The E Protein Is a Multifunctional Membrane Protein of SARS-CoV

Qingfa Wu[1,2]*, Yilin Zhang[1]*, Hong Lü[1]*, Jing Wang[3,1]*, Ximiao He[1], Yong Liu[4], Chen Ye[1], Wei Lin[1], Jianfei Hu[1,3], Jia Ji[1], Jing Xu[1], Jia Ye[1,2], Yongwu Hu[1], Wenjun Chen[1], Songgang Li[1,3], Jun Wang[1], Jian Wang[1,2], Shengli Bi[5], and Huanming Yang[1,2]#

[1] Beijing Genomics Institute, Chinese Academy of Sciences, Beijing 101300, China; [2] James D. Watson Institute of Genome Sciences, Zhijiang Campus, Zhejiang University, Hangzhou 310008, China; [3] College of Life Sciences, Peking University, Beijing 100871, China; [4] Peking Union Medical College Hospital, Chinese Academy of Medical Sciences, Beijing 100730, China; [5] Center of Disease Control and Prevention, Beijing 100050, China.

The E (envelope) protein is the smallest structural protein in all coronaviruses and is the only viral structural protein in which no variation has been detected. We conducted genome sequencing and phylogenetic analyses of SARS-CoV. Based on genome sequencing, we predicted the E protein is a transmembrane (TM) protein characterized by a TM region with strong hydrophobicity and α-helix conformation. We identified a segment (NH_2-_L-Cys-A-Y-Cys-Cys-N_-COOH) in the carboxyl-terminal region of the E protein that appears to form three disulfide bonds with another segment of corresponding cysteines in the carboxyl-terminus of the S (spike) protein. These bonds point to a possible structural association between the E and S proteins. Our phylogenetic analyses of the E protein sequences in all published coronaviruses place SARS-CoV in an independent group in *Coronaviridae* and suggest a non-human animal origin.

Key words: SARS, SARS-CoV, the E protein, envelope, TM region

Introduction

The coronaviruses are a group of enveloped viruses iruses. The putative membranous envelopes have a mosaic structure. This structure is composed of a lipid bilayer membrane that is derived from the endoplasmic reticulum (ER) and Golgi complex of the host cell and viral gene-encoded proteins (*1*).

As a small structural protein, the E (envelope) protein is so-named because it has generally been regarded as the main component of the viral envelope since its first identification in RNA viruses. In addition to the pivotal role that it purportedly plays in the assembly of the viral envelope and/or the host-derived membrane, there is accumulating evidence from research on known coronaviruses that the expression of the E protein also results in the production and release of membrane vesicles or virus-like particles (VLPs) (*2*, *3*, *4*), induction of apoptosis (*3*), and synthesis of α-interferon (*5*). Its involvement in RNA replication has also been reported (*6*, *7*). Induced mutation or recombination of the E protein may result in lethal or temperature-sensitive phenotypes and aberrant morphology (*8*).

Herein we examine the role of the E protein as a multifunctional membrane protein in SARS-CoV. We conducted comparative and phylogenetic analyses of the structure and function of the E protein in sixteen genome sequences of SARS-CoV published by Beijing Genomics Institute (BGI; ref. *9*) and other laboratories (*10-12*), and in genome sequences of all other members of *Coronaviridae* published in Genbank.

Results

Identification of the transmembrane region in the E protein

We identified a characteristic transmembrane (TM) region at the residue position 15-37 (Figure 1). This TM region is composed of 23 amino acids, occupying 30% of the total size of the E protein. The predicted

*** These authors contributed equally to this work.**

 # Corresponding author.

E-mail: yanghm@genomics.org.cn

TM region is strongly hydrophobic due to the abundance of Leu and Val residues. Computational results of three software programs (see Materials and Methods) independently supported the existence of the TM region and also yielded consistent predictions of the higher (secondary) structure of the TM region, indicating an α-helix conformation in the uncharged and highly-hydrophobic subregion of the E protein (Figure 2).

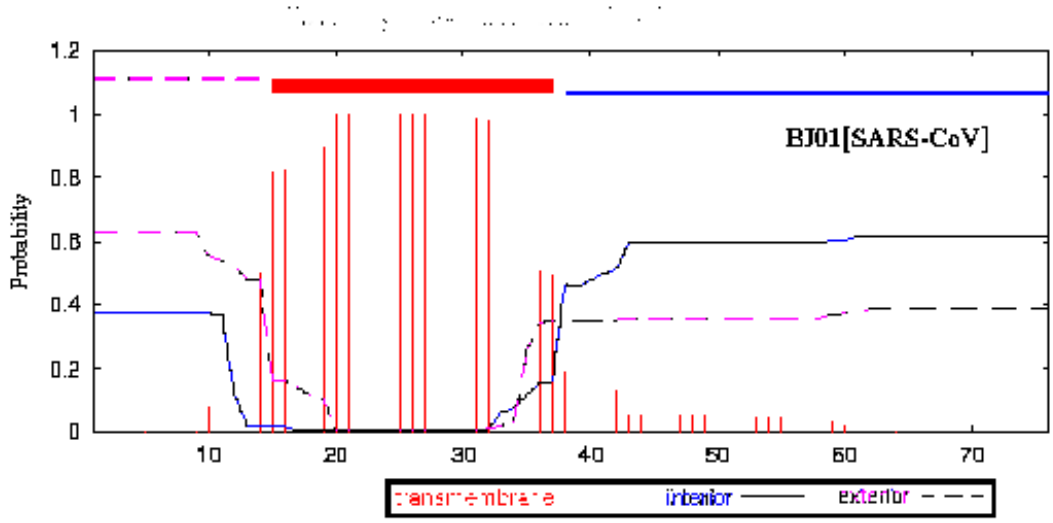

Fig. 1 The predicted TM region in the E protein of SARS-CoV by TMHMM.

