Small envelope protein E of SARS: cloning, expression, purification, CD determination, and bioinformatics analysis[1]

SHEN Xu[2*], XUE Jian-Hua[3*], YU Chang-Ying[2], LUO Hai-Bin[2], QIN Lei[4], YU Xiao-Jing[5], CHEN Jing[2], CHEN Li-Li[2], XIONG Bin[2], YUE Li-Duo[2], CAI Jian-Hua[2], SHEN Jian-Hua[2], LUO Xiao-Min[2], CHEN Kai-Xian[2], SHI Tie-Liu[5*], LI Yi-Xue[4,5*], HU Geng-Xi[6*], JIANG Hua-Liang[2*]

[2]Drug Discovery and Design Center, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 201203; [3]Shanghai Health Digit Limited, Shanghai 200233; [4]Shanghai Center for Bioinformation Technology, Shanghai 201203; [5]Bioinformation Center, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031; [6]Institute of Biochemistry and Cell Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China

KEY WORDS severe acute respiratory syndrome (SARS); small envelope protein; gene expression; bioinformatics; circular dichroism spectroscopy

ABSTRACT

AIM: To obtain the pure sample of SARS small envelope E protein (SARS E protein), study its properties and analyze its possible functions. **METHODS:** The plasmid of SARS E protein was constructed by the polymerase chain reaction (PCR), and the protein was expressed in the *E coli* strain. The secondary structure feature of the protein was determined by circular dichroism (CD) technique. The possible functions of this protein were annotated by bioinformatics methods, and its possible three-dimensional model was constructed by molecular modeling. **RESULTS:** The pure sample of SARS E protein was obtained. The secondary structure feature derived from CD determination is similar to that from the secondary structure prediction. Bioinformatics analysis indicated that the key residues of SARS E protein were much conserved compared to the E proteins of other coronaviruses. In particular, the primary amino acid sequence of SARS E protein is much more similar to that of murine hepatitis virus (MHV) and other mammal coronaviruses. The transmembrane (TM) segment of the SARS E protein is relatively more conserved in the whole protein than other regions. **CONCLUSION:** The success of expressing the SARS E protein is a good starting point for investigating the structure and functions of this protein and SARS coronavirus itself as well. The SARS E protein may fold in water solution in a similar way as it in membrane-water mixed environment. It is possible that β-sheet I of the SARS E protein interacts with the membrane surface via hydrogen bonding, this β-sheet may uncoil to a random structure in water solution.

[1] Project supported by the 863 Hi-Tech Program, No 2001AA235051, 2001AA235071 and 2001AA231111, the National Natural Science Foundation of China, No 29725203, 20072042, the State Key Program of Basic Research of China, No 002CB512801 and 002CB512802, and the special programs of oppugning SARS from the Ministry of Science and Technology, Chinese Academy of Sciences and Shanghai Basic Research Project from the Shanghai Science and Technology Commission, No 02DJ14070.

[*] Correspondence to Prof SHEN Xu, JIANG Hua-Liang, HU Geng-Xi, LI Yi-Xue, XUE Jian-Hua, and SHI Tie-Liu.

Received 2003-05-13 Accepted 2003-05-15

INTRODUCTION

An outbreak of atypical pneumonia, designated "severe acute respiratory syndrome (SARS)" by the World Health Organization (WHO) and first identified in Guangdong Province, China, has spread to several countries[1,2], and recently the SARS infection is austere in numerous places in China, including Beijing, Shanxi province and the Inner Mongolia Autonomous Region. Since the outbreak of SARS beginning early in 2003, a number of laboratories worldwide have undertaken the identification of the causative agent. Recently, it was recognized that the coronavirus was the prime criminal for SARS infection[1,2]. Afterwards, the genome sequencings for the coronaviruses from different SASR patients have been finished, which have been deposited in the GenBank already (http://www.ncbi.nlm.nih.gov/).

