 74(4): 1674-85.

Lim, K. P. and D. X. Liu (2001). "The missing link in coronavirus assembly. Retention of the avian coronavirus infectious bronchitis virus envelope protein in the pre-Golgi compartments and physical interaction between the envelope and membrane proteins." J Biol Chem 276(20): 17515-23.

Liu, D. X., D. Cavanagh, P. Green and S. C. Inglis (1991). "A polycistronic mRNA specified by the coronavirus infectious bronchitis virus." Virology 184(2): 531-44.

Liu, D. X. and S. C. Inglis (1991). "Association of the infectious bronchitis virus 3c protein with the virion envelope." Virology 185(2): 911-7.

Liu, D. X. and S. C. Inglis (1992). "Identification of two new polypeptides encoded by mRNA5 of the coronavirus infectious bronchitis virus." Virology 186(1): 342-7.

Liu, D. X., K. W. Tibbles, D. Cavanagh, T. D. Brown and I. Brierley (1995). "Involvement of viral and cellular factors in processing of polyprotein encoded by ORF1a of the coronavirus IBV." Adv Exp Med Biol 380: 413-21.

Longhi, S., Receveur-Brechot, V., Karlin, D., Johansson, K., Darbon, H., Bhella, D., Yeo, R., Finet, S., Canard, B. (2003) The C-terminal domain of the measles virus nucleoprotein is intrinsically disordered and folds upon binding to the C-terminal moiety of the phosphoprotein. J Biol Chem 278:18638-48

Luo H, Wu D, Shen C, Chen K, Shen X, Jiang H (2006) Severe acute respiratory syndrome coronavirus membrane protein interacts with nucleocapsid protein mostly through their carboxyl termini by electrostatic attraction. Int J Biochem Cell Biol 38:589-99

Lutz CS, Alwine JC (1994) Direct interaction of the U1 snRNP-A protein with the upstream efficiency element of the SV40 late polyadenylation signal. Genes Dev 8:576-86

Maeda, J., J. F. Repass, A. Maeda and S. Makino (2001). "Membrane topology of coronavirus E protein." Virology 281(2): 163-9.

Malim MH, Hauber J, Le SY, Maizel JV, Cullen BR (1989) The HIV-1 rev trans-activator acts through a structured target sequence to activate nuclear export of unspliced viral mRNA. Nature 338:254-7

Marra, M. A., S. J. Jones, C. R. Astell, R. A. Holt, A. Brooks-Wilson, Y. S. Butterfield, J. Khattra, J. K. Asano, S. A. Barber, S. Y. Chan, A. Cloutier, S. M. Coughlin, D. Freeman, N. Girn, O. L. Griffith, S. R. Leach, M. Mayo, H. McDonald, S. B. Montgomery, P. K. Pandoh, A. S. Petrescu, A. G. Robertson, J. E. Schein, A. Siddiqui, D. E. Smailus, J. M. Stott, G. S. Yang, F. Plummer, A. Andonov, H. Artsob, N. Bastien, K. Bernard, T. F. Booth, D. Bowness, M. Czub, M. Drebot, L. Fernando, R. Flick, M. Garbutt, M. Gray, A. Grolla, S. Jones, H. Feldmann, A. Meyers, A. Kabani, Y. Li, S. Normand, U. Stroher, G. A. Tipples, S. Tyler, R. Vogrig, D. Ward, B. Watson, R. C. Brunham, M. Krajden, M. Petric, D. M. Skowronski, C. Upton and R. L. Roper (2003). "The Genome sequence of the SARS-associated coronavirus." Science 300(5624): 1399-404.

Mattaj, I.W (1993) RNA recognition: a family matter. Cell 73:837-840.

Mathews MB, Shenk T (1991) Adenovirus virus-associated RNA and translation control. J Virol 65:5657-62

Meier, C., Aricescu, A.R., Assenberg, R., Aplin, R.T., Gilbert, R.J., Grimes, J.M., Stuart, D.I. (2006) The crystal structure of ORF-9b, a lipid binding protein from the SARS coronavirus. Structure 14:1157-65

Nagai, K. (1996). "RNA-protein complexes." Curr Opin Struct Biol 6(1): 53-61.
Narayanan, K., A. Maeda, J. Maeda and S. Makino (2000). "Characterization of the coronavirus M protein and nucleocapsid interaction in infected cells." J Virol 74(17): 8127-34.