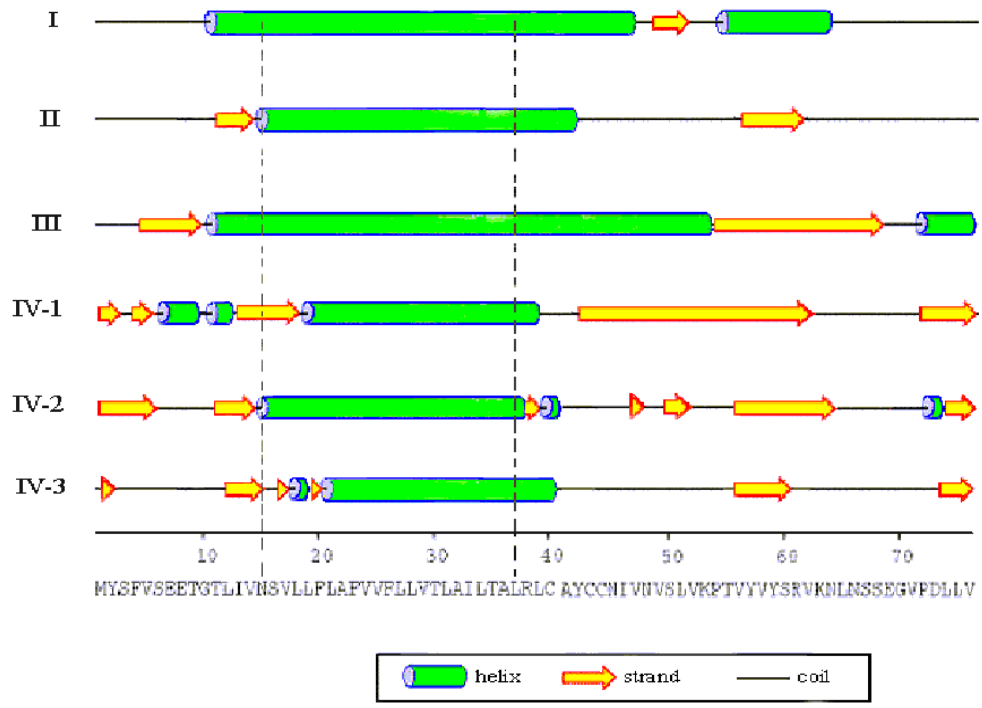

Fig. 2 Predicted secondary structures in the E protein of SARS-CoV. Software programs: I. PSIpred; II. NNPredict; III. SPLIT; IV. 1-3. Antheprot 5.0.

Our region prediction analyses using TMHMM further demonstrated that the TM region divides the E protein into three subregions and also showed that the N-terminus is probably located in the exterior of the virion. Our findings indicate that the N-terminus is composed of approximately fourteen amino acids. The N-terminus is also negatively charged and hydrophilic with a low subregional pI of 3.79. Our predictions show that the C-terminus is in the interior of the virion and has approximately 39 amino acids with a relatively higher pI of 8.61 (Table 1).

We repeated these analyses with twelve other coronaviruses that were published in GenBank (Figure 3-I, 3-II and 3-III). These analyses demonstrated that the TM region can be found in the E protein of all members of *Coronaviridae,* in spite of their low homology in the primary sequences. We also observed that several TM regions have a reversed orientation and that some coronaviruses have more than one TM region.

Table 1 The Genomic and Biochemical Features of the Entire E Protein and Its Three Subregions

	TM region	N-terminus	C-terminus	E protein
G+C (%)	42.0	35.7	40.8	40.2
A	11 (15.9%)	13 (31.0%)	30 (25.0%)	54 (23.4%)
U	29 (42.1%)	14 (33.3%)	41 (34.2%)	84 (36.4%)
C	18 (26.1%)	6 (14.3%)	23 (19.2%)	47 (20.3%)
G	11 (15.9%)	9 (21.4%)	26 (21.6%)	46 (19.9%)
Total (nt)	69 (100%)	42 (100%)	120 (100%)	231 (100%)
Leu	8 (10.6)	1 (1.3)	5 (6.6)	14 (18.4)
Val	4 (5.3)	2 (2.6)	8 (10.6)	14 (18.4)
Phe	3 (3.9)	1 (1.3)	0	4 (5.3)
Ala	3 (3.9)	0	1 (1.3)	4 (5.3)
Thr	2 (2.7)	2 (2.6)	1 (1.3)	5 (6.6)
Ile	1 (1.3)	1 (1.3)	1 (1.3)	3 (3.9)
Asn	1 (1.3)	0	4 (5.3)	5 (6.6)
Ser	1 (1.3)	2 (2.6)	4 (5.3)	7 (9.2)
Glu	0	2 (2.6)	1 (1.3)	3 (3.9)
Tyr	0	1 (1.3)	3 (3.9)	4 (5.3)
Gly	0	1 (1.3)	1 (1.3)	2 (2.6)
Met	0	1(1.3)	0	1(1.3)
Cys	0	0	3 (3.9)	3 (3.9)
Lys	0	0	2 (2.6)	2 (2.6)
Arg	0	0	2 (2.6)	2 (2.6)
Pro	0	0	2 (2.6)	2 (2.6)
Asp	0	0	1 (1.3)	1 (1.3)
Total	23 (30.3)	14 (18.4)	39 (51.3)	76 (100)
Molecular Weight (a.a.)	2491	1576	4330	8361
pI	5.52	3.79	8.61	6.01
Net Charge	0	-2	+2	0
		(-2.7%)	(-2.7%, +5.4%)	(-5.4%, +5.4%)