Open reading frames (ORF) analysis through sequence similarity to the known coronaviruses indicated that several proteins coded by SARS genome might play important functions associated with SARS infection, including replicases 1a and 1b, the spike (S) protein, the matrix (M) protein, the nucleocapsid (N) protein and the small envelope (E) protein[1,2]. To identify the functions of the SARS proteins and to establish molecular models for screening anti-SARS drugs, we are trying to overexpress several important proteins of SARS.

In general, coronaviruses are enveloped positive-strand RNA viruses that contain, at a minimum, four structural proteins[3]: the S protein, the M protein, the N protein, and the small E protein. While S, M and N proteins have been broadly studied for their important roles in receptor binding and virion budding, the significance of E protein has come to be appreciated only latterly[4,5]. With only 76 amino acids (aa) or so in length, E protein has long been taken as a voidable membrane component of coronaviruses. Yet the research revealed that E protein played an important multifunctional role in coronavirus virion life cycle[6,7]. In the present paper, we report the cloning, expression, purification, and the primary properties of the E protein of SARS.

MATERIALS AND METHODS

Materials The enzymes for the preparation of SARS E protein were purchased from Invitrogen, and the bacterial strains BL21 (DE3) and DH5alpha were from Novagen. The glutathione-sepharose 4B affinity and benzamidine-Sepharose 6B resins were purchased from Amersham Pharmacia Biotech, Inc. Except that the LMW Marker (protein ladder, 10-200 kDa) was from Fermentas, the markers for molecular weight estimation were purchased from BioRad Co. Other commercially available materials were of reagent grade or higher.

Plasmid construction The SARS E protein gene contains 231 bp nucleotides (Ch: 26102-26332, TOR2). The gene was synthesized by polymerase chain reaction (PCR). According to gene E accession number NC_004718, four oligonucleotide primers were synthesized:

E1+: GGCCGGATCCATGTACTCATTCGTTTCGGA
AGAAACAGGTACGTTAATAGTTAATAGCGTACTTCTT
E2+: GTTAATAGCGTACTTCTTTTTCTTGCTTTCG
TGGTATTCTTGCTAGTCACACTAGCCATCCTTACT
GCGCTTCGATTGTGTGCGTAC
E3–:GCGAGTAGACGTAAACCGTTGGTTTTACTAA
ACTCACGTTAACAATATTGCAGCAGTACGCACACA
ATCGAAGCGC
E4–:GCGCGAATTCTTAGACCAGAAGATCAGGAAC
TCCTTCAGAAGAGTTCAGATTTTTAACACGCGAGT
AGACGTAAA

PCR product was digested with *Bam*HI and *EcoR*I, inserted into pGEX2T expression vector and transformed into DH5alpha host cell. Eighteen positive clones were sequenced and analyzed. The results were confirmed by the gene E sequence.

Protein expression and purification Above constructed plasmid E1 was transformed into the *E coli* strain, BL21 (DE3), and single colony was grown for 12 h at 37 °C in 3 mL LB broth containing ampicillin as antibiotic. The bacterial cells were grown in LB medium at 37 °C with ampicillin (1.0 mmol/L). The protein expression was induced with 0.5 mmol/L of IPTG (isopropyl-β-*D*-thiogalasctopyranoside) after OD_{600} of the cultured medium was around 0.7. After IPTG induction at 25 °C for 5 h, cells were harvested by centrifugation at 4000×*g* at 4 °C for 30 min. Suspended the cell pellet by PBS buffer, and the suspension was centrifuged at 4000×*g* for 15 min. The supernatant was discarded and the pellet was collected and stored at -80 °C. Every 7 grams of cells were suspended in 20 mL of sonication buffer (50 mmol/L Tris-HCl, pH 8.0, 1 mmol/L edetic acid, 300 mmol/L NaCl, 1 mmol/L DDT(dithiothreitol), 1 mmol/L PMSF) and sonicated in icy bath for 30 min. The lysed cells were centrifuged at 14000 × *g*, 4 °C for 1 h and the pellets were discarded. The supernatant was applied to a glutathione Sepharose

