The Structural Characterization and Antigenicity of the S Protein of SARS-CoV

Jingxiang Li[1]*, Chunqing Luo[1]*, Yajun Deng[1,2]*, Yujun Han[1]*, Lin Tang[1], Jing Wang[1,3], Jia Ji[1], Jia Ye[1,4], Fanbo Jiang[1], Zhao Xu[4], Wei Tong[1], Wei Wei[4], Qingrun Zhang[1], Shengbin Li[1,2], Wei Li[1], Hongyan Li[1], Yudong Li[1], Wei Dong[1], Jian Wang[1,4], Shengli Bi[5], and Huanming Yang[1,4]#

[1] Beijing Genomics Institute, Chinese Academy of Sciences, Beijing 101300, China; [2] Medical College, Xi'an Jiaotong University, Xi'an 710049, China; [3] College of Life Sciences, Peking University, Beijing 100871, China; [4] James D. Watson Institute of Genome Sciences, Zhijiang Campus, Zhejiang University, Hangzhou 310008, China; [5] Center of Disease Control and Prevention, Beijing 100050, China.

The corona-like spikes or peplomers on the surface of the virion under electronic microscope are the most striking features of coronaviruses. The S (spike) protein is the largest structural protein, with 1,255 amino acids, in the viral genome. Its structure can be divided into three regions: a long N-terminal region in the exterior, a characteristic transmembrane (TM) region, and a short C-terminus in the interior of a virion. We detected fifteen substitutions of nucleotides by comparisons with the seventeen published SARS-CoV genome sequences, eight (53.3%) of which are non-synonymous mutations leading to amino acid alternations with predicted physiochemical changes. The possible antigenic determinants of the S protein are predicted, and the result is confirmed by ELISA (enzyme-linked immunosorbent assay) with synthesized peptides. Another profound finding is that three disulfide bonds are defined at the C-terminus with the N-terminus of the E (envelope) protein, based on the typical sequence and positions, thus establishing the structural connection with these two important structural proteins, if confirmed. Phylogenetic analysis reveals several conserved regions that might be potent drug targets.

Key words: SARS, coronavirus, the S protein, structure, antigenicity

Introduction

The S (spike, or E2 glycoprotein precursor) protein is the largest structural protein of Type I membrane glycoproteins and is the main component of the characteristic spikes on the surface of coronaviruses. It is composed of a knob region (S1) and a stem region (S2), and in some species of coronaviruses the precursor S protein will be cleaved into the N-terminal S1 and C-terminal S2 glycopolypeptides at a special protease cleavage site (*1-4*). The S protein is postulated to be incorporated into the viral membrane by interaction with the M (membrane) protein, and subsequent release of mature virions from the smooth vesicles (*5*). The S protein is thought to be important for viral attachment through its interaction with the cellular receptors. Previous research has shown that the half of the S protein near the C-terminus is collapsed into coiled-coils, thus bringing a fusion peptide with the transmembrane (TM) region for cellular and viral membrane fusion (*6, 7*). Furthermore, it has been suggested that the S protein induces the production of virus-neutralizing antibodies and that amino acid changes in the S protein may dramatically affect the virulence altered pathogenesis and *in vitro* host cell tropism in other coronaviruses (*6, 8*).

We report, in this paper, the results of the comparative analyses on the S protein of BJ Group (BJ01-BJ04) and 13 other reported SARS-CoV genome sequences (*9-12*) as well as 13 out of 17 members of *Coronaviridae* with the data of the S protein available.

Results and Discussion

General features

The ORF (open reading frame) of the S protein, nucleotide (nt) position 21,473-25,240, locates between the ORFs for the R (replicase) protein (nt position

*** These authors contributed equally to this work.**
 # Corresponding author.
E-mail: yanghm@genomics.org.cn

This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

246-21,466) and PUP1 (putative uncharacterized protein 1, nt position 25,249-26,073; ref. *11*). The total size, together with its defined 5′ transcription regulation sequence (TRS, 10 nt), is 3,778 nt and accounts for 12.7% of the total genome (29,725 nt). The S protein has a GC content of 38.7% (A: U: C: G = 28.0: 33.3: 18.6: 20.1). It was predicted to encode a translation product of 1,255 amino acids (a.a.; 139.17 KD). The amino acid composition is showed in Table 1. We divided the S protein into 31 subregions according to a combination of typical physiochemical features as well as defined structures (Figure 1).