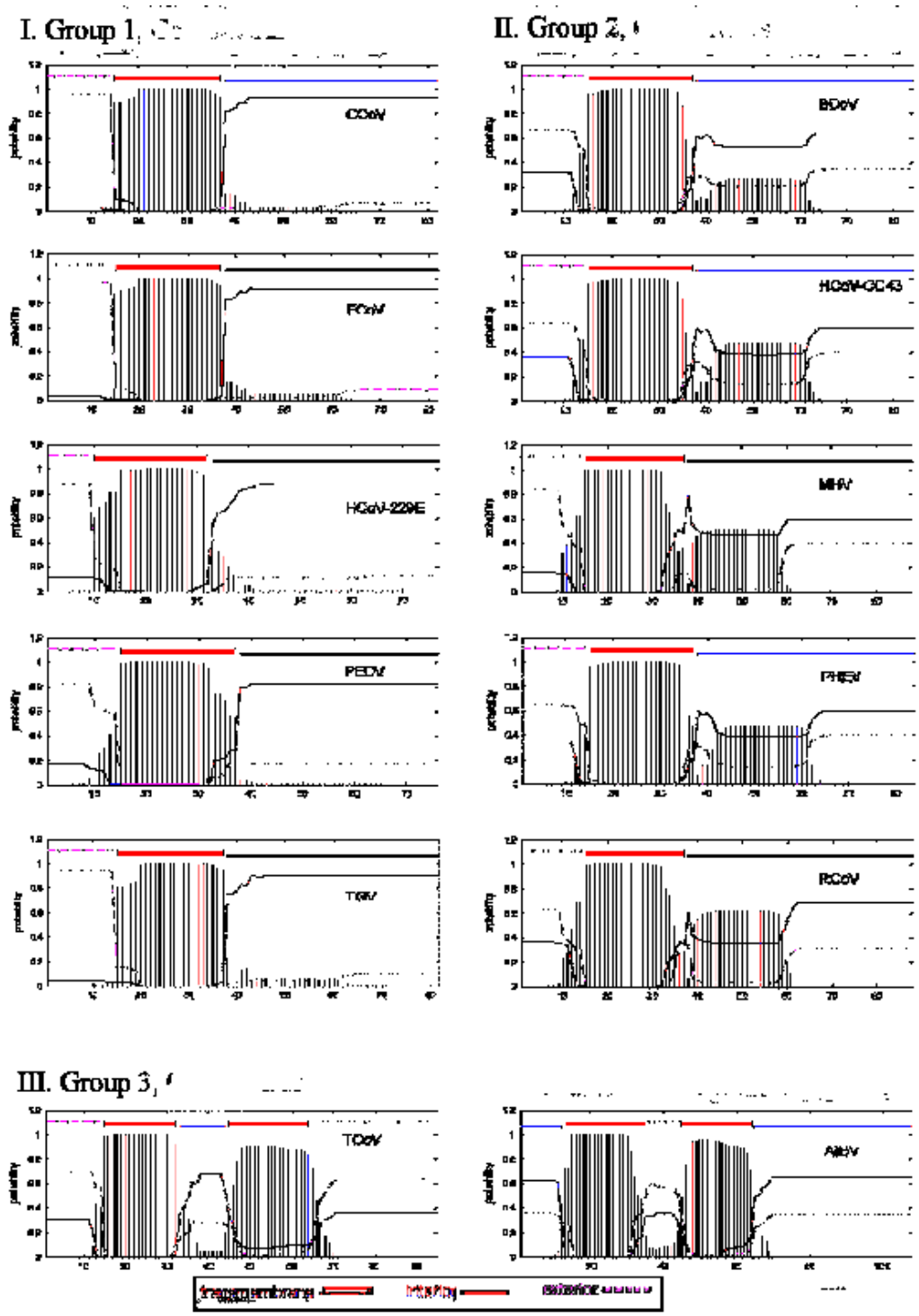

Fig. 3-I, 3-II, 3-III Predicted TM regions in the E protein of the three groups of coronaviruses.

Prediction of disulfide bonds between the E and S proteins

We conducted a subregional analysis and observed a segment containing a motif with three cysteines in the interior of the virion of the E protein. This motif is located directly after the TM region of the E protein and contains three cysteines in the order of (NH_2-_L-Cys-A-Y-Cys-Cys-N_-COOH). A corresponding segment (NH_2-_S-Cys-G-S-Cys-Cys-K_-COOH) was also found in the carboxyl inner-virion terminus of the S (spike) protein. The three cysteines that are present in both segments may form three disulfide bonds between the E and S proteins, provided that they have

the appropriate orientation and other structural features.

Other sequence and structural features of the E protein

We used ClustalW to compare the ORFs (open reading frames) for the sixteen SARS-CoV genome sequences published in GenBank. Our analyses showed that all SARS-CoV sequences have the same E protein, although its position in the genome may differ. The ORF for the E protein is 231 nucleotides (nt) in size, accounting for only 0.78% of the whole viral genome and is located at nt position from 26,098 to 26,328, between PUP2 (putative uncharacterized protein 2; nt position 25,670-26,134) and the ORF for the M (membrane) protein (nt position 26,379-27,044; ref. *9*). The E protein has a GC content of 40.2% (A: U: C: G = 54: 84: 47: 46), which is close to the average of the whole genome (40.8%; Table 1, Figure 4-I).

The E protein is believed to be the smallest protein in the viral proteome, encoding a functional protein of 76 amino acids. Two non-polar neutral amino acids (Val and Leu) constitute a substantial portion (28/76, 36.8%) of the E Protein, and contribute to its strongest hydrophobicity (47.40%) among all the viral structural proteins. The E protein also has zero net charge over the whole peptide (5.40% for both positive and negative charges) (Table 1). Using SignalP, we predicted a cleavage site of a signal peptide (AYC-CN) at the N-terminus of the E protein that is most likely located at residue position 43-44. We also examined this signal peptide in the E protein of other coronaviruses, and a similar topology was observed.

Using a single E protein, we analyzed the distributions of predicted GC content using DNA_GC_Calculator, subregional charges using EMBOSS, and hydrophobicity using Anthreprot 5.0. As shown in Figure 4, our results showed the E protein might be divided into three regions. The distribution of GC content of the ORF (Figure 4-I) reveals a GC-rich region in the middle of the E protein that is flanked by two relatively GC-poor regions. A similar topology is seen for the distribution of charges (Figure 4-II). The middle region of the E protein is uncharged and is flanked by a small, negatively charged region at the N-terminus and a region of variable charges at the C-terminus. Regarding hydrophobicity, we observed the highest hydrophobicity in the GC-rich and uncharged middle region of the E protein.

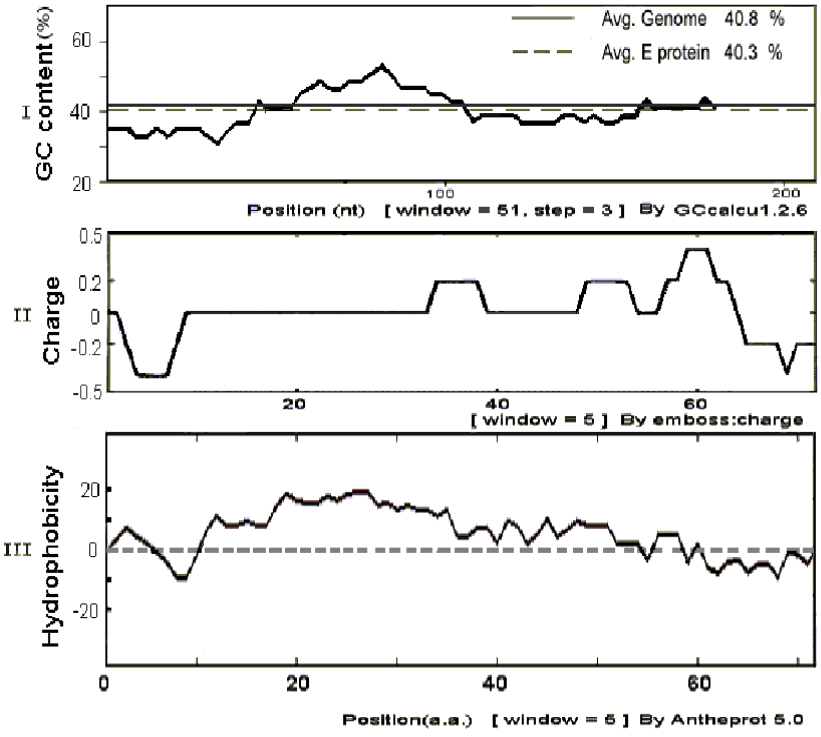

Fig. 4-I, 4-II, 4-III The distribution of GC content (I), charge (II) and hydrophobicity (III) in the E protein of SARS-CoV.