Nagai K, Oubridge C, Jessen TH, Li J, Evans PR (1990) Crystal structure of the RNA-binding domain of the U1 small nuclear ribonucleoprotein A. Nature 348:515-20

Namy, O., Moran, S.J., Stuart, D.I., Gilbert, R.J., Brierley, I. (2006) A mechanical explanation of RNA pseudoknot function in programmed ribosomal frameshifting. Nature 441:244-7

Narayanan, K., Maeda, A., Maeda, J., Makino, S (2000). Characterization of the coronavirus M protein and nucleocapsid interaction in infected cells. J Virol. 74(17):8127-34.

Narayanan, K. and S. Makino (2001). "Cooperation of an RNA packaging signal and a viral envelope protein in coronavirus RNA packaging." J Virol 75(19): 9059-67.

Narayanan, K., C. J. Chen, J. Maeda and S. Makino (2003). "Nucleocapsid-independent specific viral RNA packaging via viral envelope protein and viral RNA signal." J Virol 77(5): 2922-7.

Natarajan, P., Lander, G. C., Shepherd, C.M., Reddy, V.S., Brooks, C.L. 3rd, Johnson, J.E. (2005) Exploring icosahedral virus structures with VIPER. Nat Rev Microbiol. 3(10):809-17.

Nelson, G. W., S. A. Stohlman and S. M. Tahara (2000). "High affinity interaction between nucleocapsid protein and leader/intergenic sequence of mouse hepatitis virus RNA." J Gen Virol 81(Pt 1): 181-8.

Nelson, C. A., A. Pekosz, C. A. Lee, M. S. Diamond and D. H. Fremont (2005). "Structure and intracellular targeting of the SARS-coronavirus Orf7a accessory protein." Structure 13(1): 75-85.

Oubridge C, Ito N, Evans PR, Teo CH, Nagai K (1994) Crystal structure at 1.92 A resolution of the RNA-binding domain of the U1A spliceosomal protein complexed with an RNA hairpin. Nature 372:432-8

Peiris, J. S., S. T. Lai, L. L. Poon, Y. Guan, L. Y. Yam, W. Lim, J. Nicholls, W. K. Yee, W. W. Yan, M. T. Cheung, V. C. Cheng, K. H. Chan, D. N. Tsang, R. W. Yung, T. K. Ng and K. Y. Yuen (2003). "Coronavirus as a possible cause of severe acute respiratory syndrome." Lancet 361(9366): 1319-25.

Puglisi JD, Tan R, Calnan BJ, Frankel AD, Williamson JR (1992) Conformation of the TAR RNA-arginine complex by NMR spectroscopy. Science 257:76-80

Putics, A., W. Filipowicz, J. Hall, A. E. Gorbalenya and J. Ziebuhr (2005). "ADP-ribose-1"-monophosphatase: a conserved coronavirus enzyme that is dispensable for viral replication in tissue culture." J Virol 79(20): 12721-31.

Raamsman, M. J., J. K. Locker, A. de Hooge, A. A. de Vries, G. Griffiths, H. Vennema and P. J. Rottier (2000). "Characterization of the coronavirus mouse hepatitis virus strain A59 small membrane protein E." J Virol 74(5): 2333-42.

Ratia, K., Saikatendu, K.S., Santarsiero, B.D., Barretto, N., Baker, S.C., Stevens, R.C., Mesecar, A.D. (2006) Severe acute respiratory syndrome coronavirus papain-like protease: structure of a viral deubiquitinating enzyme. Proc Natl Acad Sci U S A 103:5717-22

Regl, G., A. Kaser, M. Iwersen, H. Schmid, G. Kohla, B. Strobl, U. Vilas, R. Schauer and R. Vlasak (1999). "The hemagglutinin-esterase of mouse hepatitis virus strain S is a sialate-4-O-acetylesterase." J Virol 73(6): 4721-7.