Table 1 The Amino Acid Composition of the S Protein

Non-polar, Neutral	Number	Percentage (%)
Ala	84	6.69
Gly	78	6.22
Ile	77	6.14
Leu	99	7.89
Met	20	1.59
Phe	83	6.61
Pro	57	4.54
Trp	11	0.88
Val	91	7.25
Total	**600**	**47.81**
Polar, Neutral		
Asn	81	6.45
Cys	39	3.11
Gln	55	4.37
Ser	96	7.65
Thr	100	7.97
Tyr	54	4.3
Total	**425**	**33.85**
Polar, Positive		
Arg	39	3.11
His	15	1.2
Lys	60	4.78
Total	**114**	**9.09**
Polar, Negative		
Asp	74	5.9
Glu	42	3.35
Total	**116**	**9.25**

Putative characteristics of the S protein

A Type I signal peptide was found at the N-terminus with the most likely cleavage site located between Codons 13 and 14 within a consensus sequence of TSG-SD (Figure 1F).

We predicted the presence of a TM region characterized by 23 highly hydrophobic amino acids. To confirm the reliability of this finding, a program, TMHMM, was used to predict the possible existence of the TM structures in the S proteins in five other coronaviruses. The results demonstrated that the S proteins in these five coronaviruses have the similar TM regions at the comparable positions, supporting the reliability of the SARS-CoV analysis.

Because the presence of TM region has been predicted, the S protein can be divided into three regions. The N-terminal exterior part is as long as 1,195 a.a. (Codons 1–1,195) and the C-terminal interior part is 37 a.a. in size (Codons 1,219–1,255). Two or three α-amphipathic regions were also predicted in the previous studies (*12*). We identified four coiled-coil-like motifs in the C-terminal region with high probability (Figure 1E). The TM region is hydrophobic and is composed of 23 a.a. (Codons 1,196–1,218; Figure 1F). The interior part has a low pI (pI 5.50) that is consistent with the hydrophilic environment in the

area close to the membrane. It is also highly conserved and is rich of Cys residues (7/37). Interestingly, comparisons of different plots reveal the relationships among GC content, charge, hydrophobicity, helix, coil and TM, and makes it possible to divide the primary structure of the S protein into 31 sub-regions (Figure 1).

Twenty-three putative N-glycosylated sites (Asp) have been detected in the long exterior region. Among them, eighteen have potential exceeding threshold (> 50%; Figure 2A). However, among 196 predicted O-glycosylated sites, merely 5 (2 Thr and 3 Ser) are reliable (potential > threshold; Figure 2B). In addition to the glycosylation sites, 57 modification sites are also predicted, including 22 N-myristoylation sites, 16 protein kinase C phosphorylation sites, 15 casein kinase II phosphorylation sites, 2 tyrosine kinase phosphorylation sites, 1 tyrosine sulfation site and 1 cAMP- and cGMP-dependent protein kinase phosphorylation site (data not shown).