4B column equilibrated with Buffer A (50 mmol/L Tris-HCl, pH 8.0, 1 mmol/L DDT, 1 mmol/L edetic acid) and the column was washed with the same buffer until the optical density of the eluted buffer returned to the baseline level. After the GST-fused SARS E protein (GST-E) bound column was equilibrated with Buffer B (20 mmol/L Tris-HCl, pH 8.0, 150 mmol/L NaCl) including 2.5 mmol/L $CaCl_2$ containing 0.5 % of thrombin at 25 °C for 3 h, the eluate was collected. The column was then washed with Buffer B, and the eluate was combined and further purified by Mono-Q ion exchanged column and Sephadex G-75 gel filtration system. Finally a benzamidine-Sepharose 6B column was used to remove thrombin completely. Fig 1 lists the results of SDS-polyacryamide gel electrophoresis of SARS E protein purification. For circular dichroism analysis, SARS E protein was dialyzed against CD buffer (20 mmol/L Na-P, 100 mmol/L NaCl, pH 7.4).

Bioinformatics analysis and molecular modeling The E protein sequence of SARS coronavirus (SARS-CoV) used in this research was extracted from NCBI GenBank (http://www.ncbi.nlm.nih.gov/), the base sequence is from NC_004718. The amino acid sequence was compared with those from other coronaviruses which represent different species, including group 1: human coronavirus 229E (HCoV), porcine respiratory coronavirus (PrCoV), feline coronavirus (FCoV), and canine coronavirus (CCoV); group 2: bovine coronavirus (BCoV), marine hepatitis virus (MHV), rat coronavirus (RtCoV), and porcine hemagglutinating encephalomyelitis virus (PHEV); and group 3: avian infectious bronchitis virus (IBV). The sequences were organized as FASTA format and the file was inputted into BioEdit to carry out the alignment manually. The secondary structure of SARS E protein was predicted by using TMHMM Server v.20 through online submission (http://www.cbs.dtu.dk/services/TMHMM), and the results were collected and analyzed by comparing the data manually.

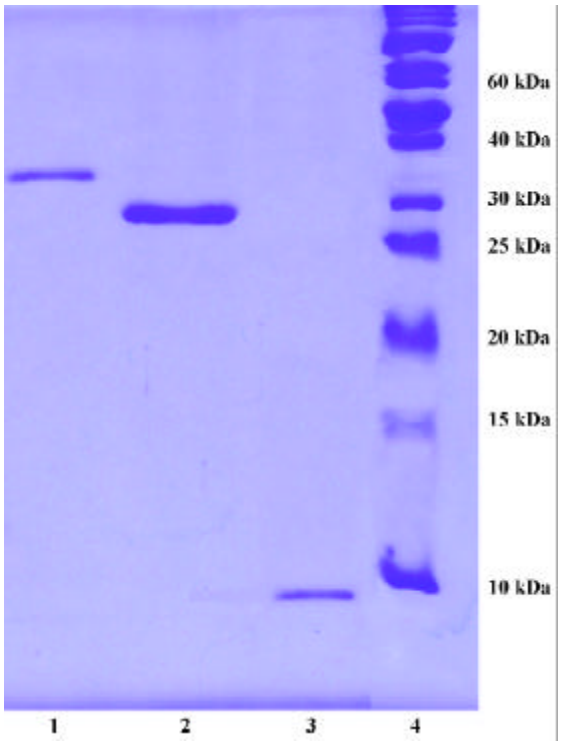

Fig 1. SDS-polyacryamide gel electrophoresis of SARS E protein (Lane 1: GST-fused SARS E protein (GST-E) purified by affinity column chromatography on glutathione sepharose 4B resin. Lane 2: GST protein obtained by washing the column with 20 mmol/L glutathione after the incubation of GST-E with thrombin. Lane 3: purified SARS E protein. Lane 4: LMW Marker.)