We identified a mutant element upstream of the E protein CDS (coding sequence). This mutation differs by one base pair from the presumed core leader sequence UCUAAAC that is located near the beginning of all other CDSs.

The GC-rich region and the predicted TM region are both located in the same subregion of the E protein. As shown in Figure 4-I and Table 2, the preferential codon usage by the TM region results in a GC-rich segment in the E protein.

Table 2 Codon Usage Frequency of the TM Region and the Entire E Protein

a.a.	TM region	E protein	Codon	TM region	E protein
Ala	3 (4.00%)	4 (5.33%)	GCA	0	0
			GCC	1 (1.33%)	1 (1.33%)
			GCG	1 (1.33%)	2 (2.67%)
			GCU	1 (1.33%)	1 (1.33%)
Cys	0	3 (4.00%)	UGC	0	2 (2.67%)
			UGU	0	1 (1.33%)
Asp	0	1 (1.33%)	GAC	0	0
			GAU	0	1 (1.33%)
Glu	0	3 (4.00%)	GAA	0	3 (4.00%)
			GAG	0	0
Phe	3 (4.00%)	4 (5.33%)	UUC	2 (2.67%)	3 (4.00%)
			UUU	1 (1.33%)	1 (1.33%)
Gly	0	2 (2.67%)	GGA	0	1 (1.33%)
			GGC	0	0
			GGG	0	0
			GGU	0	1 (1.33%)
His	0	0	CAC	0	0
			CAU	0	0
Ile	1 (1.33%)	3 (4.00%)	AUA	0	1 (1.33%)
			AUC	1 (1.33%)	1 (1.33%)
			AUU	0	1 (1.33%)
Lys	0	2 (2.67%)	AAA	0	2 (2.67%)
			AAG	0	0
Leu	8 (10.67%)	14 (18.67%)	CUA	2 (2.67%)	2 (2.67%)
			CUC	0	0
			CUG	0	2 (2.67%)
			CUU	5 (6.67%)	6 (8.00%)
			UUA	0	2 (2.67%)
			UUG	1 (1.33%)	2 (2.67%)
Met	0	1 (1.33%)	AUG	0	1 (1.33%)
Asn	1 (1.33%)	5 (6.67%)	AAC	0	2 (2.67%)
			AAU	1 (1.33%)	3 (4.00%)

Table 2 (*Continued*)

Pro	0	2 (2.67%)	CCA	0	1 (1.33%)
			CCC	0	0
			CCG	0	0
			CCU	0	1 (1.33%)
Gln	0	0	CAA	0	0
			CAG	0	0
Arg	0	2 (2.67%)	AGA	0	0
			AGG	0	0
			CGA	0	1 (1.33%)
			CGC	0	0
			CGG	0	0
			CGU	0	1 (1.33%)
Ser	1 (1.33%)	7 (9.33%)	AGC	1 (1.33%)	1 (1.33%)
			AGU	0	1 (1.33%)
			UCA	0	1 (1.33%)
			UCC	0	0
			UCG	0	2 (2.67%)
			UCU	0	2 (2.67%)
Thr	2 (2.67%)	5 (6.67%)	ACA	1 (1.33%)	2 (2.67%)
			ACC	0	0
			ACG	0	2 (2.67%)
			ACU	1 (1.33%)	1 (1.33%)
Val	4 (5.33%)	14 (18.67%)	GUA	2 (2.67%)	3 (4.00%)
			GUC	1 (1.33%)	3 (4.00%)
			GUG	1 (1.33%)	2 (2.67%)
			GUU	0	6 (8.00%)
Trp	0	0	UGG	0	0
Tyr	0	4 (5.33%)	UAC	0	4 (5.33%)
			UAU	0	0
STOP	0	1 (1.33%)	UAA	0	1 (1.33%)
			UAG	0	0
			UGA	0	0

Phylogenetic analysis of the E protein

Even with parameters of the lowest stringency, our Blast searches of nucleotide and amino acid sequences failed to detect any similarity between the E protein and any known sequences in GenBank, other than the sixteen recently published isolates of SARS-CoV.

In our comparative analysis of the E protein of SARS-CoV and other members in *Coronaviridae,* we did not observe any large homologous region (Figure 5). With regard to overall sequence similarity, the SARS-CoV E protein has the highest similarity to TGV (transmissible gastroenteritis virus; 40/82, 48.8%) and the lowest with AIBV (avian infectious bronchitis virus; 34/109, 31.2%; Figure 6). These findings are consistent with the phylogenetic tree that we proposed based on the amino acid sequence of the E protein (Figure 7-I and 7-II). No variation has yet been found in the published sequences.

Previous analyses of our group suggest that the E protein is most likely to be a mem-

ber of the well-characterized NS3_EnvE protein family (*11*), which is a family of small non-structural proteins well conserved among coronavirus (http://www.sanger.ac.uk/Software/Pfam/search.shtml), based on the reasonable similarity that was found between them (identity=21/76; similarity=13/76).