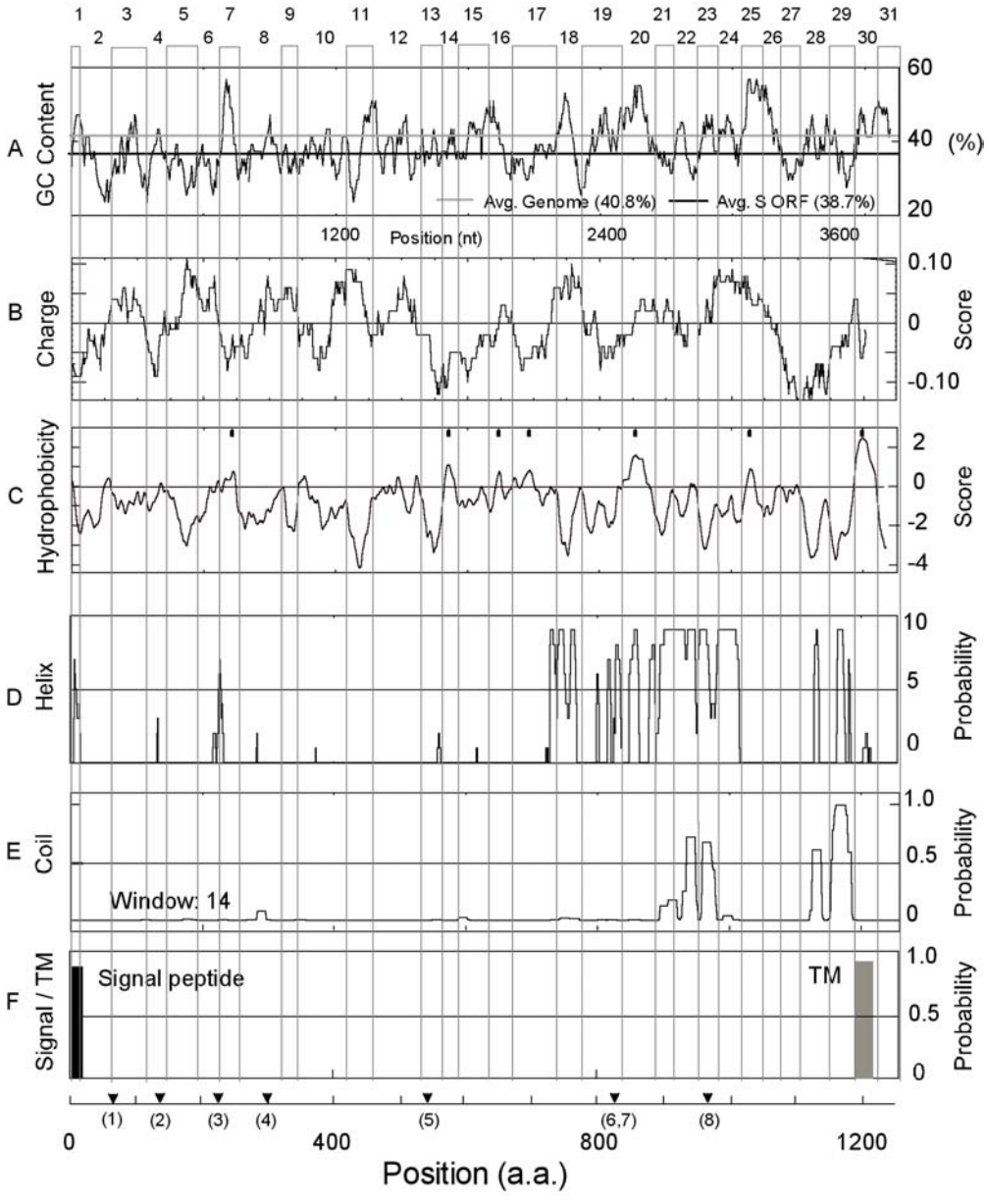

Fig.1 The predicted distributions of GC content (A), electric charge (B), hydrophobicity (C), helix (D), coil (E), signal peptide transmembrane region (F) and the points of non-synonymous substitutions (the reversed triangles) in the S protein.

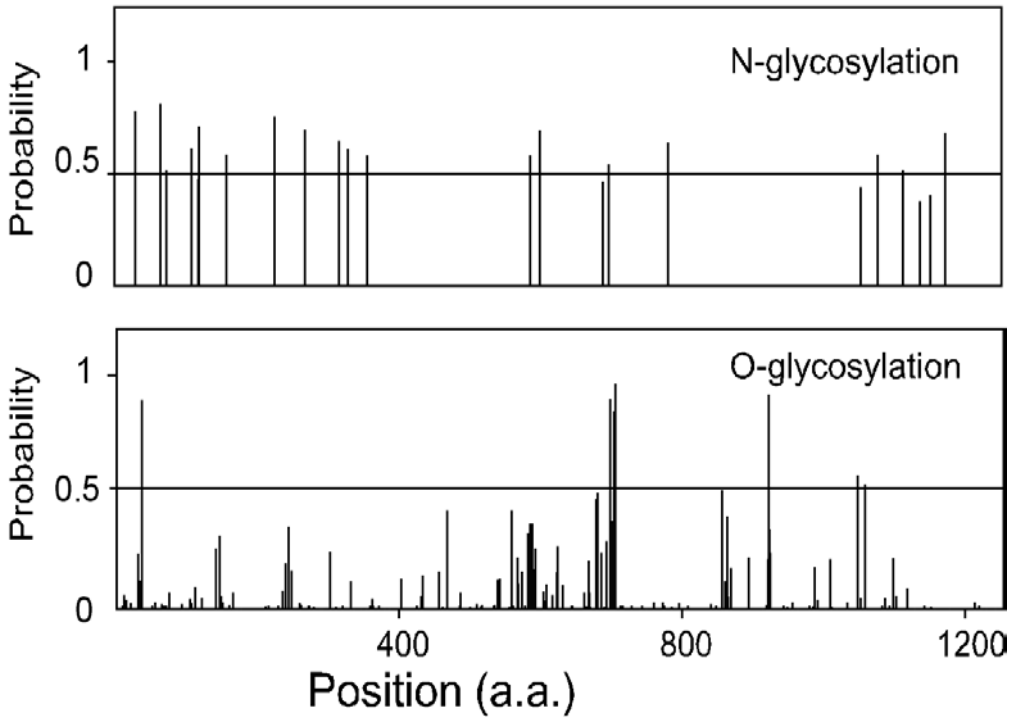

Fig.2 The predicted N-glycosylated and O-glycosylated sites in the S protein of SARS-CoV. The rectangles mean the different types of antigenic determinants.