Ricagno, S., Egloff, M.P., Ulferts, R., Coutard, B., Nurizzo, D., Campanacci, V., Cambillau, C., Ziebuhr, J., Canard. B. (2006) Crystal structure and mechanistic determinants of SARS coronavirus nonstructural protein 15 define an endoribonuclease family. Proc Natl Acad Sci U S A 103:11892-7

Risco, C., I. M. Anton, C. Sune, A. M. Pedregosa, J. M. Martin-Alonso, F. Parra, J. L. Carrascosa and L. Enjuanes (1995). "Membrane protein molecules of transmissible gastroenteritis coronavirus also expose the carboxy-terminal region on the external surface of the virion." J Virol 69(9): 5269-77.

Risco, C., I. M. Anton, L. Enjuanes and J. L. Carrascosa (1996). "The transmissible gastroenteritis coronavirus contains a spherical core shell consisting of M and N proteins." J Virol 70(7): 4773-7.

Robertson, M. P., H. Igel, R. Baertsch, D. Haussler, M. Ares, Jr. and W. G. Scott (2005). "The structure of a rigorously conserved RNA element within the SARS virus genome." PLoS Biol 3(1): e5.

Rota, P. A., M. S. Oberste, S. S. Monroe, W. A. Nix, R. Campagnoli, J. P. Icenogle, S. Penaranda, B. Bankamp, K. Maher, M. H. Chen, S. Tong, A. Tamin, L. Lowe, M. Frace, J. L. DeRisi, Q. Chen, D. Wang, D. D. Erdman, T. C. Peret, C. Burns, T. G. Ksiazek, P. E. Rollin, A. Sanchez, S. Liffick, B. Holloway, J. Limor, K. McCaustland, M. Olsen-Rasmussen, R. Fouchier, S. Gunther, A. D. Osterhaus, C. Drosten, M. A. Pallansch, L. J. Anderson and W. J. Bellini (2003). "Characterization of a novel coronavirus associated with severe acute respiratory syndrome." Science 300(5624): 1394-9.

Rottier, P., D. Brandenburg, J. Armstrong, B. van der Zeijst and G. Warren (1984). "Assembly in vitro of a spanning membrane protein of the endoplasmic reticulum: the E1 glycoprotein of coronavirus mouse hepatitis virus A59." Proc Natl Acad Sci U S A 81(5): 1421-5.

Rottier, P. J., G. W. Welling, S. Welling-Wester, H. G. Niesters, J. A. Lenstra and B. A. Van der Zeijst (1986). "Predicted membrane topology of the coronavirus protein E1." Biochemistry 25(6): 1335-9.

Rottier, P. J. M. (1995). The coronavirus membrane glycoprotein. In "The Coronaviridae". New York, Plenum press.

Rudolph, M. G., I. Kraus, A. Dickmanns, M. Eickmann, W. Garten and R. Ficner (2003). "Crystal structure of the borna disease virus nucleoprotein." Structure 11(10): 1219-26.

Sarma, J. D., E. Scheen, S. H. Seo, M. Koval and S. R. Weiss (2002). "Enhanced green fluorescent protein expression may be used to monitor murine coronavirus spread in vitro and in the mouse central nervous system." J Neurovirol 8(5): 381-91.

Schechter, I. and Berger, A (1967). “On the size of the active site in proteases. I. Papain”. Biochem. Biophys. Res. Commun. 27: 157-162.

Scherly D, Boelens W, van Venrooij WJ, Dathan NA, Hamm J, Mattaj IW (1989) Identification of the RNA binding segment of human U1 A protein and definition of its binding site on U1 snRNA. Embo J 8:4163-70

Schultze, B., K. Wahn, H. D. Klenk and G. Herrler (1991). "Isolated HE-protein from hemagglutinating encephalomyelitis virus and bovine coronavirus has receptor-destroying and receptor-binding activity." Virology 180(1): 221-8.

Seybert, A., C. C. Posthuma, L. C. van Dinten, E. J. Snijder, A. E. Gorbalenya and J. Ziebuhr (2005). "A complex zinc finger controls the enzymatic activities of nidovirus helicases." J Virol 79(2): 696-704.