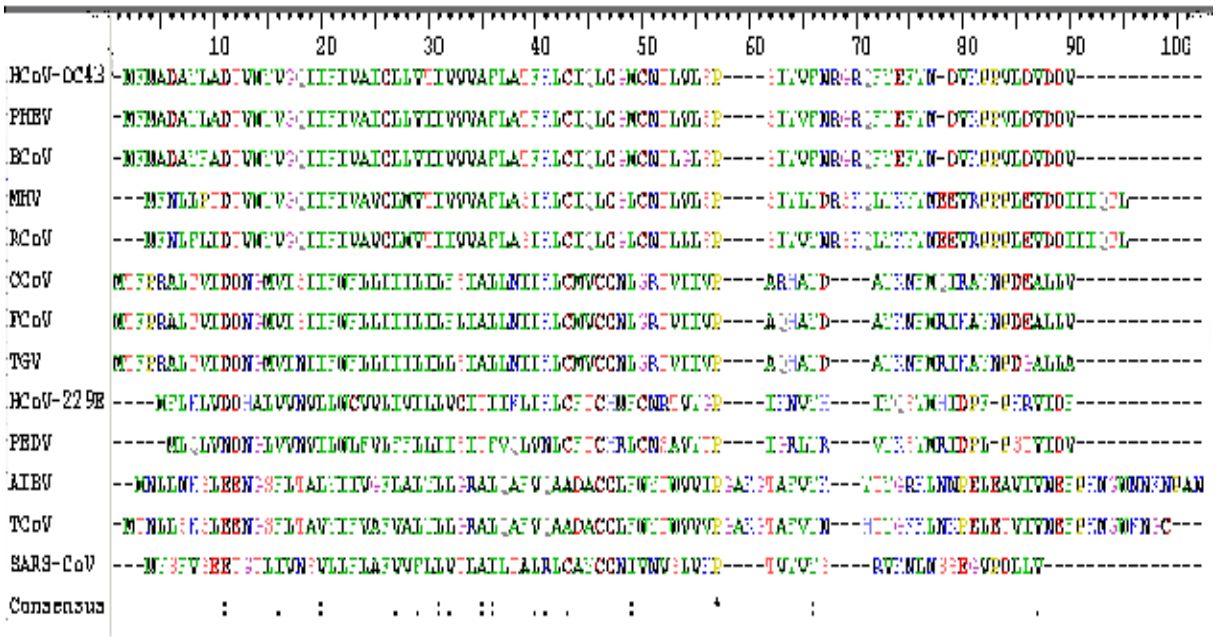

Fig. 5 Multi-alignment of amino acid sequences of the E protein between SARS-CoV and other coronaviruses. AIBV: avian infectious bronchitis virus; BCoV: bovine coronavirus; CcoV: canine coronavirus; FCoV: feline coronavirus; HCoV-229E: human coronavirus 229E; HCoV-OC43: human coronavirus OC43; MHV: murine hepatitis virus; PEDV: porcine epidemic diarrhea virus; PHEV: porcine hemagglutinating encephalomyelitis virus; RCoV: rat coronavirus; TCoV: turkey coronavirus; TGV: transmissible gastroenteritis virus.

	BJ01												
BJ01	100	TGV											
TGV	48.8	100	CCoV										
CCoV	47.6	96.3	100	FCoV									
FCoV	47.6	95.1	98.8	100	HCoV-E								
HCoV-E	47.4	56.1	54.9	54.9	100	PEDV							
PEDV	43.6	52.9	52.9	51.2	71.9	100	HCoV-O						
HCoV-O	46.4	42.5	44.8	44.8	39.3	39.3	100	PHEV					
PHEV	46.4	43.7	46.0	46.0	40.5	40.5	98.8	100	BCoV				
BCoV	45.2	41.9	43.7	43.7	38.1	38.1	96.4	97.6	100	MHV			
MHV	44.3	41.8	46.2	44.0	40.9	44.3	82.2	81.1	80.0	100	RCov		
RCov	44.3	44.0	47.3	46.2	42.0	44.3	84.4	83.3	81.1	96.6	100	TCoV	
TCoV	42.3	38.5	41.7	41.7	33.7	34.4	36.0	35.0	35.1	37.6	36.6	100	AIBV
AIBV	31.2	31.8	34.5	34.5	32.4	28.7	31.5	32.4	30.6	31.8	30.3	78.9	100

Fig. 6 Pairwise similarity between coronaviruses based on the E protein. The numbers indicate the percentage of similarity between each listed coronavirus and BJ01. (HCoV-E: HCoV-229E; HCoV-O: HCoV-OC43.)

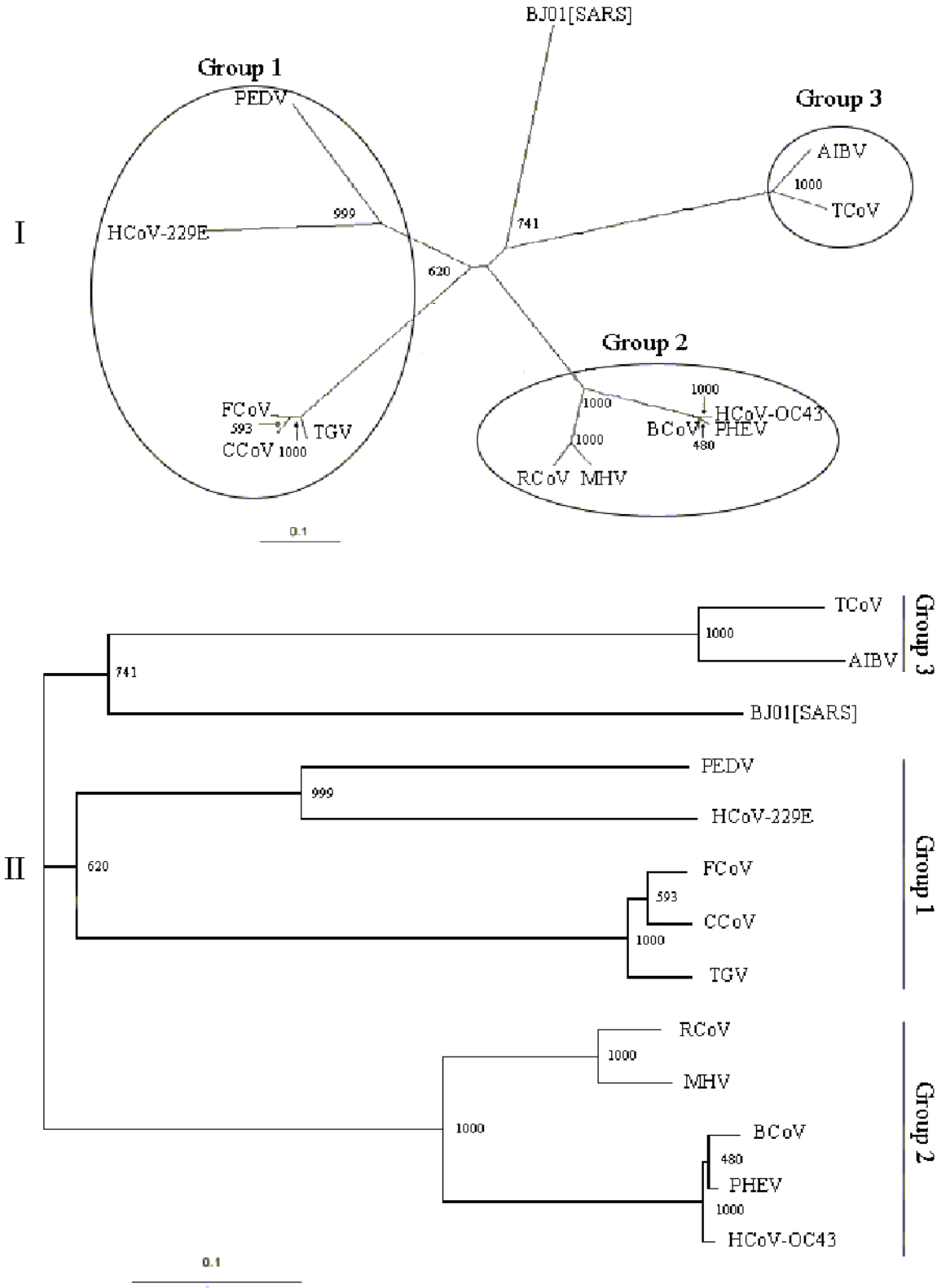

Fig. 7 An unrooted phylogenetic tree of the coronaviruses based on the amino acid sequence of the E protein by ClustalW and TreeView.