Seven putative conserved domains (E value < 0.1) are determined by RPS-BLAST against database CDD v1.62. The most significant ones are a partial S1 (Condons 264–578) and an entire S2 (Condons 684–1,236). The former has a significantly higher E value (0.028) than the latter (5e–118). We believe there is more sequence divergence in the S1 than in the S2 region that is the main reason why only a portion of S1 region is predicted. The entire S1 region should be, roughly, from Codon 1 to around 680 (*4*). In addition to these regions, three low complexity domains are found, located at Codons 111–125 (SQSVIIINNSTNVVI), 862–875 (GTATAGWTFGAGAA) and 1,215–1,226 (LLCCMTSCCSCL, Cys-rich region).

In the S protein of SARS-CoV, we have not found any known cleavage site (RRAHR, RRARR, RRSRR, HRARS, HRARR, RRARS and HRARR). From the separate analysis result of the R protein of SARS-CoV in the same patent samples, one 3CLP (chymotrypsin-like protein) cleavage site (Condon 984, LQ|S), one reliable (Condon 39, RG|V) and 17 putative PLP (papain-like protein) cleavage sites were found. Thus, further experimental evidence is needed to determine whether the S protein is cleaved.

Antigenicity of the S protein

In contrast to the other structural proteins of SARS-CoV, the S protein of AIBV (avian infectious bronchitis virus) has the potential to induce serotype-specific and cross-reactive antibodies (*13*). In our bioinformatic analyses, we predicted the possible antigenic determinants of the S protein of SARS-CoV, and sixty-six possible antigenic determinants were determined with high probability. The average size of the predicted antigenic determinants is 13 a.a. and the sum length is 858 a.a., which covers 68.4% of the S protein. The ELISA (enzyme-linked immunosorbent assay) results confirmed our bioinformatics prediction (Table 2). Thus we are sure that the S protein of SARS-CoV contains many antigenic determinants and have close relations with the immunoreactions.

Table 2 The ELISA Checking Result of the Synthesized Peptides of the S Protein

Peptide	Position (a.a.)	Sequence	Positive
S67	67 - 89	TGFHTINHTFDNPVIPFKDGIYF	-
S84	84 - 108	KDGIYFAATEKSNVVRGWVFGSTMN	+
S101	101 - 122	WVFGSTMNNKSQSVIIINNSTN	+
S118	118 - 142	NNSTNVVIRACNFELCDNPFFAVSK	+
S260	260 - 281	TTFMLKYDENGTITDAVDCSQN	-
S277	277 - 298	DCSQNPLAELKCSVKSFEIDKG	+
S301	301 - 322	QTSNFRVVPSGDVVRFPNITNL	+++
S323	324 - 344	PFGEVFNATKFPSVYAWERKK	+
S345	345 - 366	ISNCVADYSVLYNSTFFSTFKC	-
S582	582 - 604	SVITPGTNASSEVAVLYQDVNCT	+
S599	599 - 620	QDVNCTDVSTAIHADQLTPAWR	+++
S624	624 - 645	TGNNVFQTQAGCLIGAEHVDTS	+
S645	645 - 667	SYECDIPIGAGICASYHTVSLLR	++
S886	886 - 906	YRFNGIGVTQNVLYENQKQIA	-
S1130	1130 -1147	FKEELDKYFKNHTSPDVD	+++
S1211	1121 - 1232	MVTILLCCMTSCCSCLKGACSC	-
S1227	1227 - 1248	KGACSCGSCCKFDEDDSEPVLK	+
S1234	1234 - 1255	SCCKFDEDDSEPVLKGVKLHYT	++

Mutations detected in the S protein

By comparing the published genome sequences of 16 isolates, 15 substitutions were detected at 15 positions. Among them, two were detected in more than one independent isolate. Eight (53.3%) are transitions (5 C/Ts and 3 A/Gs), and the other seven (46.7%) are transversions (4 A/Cs, 1 C/G and 2 T/Gs). Additionally, eight of fifteen are non-synonymous substitutions, some of which would lead to changes with regard to pI, polarity and charge (Table 3). There are more substitutions (10; 66.7%) in the S1 region than in the S2 (5, 33.3%). Considering that S1 is located outside of the virion and is believed to function by binding to the surface receptor of cells, the high mutation rate may allow the virion to escape from recognizing of immune system and also provide extraordinary variations in host range and tissue tropism.