Shi, S. T., P. Huang, H. P. Li and M. M. Lai (2000). "Heterogeneous nuclear ribonucleoprotein A1 regulates RNA synthesis of a cytoplasmic virus." Embo J 19(17): 4701-11.

Sims, A. C., J. Ostermann and M. R. Denison (2000). "Mouse hepatitis virus replicase proteins associate with two distinct populations of intracellular membranes." J Virol 74(12): 5647-54.

Siomi H, Matunis MJ, Michael WM, Dreyfuss G (1993) The pre-mRNA binding K protein contains a novel evolutionarily conserved motif. Nucleic Acids Res 21:1193-8

Skehel, J.J., and Wiley, D.C. (1998). Coiled coils in both intracellular vesicle and viral membrane fusion. Cell 95, 871–874.

Skehel, J.J., and Wiley, D.C. (2000). Receptor binding and membrane fusion in virus entry: the influenza hemagglutinin. Annu. Rev. Biochem. 69, 531–569.

Snijder, E. J., P. J. Bredenbeek, J. C. Dobbe, V. Thiel, J. Ziebuhr, L. L. Poon, Y. Guan, M. Rozanov, W. J. Spaan and A. E. Gorbalenya (2003). "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(5): 991-1004.

Spaan, W., H. Delius, M. Skinner, J. Armstrong, P. Rottier, S. Smeekens, B. A. van der Zeijst and S. G. Siddell (1983). "Coronavirus mRNA synthesis involves fusion of non-contiguous sequences." Embo J 2(10): 1839-44.

Stadler, K., V. Masignani, M. Eickmann, S. Becker, S. Abrignani, H. D. Klenk and R. Rappuoli (2003). "SARS--beginning to understand a new virus." Nat Rev Microbiol 1(3): 209-18.

St Johnston D, Brown NH, Gall JG, Jantsch M (1992) A conserved double-stranded RNA-binding domain. Proc Natl Acad Sci U S A 89:10979-83

Stohlman, S. A., J. O. Fleming, C. D. Patton and M. M. Lai (1983). "Synthesis and subcellular localization of the murine coronavirus nucleocapsid protein." Virology 130(2): 527-32.

Stohlman, S. A., R. S. Baric, G. N. Nelson, L. H. Soe, L. M. Welter and R. J. Deans (1988). "Specific interaction between coronavirus leader RNA and nucleocapsid protein." J Virol 62(11): 4288-95.

Sturman, L. S., K. V. Holmes and J. Behnke (1980). "Isolation of coronavirus envelope glycoproteins and interaction with the viral nucleocapsid." J Virol 33(1): 449-62.

Supekar, V. M., C. Bruckmann, P. Ingallinella, E. Bianchi, A. Pessi and A. Carfi (2004). "Structure of a proteolytically resistant core from the severe acute respiratory syndrome coronavirus S2 fusion protein." Proc Natl Acad Sci U S A 101(52): 17958-63.

Surjit, M., Liu, B., Kumar, P., Chow, V.T., Lal, S.K (2004).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. 14;317(4):1030-6.

Sutton, G., E. Fry, L. Carter, S. Sainsbury, T. Walter, J. Nettleship, N. Berrow, R. Owens, R. Gilbert, A. Davidson, S. Siddell, L. L. Poon, J. Diprose, D. Alderton, M. Walsh, J. M. Grimes and D. I. Stuart (2004). "The nsp9 replicase protein of SARS-coronavirus, structure and functional insights." Structure 12(2): 341-53.

Tan R, Chen L, Buettner JA, Hudson D, Frankel AD (1993) RNA recognition by an isolated alpha helix. Cell 73:1031-40

Terwilliger, T. C. (2003). "SOLVE and RESOLVE: automated structure solution and density modification." Methods Enzymol 374: 22-37.

Theunissen O, Rudt F, Guddat U, Mentzel H, Pieler T (1992) RNA and DNA binding zinc fingers in Xenopus TFIIIA. Cell 71:679-90

Thiel, V., J. Herold, B. Schelle and S. G. Siddell (2001). "Viral replicase gene products suffice for coronavirus discontinuous transcription." J Virol 75(14): 6676-81.