Based on the secondary structural prediction, the transmembrane segment was identified from the residues 12 to 34; the two short β-sheets were composed by residues 45 to 51 (β-sheet I) and residues 55 to 61 (β-sheet II), respectively. Firstly, the 3D models of the N terminal, transmembrane and C terminal segments were constructed separately. The transmembrane was modeled as an α-helix structure; the N terminal and C terminal segments were built as a random structure except the two short β-sheets. Then these three models were assembled as a whole structure. The models were constructed by using the Bioploymer module encoded in Sybyl6.8[8]. The whole structural model was optimized by the simulated annealing (SA)[9]. The total SA run cycle was set up as 50, the high and low temperatures were respectively set up as 1000 and 100 °C. Afterwards, the structure model was minimized by steepest descent first, then by conjugate gradient method to the energy gradient root-mean-square (RMS) <0.05 kcal/(mol·Å). Amber force field[10] was used in SA simulation and structure minimization (a distance-dependent dielectric constant of 4.0, nonbonded cut-off 8 Å, Kollman-all-atom charges). Finally, α-helix of the TM segment was embedded into the palmitoyloleoylphosphatidylcholine (POPC) lipid bilayer manually. The coordinates of POPC lipid bilayer were downloaded from (http://moose.bio.ucalgary.ca)[11].

RESULTS AND DISCUSSION

Bioinformatics annotation As a membrane protein, the major biological function of the E protein is

to participate in the formation of viral envelope. Meanwhile, it also plays an important role in the viral replication in some coronavirus viruses, such as in transmissible gastroenteritis coronavirus (TGEV) and in marine hepatitis virus (MHV)[12,13]. In addition, it is proved that E protein is related to the apoptosis of the E-protein-expressing cells in MHV[14].

The bioinformatics analysis result is listed in Tab 1. The homology of E protein for SARS coronavirus to other coronavirus is very low, ranging from ~17 % to ~23 % based on the amino acid sequence[1]. The prediction for the second structure of SARS E protein by using TMHMM system shows that there is a transmembrane structure in the protein and the position in the amino acid sequence ranges from residues 12 to 34, which is a hydrophobic region. The first 11 amino acids of N-terminus are in the virion, whereas the hydrophilic tail is exposed to the cytoplasmic side. Previous experimental results indicated that there was no signal peptide cleavage process happening during or after membrane integration of the protein[15]. Since the E protein sequence lacks a predicted cleavage site (data are not shown here), it seems mostly that the first 11 amino acids go for membrane integration directly.

Tab 1. The transmembrane (TM) region of E proteins[a]

Virus strain	Group	Protein ID	E protein length (in aa)	Predicted TM region
SARS-CoV	1 or 2	NP_828854	76	12-34
PrCoV	1	Z24675.1	82	15-17
FCoV	1	CAA74228.1	82	15-37
CCoV	1	BAA02412.1	82	15-37
HCoV	1	NP_073554.1	77	10-32
BCoV	2	AAL40404.1	84	15-37
PHEV	2	AAM77003.1	84	15-37
RtCoV	2	AAF97741.1	88	15-37,39-59
MHV	2	NP_068673.1	83	15-37,39-59
IBV	3	AAO33465.1	103	13-32,47-65

[a] The E protein sequences used were retrieved from the GenBank. Sequences used for the prediction included group 1: human coronavirus 229E (HCoV-229E, NP_073554.1), porcine respiratory coronavirus (PrCoV, Z24675.1), canine coronavirus (CCoV, BAA02412.1), and feline coronavirus (FCoV, CAA74228.1); group 2: bovine coronavirus (BCoV, AAL40404.1), murine hepatitis virus (MHV, NP_068673.1), porcine hemagglutinating encephalomyelitis virus (PHEV, AAM77003.1), and rat coronavirus (RtCoV, AAF97741.1); group 3: infectious bronchitis virus (IBV, AAO33465.1).