Discussion

Understanding the pathogenesis of SARS and the properties of SARS-CoV is of great significance for public health. A crucial step toward this goal is to increase our knowledge of the origin, components, structure, and underlying functions of the viral envelope, from which the E protein derives its name. In addition, more information is needed about the relationship of the E protein to the M protein of the host-derived membrane and the N protein (nucleoprotein) of the viral capsid.

The coronavirus is generally understood to be a virus-derived nucleocapsid, composed of RNA and the N protein, and wrapped by a host-derived lipid bilayer membrane that is made up of the viral S and M proteins, perhaps also the N protein (*3, 8, 11, 13*). Our results demonstrate that the E protein is another component protein of the host-derived membrane and has a potential structural link with the S protein through predicted disulfide bonds.

Combined evidence for the TM region in the E protein

Our conclusions that the E protein in SARS-CoV genome contains a TM region and is one of the major structural components of the host-derived lipid bilayer membrane are well supported by evidence from two perspectives.

First, the size of the predicted TM region, which spans 20-24 residues (±2 for the marginal residues of each predicted boundary), is consistent with the size of common membrane domains.

Secondly, bioinformatics analyses of the nucleotide/amino acid sequence and its related physiochemical features of the region demonstrate that the TM region is present in the E protein. This predicted TM region is strongly hydrophobic and characterized by two predominant non-polar and uncharged residues, Leu and Val (13/23). Evidence derived from the primary structure, such as a near-zero net charge (Figure 4-II) and a relatively higher pI than either of the two flanking regions (Table 1), provides further confirmation of the TM region in the E protein. The GC-rich region and the predicted TM region are both located in the same place of the E protein, a result of the preferential codon usage by the TM region. A mutant element was identified upstream of the E protein CDS instead of the presumed core leader sequence UCUAAAC, which is located near the beginning of all other CDSs. This mutant element was previously reported by us (UCUACAC at -200 nt upstream of the E protein CDS; ref. *9*) and others (UACGAAC at -2 nt upstream of the start codon; ref. *11*). This mutant element is homologous to the core leader sequence. However, Previous experimental evidence suggested that a point mutation presumably located within this region increased the translation efficiency of RNA transcripts from the E protein (*3*). The role of the leader sequence of the E protein in its transcription regulation still awaits further experimental proof.

Thirdly, prediction of secondary structure is also an important parameter for bioinfomatics determination of the TM region. Our results consistently demonstrated the presence of an α-helical conformation of the postulated TM region. The only differences yielded by various bioinformatics programs were the boundary determinations, which are acceptable in region prediction analysis. The α-helical conformation fits in the Type II TM region (*14*).

Serious attention has been paid to possible mistakes by the prediction software. We previously observed errors when analyses were performed on the M protein (see the article about the M protein in this issue). We therefore evaluated different software programs to see whether alteration of a single amino acid will change the orientation of the predicted TM region. Our results showed that this change takes place. In order to avoid or to minimize the possible prediction errors, two strategies were adopted: first, we applied different analysis software with different functional features and parameters to the analysis of the SARS-CoV E protein (e.g. TM helix: 11-35, predicted by Hmmtop, TM helix from outside to inside: 17-34, predicted by Tmpred); secondly, we used the same software previously used for the analysis of the E protein in other known members of *Coronaviridae* (Figure 3) .

Different analytic methods have not offered any evidence that the TM region is an artifact, whether by differences in boundary determination, possible orientation judged by various relevant parameters, or other physical and/or chemical features. This conclusion was reached after repeated analysis, comparison, and conservative interpretation of our results. Furthermore, analyses of the E protein in all the other members of *Coronaviridae* also support the existence of the TM region. An extra TM region has been detected in a few members in Group II, but the existence of such a second TM region in the SARS-CoV E protein has

been firmly excluded.

We propose that the orientation of the predicted TM region would bring the N-terminus in the exterior of the virion. It is consistent with the data previously reported in MHV and TGV (*8*, *14*) and with our analysis result of the M protein (see the article about the M protein in this issue). The C-terminus is postulated to be in the interior of the virion. The presence of postulated glycosylation sites in the C-terminal region of the E protein should be noticed. This phenomenon is also observed in HCoV-229E and RCoV, which share the same topology with the E protein of SARS-CoV and have been supported by experimental and bioinformatic evidence.

The orientation is crucial to the understanding of the structure, the function of the E protein and its relevance to other viral components, as well as the whole picture of the entire virus. Therefore, it is necessary to emphasize again that this conclusion, as well as others proposed in this study, should be regarded as a working model based on bioinformatics analysis of the sequence data and be interpreted in the right perspective. It will be tested, updated, revised, corrected, or even discarded to follow the constantly emerging new data. The conclusion should not be finalized until the experimental data are supportive.

Possible structural link between the E and S proteins

The E protein is postulated to have the structural link with the S protein through the predicted disulfide bonds, even awaiting the experimental data to confirm, providing a new way to think about the viral structure and function which are based on the interaction of its proteins, as well as its relationship to host cells.

The S protein is the largest structural protein in coronaviruses, and has been regarded as the major protein responsible for viral attachment and entry into the host cell (*16*), antigenicity, host range and tissue tropism, virulence and pathogenesis, in addition to the constitution of the characteristic spikes on the surface of the virion (*17-20*). The S protein is postulated to incorporate into the viral envelope with its TM region anchored into the host-derived membrane. It also interacts with the M protein in the release of mature virions from the smooth vesicles (*20*).

A motif characterized by three cysteines (NH_2-_L-Cys-A-Y-Cys-Cys-N_-COOH) has been discovered immediately following the TM region in the interior of the E protein. This motif is absolutely conserved in CCoV, FCoV and TGV (Figure 5) and has been suggested as palmitoylation (*8*). The distance between the first Cys in the motif and the last residue of the TM region toward interior of the virion is 1 amino acid in the E protein, and 12 amino acid while considering the corresponding motif (NH_2-_S-Cys-G-S-Cys-Cys-K_-COOH) in the most C-terminus of the S protein. Based on the predicted orientation, position and composition of this motif, we propose that the Cys residues in the motif of the E protein might provide sites to form possible disulfide bonds with the corresponding motif of the S protein (Figure 8), with reference to the well-known molecule model derived from insulin. Whether the E and S proteins are co-expressed from a poly-cistron or post-translationally modified in a way similar to that for insulin still need further approval.