Table 3 The Distribution of Substitutions in the S Protein

Subregions (a.a.)	Position (a.a./nt)	Substutution (a.a./nt)	pI value	Hydrophobicity (%)	Hydrophilicity (%)	Charge (+)	Charge (-)	Isolate
1 (60–115)	77(21,702)	Asp-Gly (A/G)	7.9–9.0	(33.9)	46.4–44.6	(10.7)	5.4–3.6	BJ04, CUHK, HKU, SIN1, SIN2, SIN3, TOR2, SIN4, SIN5, TW1, Urbani, ZJ01
2 (115–145)	144(21,902)	Met-Leu (A/C)	(5.8)	(38.7)	(38.7)	(6.5)	(6.5)	BJ03
3 (220–255)	244(22,203)	Thr-Ile (C/T)	(8.0)	33.3–36.1	33.3–30.6	(5.6)	(2.8)	CUHK, HKU, SIN1, SIN2, SIN3, SIN4, SIN5, TOR2, TW1, Urbani, ZJ01
4 (255–320)	311(22,403)	Gly-Arg (G/A)	5.1–6.0	(28.8)	48.5–50.0	10.6–12.1	(12.1)	BJ02
5 (560–580)	577(23,201)	Ser-Ala (T/G)	(4.3)	(23.8)	52.4–47.6	(9.5)	(19.0)	TOR2
6 (830–880)	860(24,050) 861(24,051)	ValSer-LeuArg (AG/CT)	4.1–4.6	(33.3)	(27.5)	2.0–3.9	(5.9)	BJ03
7 (975–1,010)	1,001(24,474)	Arg-Met (G/T)	10.8–9.7	27.8–30.6	50.0–47.2	13.9–11.1	(5.6)	BJ04

CUHK: CUHK-Su10; HKU: HKU-39849; SIN1: SIN2500; SIN2: SIN2677; SIN3: SIN2679; SIN4: SIN2748; SIN5: SIN2774

Table 4 Pairwise Identity/Similarity Between Coronaviruses Based on the S Protein

	SARS_CoV												
SARS-CoV	100.0/100.0	MHV											
MHV	30.2/46.2	100.0/100.0	RSCoV										
RSCoV	30.3/45.6	79.2/87.4	100.0/100.0	BCoV									
BCoV	30.1/44.9	63.7/76.2	65.2/78.2	100.0/100.0	PHEV								
PHEV	29.8/44.8	63.5/76.5	65.0/78.1	82.3/89.6	100.0/100.0	HCoV-OC43							
HCoV-OC43	29.4/44.6	62.9/75.6	64.4/77.1	91.7/94.3	80.7/87.9	100.0/100.0	HECoV						
HECoV	29.9/44.6	63.4/76.3	65.1/78.4	97.9/98.9	82.2/89.6	91.2/94.1	100.0/100.0	HCoV-229E					
HCoV-229E	26.7/41.0	24.3/37.9	23.5/36.4	25.8/38.2	25.5/38.0	25.5/37.9	25.6/37.8	100.0/100.0	CCoV				
CCoV	24.4/38.4	21.8/33.4	21.8/32.6	23.4/34.8	23.4/35.6	25.4/37.7	24.1/35.9	38.8/52.1	100.0/100.0	AIBV			
AIBV	25.2/36.6	24.3/37.5	26.1/38.3	24.0/36.5	22.6/34.5	21.6/34.3	24.2/36.9	28.3/43.6	24.7/36.7	100.0/100.0	PEDV		
PEDV	23.5/35.3	24.7/38.5	21.0/31.6	24.9/38.7	24.5/39.5	25.0/38.9	24.5/37.4	42.4/56.7	43.7/60.7	25.6/38.8	100.0/100.0	FIPV	
FIPV	22.9/34.8	24.9/38.1	22.0/32.8	22.2/32.7	22.7/34.6	23.8/33.8	22.3/32.9	39.8/53.4	91.2/95.0	24.2/36.8	44.1/60.2	100.0/100.0	PTGV
PTGV	21.7/33.8	24.6/38.3	25.2/39.0	25.1/38.3	25.5/39.0	24.5/37.4	25.2/38.9	39.2/52.9	78.8/84.1	23.6/35.3	43.2/59.6	80.5/85.3	100.0/100.0