Even though the homology of E proteins between HCoV and SARS is low, the predicted transmembrane structures are very similar, the crossing membrane amino acid sequence for HCoV is from 10-30 and the transmembrane sequences between these two coronaviruses are conserved at most positions (Tab 1, Fig 2). Also the transmembrane regions for those studied coronaviruses are very similar (Tab 1), although the variations for those amino acid sequences are significant. These results imply that the transmembrane structure of SARS E protein should share the similar topology to the E proteins of other coronaviruses.

The phylogenetic study with the structure proteins and some non-structure proteins indicated that the SARS coronavirus should be separated from other three groups and listed as a new group[1,2]. Interestingly, we found that the key amino acids in the SARS E protein were well conserved compared with the groups 1 and 2 coronaviruses (Fig 2). Previous results have demonstrated that the lysine following the transmembrane sequence is conserved in the E proteins, and the proline located in amino acids 50-60 is absolutely conserved among all coronavirus E proteins. The alignment of the E protein sequences in this study confirmed that the SARS E protein kept this conservation. The conserved proline in SARS E protein is located at position 54, the same position as it in MHV and RtCoV E proteins (Fig 2). For the SARS E protein, arginine is located at the position of the conserved lysine in other coronavirus E proteins. This minor difference would not change the conservation dramatically, for lysine and arginine have similar structures and properties. This indicates that the SARS coronavirus is probably related closer to the group 1 or group 2.

In addition, the secondary structure prediction for SARS E protein indicated that the percentages of α-helix, β-sheet, and random coil are 30.26 %, 18.43 %, and 51.32 %, respectively.

Circular dichroism (CD) analysis CD spectrum of the SARS E protein (0.3 g/L) in solution (escaped from lipid environment) at 25 °C was obtained by means of a JASCO 715 spectropolarimeter equipped with a Neslab water bath. The CD spectrum was recorded using an optical cell with a 0.1 mmol/L path-length for the far-UV region, which is shown in Fig 3. Its secondary structural feature is estimated by using the CDFIT program (http://www-structure.llnl.gov/cd/cdtutorial.htm). CD spectrum revealed that the per-