Thiel, V., K. A. Ivanov, A. Putics, T. Hertzig, B. Schelle, S. Bayer, B. Weissbrich, E. J. Snijder, H. Rabenau, H. W. Doerr, A. E. Gorbalenya and J. Ziebuhr (2003). "Mechanisms and enzymes involved in SARS coronavirus genome expression." J Gen Virol 84(Pt 9): 2305-15.

Tresnan, D. B., R. Levis and K. V. Holmes (1996). "Feline aminopeptidase N serves as a receptor for feline, canine, porcine, and human coronaviruses in serogroup I." J Virol 70(12): 8669-74.

Tung, F. Y., S. Abraham, M. Sethna, S. L. Hung, P. Sethna, B. G. Hogue and D. A. Brian (1992). "The 9-kDa hydrophobic protein encoded at the 3' end of the porcine transmissible gastroenteritis coronavirus genome is membrane-associated." Virology 186(2): 676-83.

Valegard K, Murray JB, Stockley PG, Stonehouse NJ, Liljas L (1994) Crystal structure of an RNA bacteriophage coat protein-operator complex. Nature 371:623-6

Valegard, K., J. B. Murray, N. J. Stonehouse, S. van den Worm, P. G. Stockley and L. Liljas (1997). "The three-dimensional structures of two complexes between recombinant MS2 capsids and RNA operator fragments reveal sequence-specific protein-RNA interactions." J Mol Biol 270(5): 724-38.

van der Meer, Y., E. J. Snijder, J. C. Dobbe, S. Schleich, M. R. Denison, W. J. Spaan and J. K. Locker (1999). "Localization of mouse hepatitis virus nonstructural proteins and RNA synthesis indicates a role for late endosomes in viral replication." J Virol 73(9): 7641-57.

Vennema, H., G. J. Godeke, J. W. Rossen, W. F. Voorhout, M. C. Horzinek, D. J. Opstelten and P. J. Rottier (1996). "Nucleocapsid-independent assembly of coronavirus-like particles by co-expression of viral envelope protein genes." Embo J 15(8): 2020-8.

Verma, S., V. Bednar, A. Blount and B. G. Hogue (2006). "Identification of functionally important negatively charged residues in the carboxy end of mouse hepatitis coronavirus A59 nucleocapsid protein." J Virol 80(9): 4344-55.

Vlasak, R., W. Luytjes, J. Leider, W. Spaan and P. Palese (1988). "The E3 protein of bovine coronavirus is a receptor-destroying enzyme with acetylesterase activity." J Virol 62(12): 4686-90.

von Grotthuss, M., L. S. Wyrwicz and L. Rychlewski (2003). "mRNA cap-1 methyltransferase in the SARS genome." Cell 113(6): 701-2.

Weissenhorn, W., Dessen, A., Harrison, S.C., Skehel, J.J., and Wiley, D.C. (1997). Atomic structure of the ectodomain from HIV-1 gp41. Nature 387, 426–430.

Wu, H. Y., A. Ozdarendeli and D. A. Brian (2006). "Bovine coronavirus 5'-proximal genomic acceptor hotspot for discontinuous transcription is 65 nucleotides wide." J Virol 80(5): 2183-93.

Wurm, T., H. Chen, T. Hodgson, P. Britton, G. Brooks and J. A. Hiscox (2001). "Localization to the nucleolus is a common feature of coronavirus nucleoproteins, and the protein may disrupt host cell division." J Virol 75(19): 9345-56.

Wurzer, W. J., K. Obojes and R. Vlasak (2002). "The sialate-4-O-acetylesterases of coronaviruses related to mouse hepatitis virus: a proposal to reorganize group 2 Coronaviridae." J Gen Virol 83(Pt 2): 395-402.

Xu, Y., Y. Liu, Z. Lou, L. Qin, X. Li, Z. Bai, H. Pang, P. Tien, G. F. Gao and Z. Rao (2004). "Structural basis for coronavirus-mediated membrane fusion. Crystal structure of mouse hepatitis virus spike protein fusion core." J Biol Chem 279(29): 30514-22.