AIBV: avian infectious bronchitis virus; BCoV: bovine coronavirus; CCoV, canine coronavirus; FCoV: feline coronavirus; FIPV: feline infectious peritonitis virus; HCoV-229E: human coronavirus 229E; HCoV-OC43, human coronavirus OC43; MHV: murine hepatitis virus; PEDV: porcine epidemic diarrhea virus; PHEV: porcine hemagglutinating encephalomyelitis virus; PRCoV: porcine respiratory coronavirus; PTGV: porcine transmissible gastroenteritis virus; RCoV: rat coronavirus; RSCoV: rat sialodacryoadenitis coronavirus; HECoV: human enteric coronavirus; SARS-CoV: the human SARS-associated coronavirus; TCoV: turkey coronavirus.

We analyzed the physiochemical characteristics of adjacent regions of the eight non-synonymous mutations (Table 4). The results show that some subregions' characteristics are changed markedly due to the mutations taking place in them. It has been reported that clinical symptoms and severity differ in some SARS patients, and mutations of the SARS-CoV genome are suspected to be the underlying mechanism for these clinical variations. We believe the corresponding mutations should be located in the S protein, especially the S1 domain, because previous studies of MHV (murine hepatitis virus) show that some mutations in the S protein can induce changes in antigenicity, tissue tropism, cell–cell fusion and virulence (*8, 14*).

Sequences similarity analysis

A phylogenetic comparative analysis previously conducted by our research team placed the SARS-CoV outside the three known groups in the family of *Coronaviridae*. The unrooted phylogenetic tree and pairwise alignment indicated that SARS-CoV is closer to Group 2 than to the others (Figures 3 and 5). This result is different from those proposed for the other structural proteins of SARS-CoV (*11*). To extend these earlier analyses by our group, a pairwise alignment was performed among the S protein of coronaviruses. The pairwise alignment yielded similar results to that of the unrooted tree (Table 3). Based on the similarity observed among the S proteins of 13 coronaviruses, we determined that the N-terminal half is highly variable while the C-terminal half is relatively conserved. Five more homologous regions, including the TM region, could be identified. (Figure 4)

To find a conserved region involved in antigenicity is important for designing effective and sustainable vaccines, although caution should be paid to possible cross reactions or over response. Vaccines for preventing coronavirus infections, such as those that have been developed for treating chicken bronchitis caused by AIBV, are not always effective due to mutation of the virus (*15*). Our similarity analysis showed that the S1 region, which is responsible for the attachment of the virus to cells, has undergone significant alteration (Figure 4). Our efforts to locate conservative exterior regions of the S protein revealed that the subregion at amino acid position 1,191–1,200 appears be the most conserved. There are no matches between this peptide and adjacent regions with the human protein database. Thus, this conserved region and the other less conserved regions may be important for designing sub-fractional natural vaccines or genetically engi-

neered vaccines. We also identified four coiled-coil-like motifs in the C-terminal region with high probability (Figure 1E). Similar structures have been previously detected in some other coronaviruses (*16*). We believe that these regions are highly conserved and may be a potent drug target (*17*).

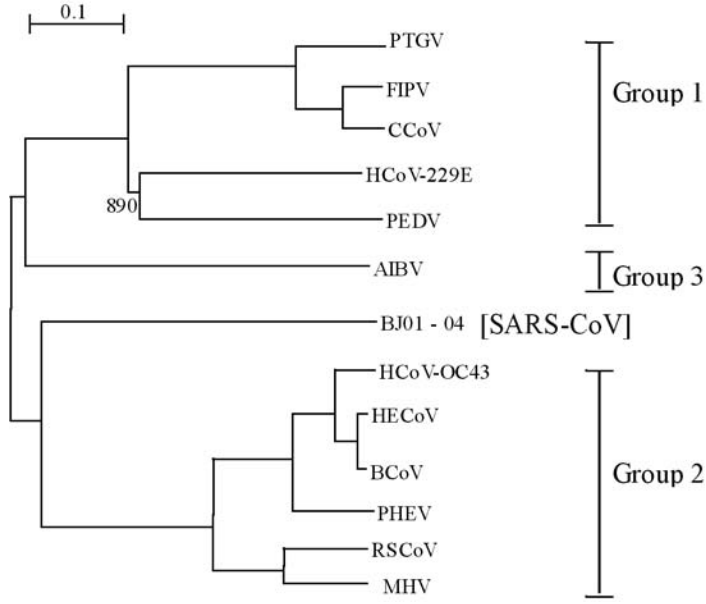

Fig.3 A phylogenetic tree of the coronaviruses based on the S protein.