The inter-molecular reactions between viral structural proteins have been reported (*5*, *21*). M-M homodimer is known to be a prerequisite for particle formation (*21*). The interaction between the E and M proteins has been proposed to be involved in formation and extracellular release of the viral particles and the induction of α-interferon (*3, 5, 8, 22, 23*). Several domains of the E and M polypeptide chains might be implicated in stereospecific interactions (*5*).

We have also explored the possible direct peptide-peptide interaction between the E and M proteins. We have detected three Cys dispersing in the whole M protein (Cys^{63}, Cys^{85}, and Cys^{158}; see the article about the M protein in this issue), but no evidence has been established for possible disulfide bonds with the E protein. The possible interpretation would be the spacing of Cys residues (two of the three are located in the TM region), even if it should not be regarded as prerequisite for disulfide bonds. The possible existence of the E protein dimer or polymer through formation of the disulfide bonds at the Cys sites should also be considered with reference to the coming proteomics data. The possibility of such a covalent interaction with the last known structural protein, the N protein, can be excluded because it is Cys-free, and that with other PUPs or NSPs (non-structural proteins) or proteolytic products derived from viral structural proteins, even not very likely, might be considered.

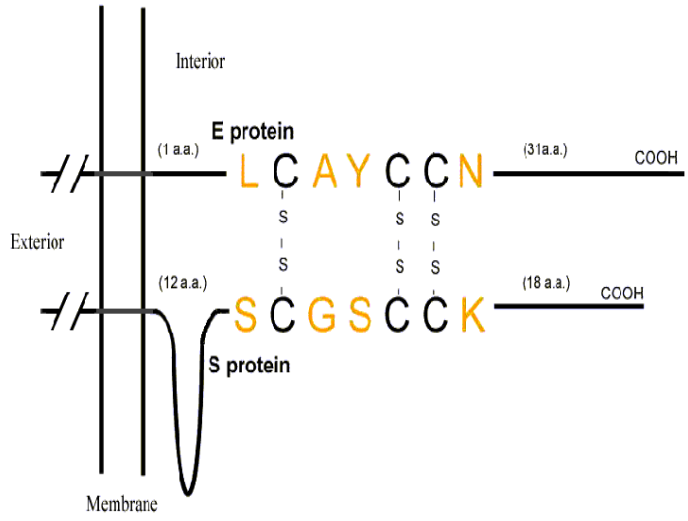

Fig. 8 The predicted disulfide bonds between the E and S proteins of SARS-CoV.

The direct or indirect, homo- or hetero-polymer interaction of viral proteins would be important to the functions and architecture of the virus. The biological significance of this discovery is to demonstrate the first structural link between the two viral structural proteins of SARS-CoV. Once it is confirmed, it might help the interpretation of the immunointeraction between the virus and its host, possible signal transduction pathways, and the involvement of the two proteins in pathogenesis and antigenicity.

Localization of other possible functions outside the TM region

Evidence for functions by the E protein, or by a monomer, or by interactions among the E monomers, or by cooperation with other structural viral proteins, has been accumulated (*8*). Even at low expression levels in the virion (*2*, *8*), the expression of the E protein in the previously known coronaviruses has been reported to be involved in replication (*6*, *7*), apoptosis (*3*), synthesis of α-interferon (*5*), lethal or temperature-sensitive phenotypes and aberrant morphology (*8*), and perhaps many other functions, demonstrating that the E protein is a multifunctional protein (*3*).

If so, all these functions of the E protein would be carried out by the short (approximately 14 residues) N-terminal region in the exterior and the 39-residue C-terminal region in the interior of the virion, since the TM region accounts for about one third of the E protein (23/76), suggesting that the other two thirds would be responsible for all the other possible functions of the E protein. It has been reported that the mutations introduced into the C-terminal region of the E protein in MHV yielded thermolabile viruses (*8*).

The predicted signal peptides at the N-terminal region of the E protein, with a most likely cleavage site (AYC-CN) at residue position 43-44, would be functional if it could be released by either a viral endogenous or cellular mechanism. The supportive evidence is from its presence in the E protein and similar topology in other coronaviruses. But the E protein in MHV was found to integrate in membrane without involvement of a cleaved signal peptide (*23*). Proteomics data are urgently needed for any predicted function of the E protein.

Recent non-human origin hypotheses of the E protein

One of the discoveries by sequence analysis is that the E protein of coronaviruses is so unique that it is divergent from any other known sequences including the E protein of other enveloped (or membrane) viruses other than *Coronaviridae*.

Even many hypotheses have been proposed, the origin of the SARS-CoV remains a mystery. If the evolution clock model that mutation frequency is proportional to evolution rate were taken, it would have taken millions of years for an ancestral E protein to evolve into this type, putting the selection pressure on coronaviruses aside for the simplest estimation, ei-

ther positive (host range, tissue tropism, replication rate, etc.) or negative (replication errors leading to abortion, the too high fatality of hosts, etc.).

Based on the sequence data, especially the uniqueness of the E protein, we postulate that the E protein would have been evolved independently from any other viral proteins in coronaviruses and from a non-human host. It would not be originated by accumulation of point mutation from an ancestor, neither by a single or a few recombination events with parts of any other viral genomes, nor by shuffling (insertion/deletion and segmental duplication) of reasonably large segments by recent horizontal transfer. It would have been existing in a non-human animal for a long time, probably latent, and getting into humans through its recent contact or established a new link with humans.

Some evidence would support our hypothesis. Firstly, no antibody against the virus has been detected from normal individuals, suggesting that it might not be a latent virus in humans(11). Secondly, the significant similarity in its secondary structure and the similar function demonstrated in all E proteins of coronaviruses, in spite of the divergence of the primary structure, suggesting that it has experienced the selection in nature for a long time. Thirdly, relatively rare variations were detected by comparison of all the published sequences of SARS-CoV. No sequence variation has yet been reported. Even the size of the E protein is only approximately 0.2% (228/29725) of the genome, the rare variation of the E protein still could not be satisfactorily interpreted by simple size ratio because of the high mutation rata reported. In other coronaviruses, the E protein has always been found to be entire, while various types of mutation have been detected in other regions (*8*). The relative stability of the E protein in evolution or replication in either host tissues or cell cultures would mean its important function besides its evolutionary significance.

As the smallest known structural protein in the tiny SARS-CoV genome, the E protein leaves the biggest mystery to explore.