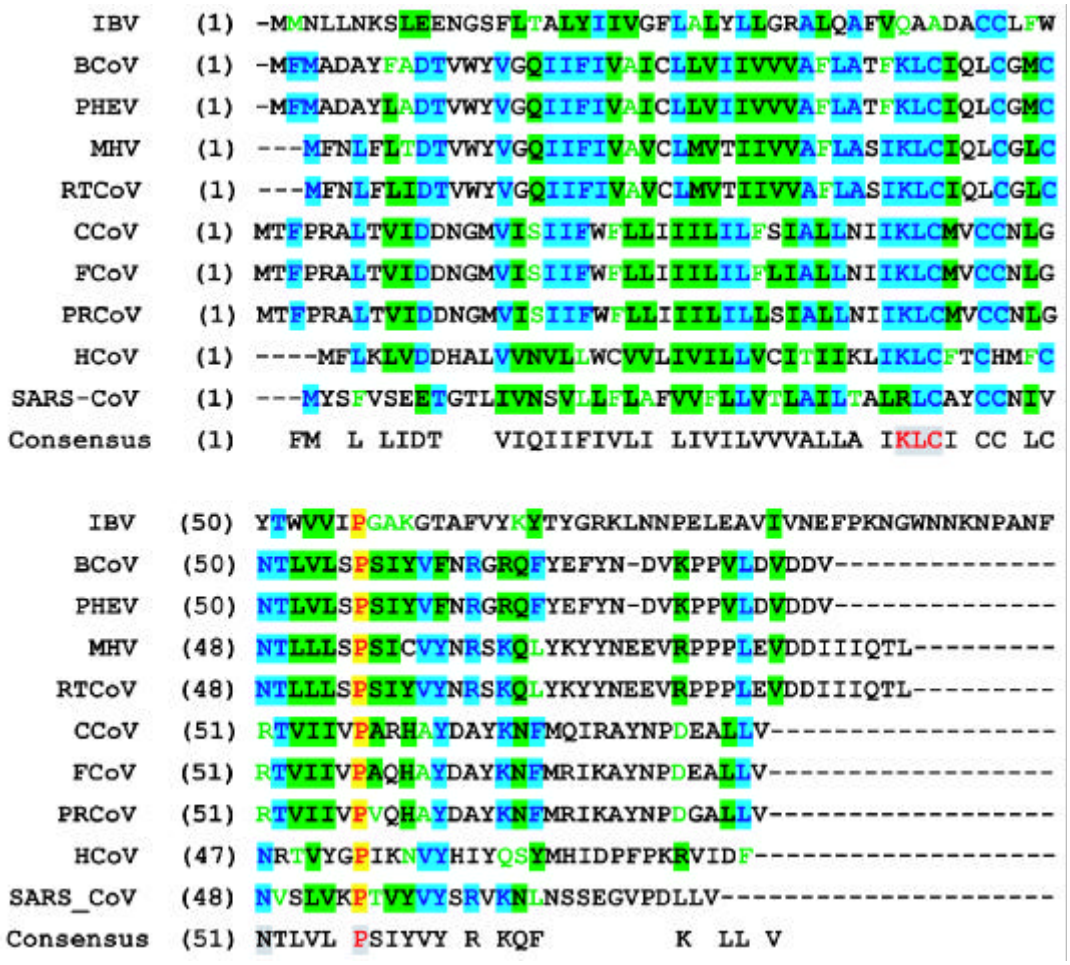

Fig 2. Conserved motifs in coronavirus E proteins. The portion of the conserved sequences is shown. The conserved amino acids are highlighted in colors. Sequences used in the alignments include: group 1–human coronavirus 229E (HCoV-229E, NP_073554.1), Porcine respiratory coronavirus (PrCoV, Z24675.1), canine coronavirus (CCoV, BAA02412.1), and feline coronavirus (FCoV, CAA74228.1); group 2–bovine coronavirus (BCoV, AAL40404.1), murine hepatitis virus (MHV, NP_068673. 1), porcine hemagglutinating encephalomyelitis virus (PHEV, AAM77003.1), and rat coronavirus (RtCoV, AAF97741.1); group 3–infectious bronchitis virus (IBV, AAO33465.1).

centages for α-helix, β-sheet, and random coil were 34.82 %, 10.76 %, and 54.42 %, respectively, which are in general agreement with the above secondary structure prediction.

A primary 3D model According to the secondary structure prediction, we constructed a primary 3D model for the SARS E protein (Fig 4). This 3D model demonstrates that the TM segment of the SARS E protein adopts an α-helix conformation, and inserts into the lipid bilayer well. The two short β-sheets do not form an anti-parallel hairpin structure; the first β-sheet (β-sheet I) forms hydrogen bonds with the surface of the lipid bilayer (Fig 4). CD spectrum only gives the secondary structure information in water solution, which should differ from that in membrane-water mixed

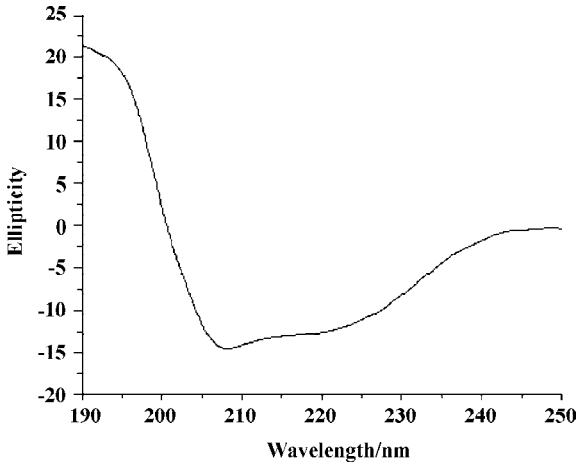

Fig 3. CD spectrum of the SARS E protein at 25 °C.