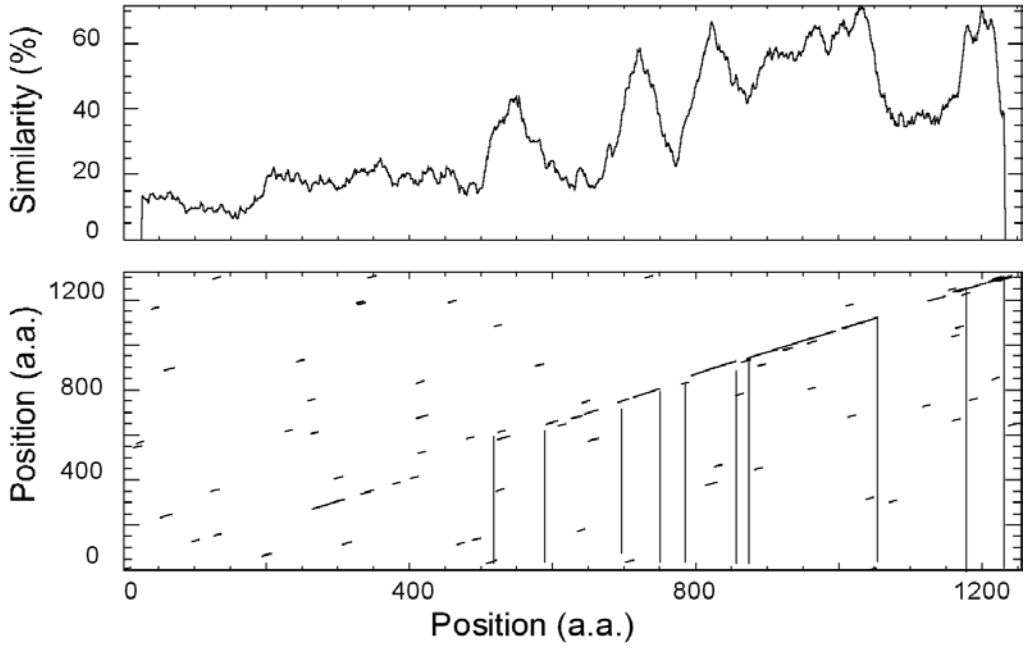

Fig.4 The similarity among the S proteins of 13 coronaviruses and dotplot result between MHV and SARS-CoV (BJ01 - 04).

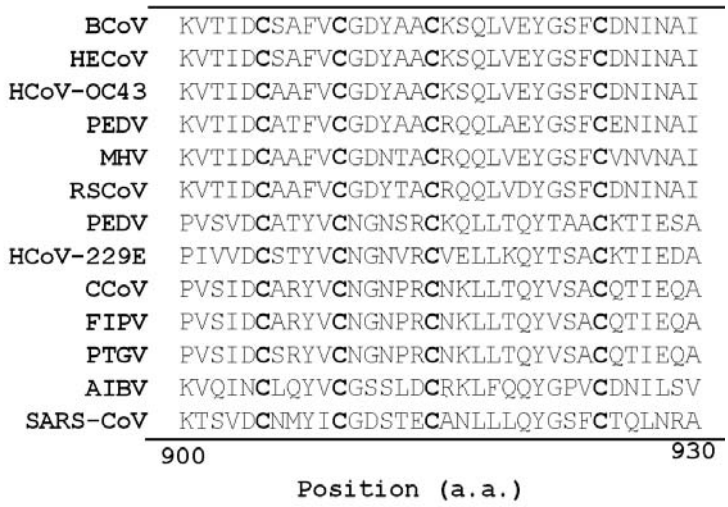

Fig.5 A typical Cys conserved region of coronaviruses.

In our similarity analysis, we paid particular attention to the presence of Cys, which could provide sites for disulfide bonds to link two peptides structurally. Among 39 cysteins in the S protein of SARS-CoV, 14 are highly conserved in comparison with other coronaviruses, especially the 4 adjacent cysteins that are perfectly aligned in all the coronaviruses we analyzed, locating from Condons 720 to 742 (Figure 5).