Xu, T., A. Ooi, H. C. Lee, R. Wilmouth, D. X. Liu and J. Lescar (2005). "Structure of the SARS coronavirus main proteinase as an active C2 crystallographic dimer." Acta Crystallograph Sect F Struct Biol Cryst Commun 61(Pt 11): 964-6.

Yamada, Y. K., M. Yabe, T. Ohtsuki and F. Taguchi (2000). "Unique N-linked glycosylation of murine coronavirus MHV-2 membrane protein at the conserved O-linked glycosylation site." Virus Res 66(2): 149-54.

Yang, H., M. Yang, Y. Ding, Y. Liu, Z. Lou, Z. Zhou, L. Sun, L. Mo, S. Ye, H. Pang, G. F. Gao, K. Anand, M. Bartlam, R. Hilgenfeld and Z. Rao (2003). "The crystal structures of severe acute respiratory syndrome virus main protease and its complex with an inhibitor." Proc Natl Acad Sci U S A 100(23): 13190-5.

Yeager, C. L., R. A. Ashmun, R. K. Williams, C. B. Cardellichio, L. H. Shapiro, A. T. Look and K. V. Holmes (1992). "Human aminopeptidase N is a receptor for human coronavirus 229E." Nature 357(6377): 420-2.

Yin, H.S., Paterson, R.G., Wen, X, Lamb, R.A., Jardetzky, T.S (2005). “Structure of the uncleaved ectodomain of the paramyxovirus (hPIV3) fusion protein”. Proc Natl Acad Sci U S A. 102(26):9288-93.

Yokomori, K., L. R. Banner and M. M. Lai (1991). "Heterogeneity of gene expression of the hemagglutinin-esterase (HE) protein of murine coronaviruses." Virology 183(2): 647-57.

Yu, X., W. Bi, S. R. Weiss and J. L. Leibowitz (1994). "Mouse hepatitis virus gene 5b protein is a new virion envelope protein." Virology 202(2): 1018-23.

Yu, I. M., C. L. Gustafson, J. Diao, J. W. Burgner, 2nd, Z. Li, J. Zhang and J. Chen (2005). "Recombinant severe acute respiratory syndrome (SARS) coronavirus nucleocapsid protein forms a dimer through its C-terminal domain." J Biol Chem 280(24): 23280-6.

Yuan, Q., Y. Liao, J. Torres, J. P. Tam and D. X. Liu (2006). "Biochemical evidence for the presence of mixed membrane topologies of the severe acute respiratory syndrome coronavirus envelope protein expressed in mammalian cells." FEBS Lett 580(13): 3192-200.

Zhai, Y., F. Sun, X. Li, H. Pang, X. Xu, M. Bartlam and Z. Rao (2005). "Insights into SARS-CoV transcription and replication from the structure of the nsp7-nsp8 hexadecamer." Nat Struct Mol Biol 12(11): 980-6.

Zhou, M., A. K. Williams, S. I. Chung, L. Wang and E. W. Collisson (1996). "The infectious bronchitis virus nucleocapsid protein binds RNA sequences in the 3' terminus of the genome." Virology 217(1): 191-9.

Zhou, M. and E. W. Collisson (2000). "The amino and carboxyl domains of the infectious bronchitis virus nucleocapsid protein interact with 3' genomic RNA." Virus Res 67(1): 31-9.

Ziebuhr, J., E. J. Snijder and A. E. Gorbalenya (2000). "Virus-encoded proteinases and proteolytic processing in the Nidovirales." J Gen Virol 81(Pt 4): 853-79.

Ziebuhr, J., V. Thiel and A. E. Gorbalenya (2001). "The autocatalytic release of a putative RNA virus transcription factor from its polyprotein precursor involves two paralogous papain-like proteases that cleave the same peptide bond." J Biol Chem 276(35): 33220-32.

Zuniga, S., Sola, I., Moreno, J.L., Sabella, P., Plana-Duran, J., Enjuanes, L. (2007) Coronavirus nucleocapsid protein is an RNA chaperone. Virology 357:215-27