Methods and Materials

Sources of samples and sequences

The procedures for collecting SARS-CoV samples from patients and preparing RNA for sequencing have been described previously (*9*). We performed sequencing using MegaBACE1000 (Amersham, New Jersey, USA). Base calling, contamination removal and assembly were performed by Phred, CrossMatch and Phrap (http://www.phrap.org), respectively. The sequences of the two complete and four draft genomes assembled by BGI have been deposited in GenBank (accession numbers: AY278488, AY279354, AY278490, AY278489, AY278487, and AY297028.1) are available freely. All the experimental materials, including the cDNA clones representing various segments of the viral genome with known sequences, are available for collaborators (see Supplementary Database: http://www.genomics.org.cn/SARS/).

Analysis of structure and function

We used our own custom software and published freeware programs for this study. Briefly, these programs included ORF Finder (http://www.ncbi.nlm.nih.gov/gorf/gorf.html) to determine ORF, DNA_GC_Calculator (http://www.genome.iastate.edu/ftp/share/DNAgcCal/) to analyze the GC content, and the EMBOSS program (http://www.hgmp.mrc.ac.uk/Software/EMBOSS/Apps/) and Antheprot 5.0 (http://antheprot-pbil.ibcp.fr/) for physical and biochemical features prediction. For TM region prediction we used TMHMM (http://www.cbi.dtu.dk/services/TMHMM/), TopPred2 (http://bioweb.pasteur.fr/seqanal/protein/intro-uk.html), TMpred (http://www.ch.embnet.org/software/TMPRED_form.html) and Hmmtop (http://bioresearch.ac.uk/whatsnew/detail/3022811.html) simultaneously for comparative analysis and independent computational confirmation. Three programs were selected for secondary structure prediction: PSIpred (http://bioinf.cs.ucl.ac.uk/psipred/), NNpredict (http://www.cmpharm.ucsf.edu/~nomi/nnpredict.html), SPLIT (http://garlic.mefos.hr/split/) and Antheprot 5.0. As for the phylogenetic analysis, we used ClustalW (http://www-igbmc.u-strasbg.fr/BioInfo/ClustalW) for mutiple sequence alignment as well as PFAM searching (http://www.sanger.ac.uk/Software/Pfam/search.shtml) for protein family classification.

Acknowledgements

We thank Ministry of Science and Technology of China, Chinese Academy of Sciences, and National

Natural Science Foundation of China for financial support. We also want to express our thankfulness to collaborators and clinicians from Peking Union Medical College Hospital, National Center of Disease Control of China, and the Municipal Governments of Beijing and Hangzhou.

References

1. Cavanagh, D. and Brown, T.D.K. (ed.) 1990. *Coronaviruses and their diseases.* Plenum Press, New York, USA.
2. Corse, E. and Machamer C.E. 2000. Infectious bronchitis virus E protein is targeted to the Golgi complex and directs release of virus-like particles. *J. Virol.* 74: 4319-4326.
3. An, S., *et al.* 1999. Induction of apoptosis in murine coronavirus-infected cultured cells and demonstration of E protein as an apoptosis inducer. *J. Virol.* 73: 7853-7859.
4. Maeda, J., *et al.* 2001. Membrane topology of coronavirus E protein. *Virology* 281: 163-169.
5. Baudoux, P., *et al.* 1998. Coronavirus pseudoparticles formed with recombinant M and E protein induce alpha interferon synthesis by leukocytes. *J. Virol.* 72: 8638-8643.
6. Ortego, J., *et al.* 2002. Generation of a replication-competent, propagation-deficient virus vector based on the transmissible gastroenteritis coronavirus genome. *J. Virol.* 76: 11518-11529.
7. Kuo, L. and Masters, P.S. 2002. Genetic evidence for a structural interaction between the carboxy termini of the membrane and nucleocapsid proteins of mouse hepatitis virus. *J. Virol.* 76: 4987-4999.
8. Fischer, F., *et al.* 1998. Analysis of constructed E gene mutants of mouse hepatitis virus confirms a pivotal role for E protein in coronavirus assembly. *J. Virol.* 72: 7885-7894.
9. Qin, E.D., *et al.* 2003. A complete sequence and comparative analysis of a SARS-associated virus (Isolate BJ01). *Chin. Sci. Bull.* 48: 941-948.
10. Rota, P.A., *et al.* 2003. Characterization of a novel coronavirus associated with severe acute respiratory syndrome. *Science* 300: 1394-1399.
11. Marra, M.A., *et al.* 2003. The genome sequence of the SARS-associated coronavirus. *Science* 300: 1399-1404.
12. Ruan, Y., *et al.* 2003. Comparative full-length genome sequence analysis of 14 SARS coronavirus isolates and common mutations associated with putative origins of infection. *Lancet* 361: 1779-1785.
13. Holmes, K.V. 2003. SARS-associated coronavirus. *N. Engl. J. Med.* 348:1948-1951.
14. Hasler, U., *et al.* 2000. Determinants of topogenesis and glycosylation of type II membrane proteins. *J. Biol. Chem.* 275: 29011-29022.
15. Godet, M., *et al.* 1992. TGEV corona virus ORF4 encodes a membrane protein that is incorporated into virions. *Virology* 188: 666-675.
16. Ballesteros, M.L., *et al.* 1997. Two amino acid changes at the N-terminus of transmissible gastroenteritis coronavirus spike protein result in the loss of enteric tropism. *Virology* 227: 378-388.
17. Sanchez, C.M., *et al.* 1999. Targeted Recombination demonstrates that the spike gene of transmissible gastroenteritis coronavirus is a determinant of its enteric tropism and virulence. *J. Virol.* 73: 7607-7618.
18. Leparc-Goffart, I., *et al.* 1998. Targeted recombination within the spike gene of murine coronavirus mouse hepatitis virus-A59: Q59 is a determinant of hepatotropism. *J. Virol.* 72: 9628-9636.
19. Gallagher, T.M. and Buchmeier, M.J. 2001. Coronavirus Spike Proteins in Viral Entry and Pathogenesis. *Virology* 279: 371-374.
20. Garoff, H., *et al.* 1998. Virus maturation by budding. *Microbiol. Mol. Biol. Rev.* 62: 1171-1190.
21. Locker, J.K., *et al.* 1995. The organization of the endoplasmic reticulum and the intermediate compartment in cultured rat hippocampal neurons. *Mol. Biol. Cell* 6: 1315-1332.
22. de Haan, C.A.M., *et al.* 1998. Coronavirus particle assembly: primary structure requirements of the membrane protein. *J. Virol.* 72: 6838-6850.
23. Raamsman, M.J.B., *et al.* 2000. Characterization of the coronavirus mouse hepatitis virus strain A59 small membrane protein E. *J. Virol.* 74: 2333-2342.