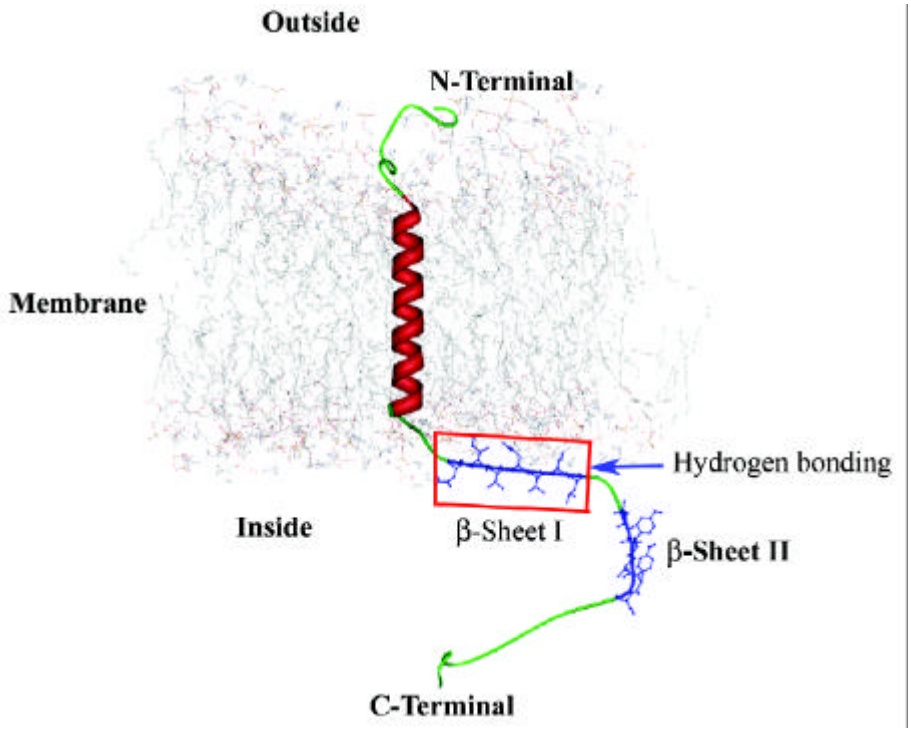

Fig 4. A primary 3D model of the SARS E protein. The transmembrane segment was embedded in the POPC lipid bilayer, the two short b-sheets were represented by ball-and-stick model.

environment. However, the comparison of the CD spectrum with the secondary structure prediction result shows that SARS E protein might adopt a similar fold in water solution to that in membrane-water mixed environment except the two β-sheets. In water solution, the percentage of β-sheet of SARS E protein (10.76 %) is as about half as that in the membrane-water mixed environment (18.43 %). From this result and the 3D model, one hypothesis could be tentatively proposed: β-sheet I is not stable in water solution due to the absence of the stabilization of its hydrogen-bonding interaction to the membrane, it may thus exist in a ramdom coil structure. Nevertheless, we just obtained a very crude 3D model. More robust 3D model of SARS E protein is being constructed by the long-time molecular dynamics simulation, X-ray crystallographic, and nuclear magnetic resonance (NMR) methods.

CONCLUSIONS

Bioinformatics analysis indicated that SARS E protein was much conserved compared to the E proteins of other coronaviruses; in particular, the primary amino acid sequence of SARS E protein is much more similiar to MHV and other mammal coronaviruses than to avian IBV E protein (Fig 2, Tab 1). The transmembrane (TM) segment of SARS E protein is relatively more conserved in the whole protein than other region, while the N terminal homology is the poorest. For the C terminal region, the relatively conserved "cystine-cystine (CC)" pattern has been reported to be a palmitoylation site[16]; the most conserved "proline (P)" residue makes the C terminal structure more flexible. These conserved sites may be essentially associated with the function of E protein.

We have successfully constructed the plasmid of SARS E protein, established the expression system and obtained the pure protein sample. The CD spectrum determination and the secondary structure prediction indicated that SARS E protein might fold in water solution in a similar way as it in membrane-water mixed environment except the β-sheet. Based on the secondary structure prediction, a primary 3D structure model of SARS E protein embedded in a lipid bilayer was constructed by molecular modeling method. This model gives a clue for investigating the structure properties, it is possible that β-sheet I interacts with the membrane surface via hydrogen bonding; this β-sheet may uncoil to a random structure in the water solution.

REFERENCES

1 Paul A. Rota, M. Steven Oberste, Stephan S. Monroe, W. Allan Nix, Ray Campagnoli, *et al.* Characterization of a novel coronavirus associated with severe acute respiratory syndrome. Science (Sciencexpress) 2003. May 1.

2 Marra MA, Jones SJ, Astell CR, Holt RA, Brooks-Wilson A, *et al*. The genome sequence of the SARS-associated coronavirus. Science (Sciencexpress) 2003. May 1.

3 Siddell SG. The coronaviridae: an introduction. In: Siddell SG, editor. *The Coronaviridae*. New York: Plenum Press; 1995. p1-10.

4 Tung FY, Abraham S, Sethna M, Hung SL, Sethna P, Hogue BG, *et al*. The 9-kDa hydrophobic protein encoded at the 3' end of the porcine transmissible gastroenteritis coronavirus genome is membrane-associated. Viroolgy 1992; 186: 676-83.

5 Maeda J, Repass JR, Maeda A, Makino S. Membrane topology of coronavirus E protein. Virology 2001; 281: 163-9.

6 Bos ECW, Luytjes W, van der Meulen H, Koerten HK, Spaan WJM. The production of recombinant infectious DI-particles of a murine coronavirus in the absence of helper virus. Virology 1996; 218: 52-60.

7 Vennema H, Godeke GJ, Rossen JWA, Voorhout MF, Horzinek MC, Opstelten DJE, *et al*. Nucleocapsid-independent assembly of coronavirus-like particles by co-expression of viral envelope protein genes. EMBO J 1996; 15: 2020-8.

8 Sybyl [molecular modeling package], version 6.8; St. Louis (MO): Tripos Associates; 2000.

9 Kirpatrick S, Gelatt CD, Vecchi MP. Optimization by simulated annealing. Science 1983; 220: 671-80.

10 Cornell WD, Cieplak P, Bayly CI, Gould IR, Merz KM, *et al.* A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. J Am Chem Soc 1995; 117: 5179-97.

11 Tieleman DP, Berendsen HJC, Sansom MSP. An alamethicin channel in a lipid bilayer: molecular dynamics simulations. Biophys J 1999; 76: 1757-69.

12 Godet M, Haridon RL, Vautherot JF, Laude H. TGEV coronavirus ORF4 encodes a membrane protein that is incorporated into virion. Virol 1992; 188: 666-75.

13 Kuo L, Masters PS. The small envelope protein E is not essential for murine coronavirus replication. J Virol 2003; 77: 4597-608.

14 An SW, Chen CJ, Yu X, Leibowitz JL, Makino S. Induction of apoptosis in murine coronavirus-infected cultured cells and demonstration of E protein as an apoptosis Inducer J. Virol 1999; 73: 7853–9.

15 Raamsman JB, Locker JK, de Hooge A, de Vries AAF, Griffiths G, *et al*. Characterization of the coronavirus mouse hepatitis virus strain A59 small membrane protein E. J Virol 2000; 74: 2333-42.

16 Anderson AM, Melin L, Bean A, Pettersson RF. A retention signal necessary and sufficient for Golgi localization maps to the cytoplasmic tail of a Bunyaviridae (Uukuniemi virus) membrane glycoprotein. J Virol 1997; 71: 4717-27.