In addition to the known functions of the S protein in binding the virus to the host cellular receptor and inducing the fusion with the host cellular membrane, we believe that this large protein may have other functions in combination with other structural proteins or PUPs. No obvious connections, such as disulfide bonds, have been identified between the S and M proteins. However, the formation of a complex of the S protein and the M protein in the process of MHV assembly has been previously reported (*18*). The three cysteins in the M protein did not show any phenomena observed in other proteins suggesting that they contribute to any disulfide bond, but further studies are essential to understand the possible relationship between them (*11*).

We have found an interesting potential hypothesized connection between the S and E proteins. The E protein has been regarded as the main component for the viral inner or small envelope. In our study, we identified a strong hydrophobic region with a length of 23 amino acids, indicating the presence of a typical TM region in this small protein. The middle part has a LCAYCCM just outside the TM region in the predicted structure, which could form three disulfide bonds connecting the SCGSCCK near the C-terminus of the S protein. This can be interpreted as the possible signal pathway leading from the exterior of the virus through the S protein to the E protein and then to the inside of the viron. Biochemical and molecular biological studies are needed to support this finding.

Since the S protein is the main component of the spike as well as the largest structural protein in the virus, additional comparative analyses on the S protein involving more genomes will be essential for understanding its antigenicity, host range, and its possible signal transduction These studies will help to the diagnosis and vaccine development against SARS.

Materials and Methods

Genome sequences

The samples of the SARS patients have been described previously (*11*). Four isolates were prepared from Vero-E6 cell culture. Viral RNA preparation, RT-PCR and cloning were made according to general protocols at BGI (Beijing Genomics Institute, http://www.genomics.org.cn). The sequencing was performed by using MegaBACE 1000 (Amersham, New Jersey, USA). Sequences of the four complete SARS-CoV genomes (AY278488, AY279354, AY278490, and AY278487) have been deposited in GenBank (http://www.ncbi.nlm.nih.gov/). The complete and/or partial sequences of 1,355 RNA viruses and 13 coronaviruses (NC_003436, NC_002306, NC_003045, NC_002645, NC_001846, NC_001451, AY078417, AF207551, P36300, P10033, HOBOC43S, L07748, and AY278488) were downloaded from GenBank for comparative analysis.

Design and synthesis of peptides

The peptides were designed on the basis of the combined analyses of the proteins and were synthesized

by Hangzhou Zhongtai Inc. (Hangzhou, China).

ELISA test

Blood samples of two normal controls and nine SARS patients were provided by Beijing Plastic Surgery Hospital, Beijing Peoples' Hospital, and Beijing Tiantan Hospital, China. Peroxidase-conjugated mouse anti-human IgG and peroxidase-HRP (P6782) were purchased from Sigma (New Jersey, USA). Peptides (1 μg/mL, in 0.5 M carbonate buffer, pH 9.6) were dispensed into a 96-well microplate (100 μL/well) and then incubated at 4 $^\circ$C overnight. After being washed with PBS containing 0.5 M Tween-20, BSA (2 mg/mL) was added and the plates were incubated at 37 $^\circ$C for 1 h for blocking. The patient serum sample (10 μL), diluted with 100 μL of sample buffer, was added into each well and incubated at 37 $^\circ$C for 30 min. After being further washed with PBS-T, 100 μL mouse anti-human IgG was added and incubated at 37 $^\circ$C for 20 min. Finally, the wells were washed with PBS-T. The reaction was observed by adding the TMB solution as substrate, after incubation at 37 $^\circ$C for 10 min. The reaction was stopped by adding 50 μL 4 M sulphuric acid, and optical density at 450 nm (ref. 630 nm) was measured with an automatic ELISA reader (Multiskan Ascent, Finland).

Bioinformatic analysis

ORF and amino acid sequences were determined with ORF Finder (http://www.ncbi.nlm.nih.gov/gorf/gorf.html). The hypothesized physiochemical features of the putative protein were examined by using Compute PI/MW (http://us.expasy.org/tools/pi_tool.html). The glycosylation sites were predicted by NetNGlyc1.0 (http://www.cbs.dtu.dk/services/NetNGlyc) and NetOGlyc (http://www.cbs.dtu.dk/services/NetOGlyc/). The other modification sites of the S protein were predicted by Scanprosite (http://us.expasy.org/tools/scanprosite/). The electric charge and hydropathicity analyses were performed by the software of charge (http://www.hgmp.mrc.ac.uk/Software/EMBOSS) and toppred (http://bioweb.pasteur.fr/seqanal/interfaces/toppred.html). The antigenic determinants were predicted by Antigenic (http://www.hgmp.mrc.ac.uk/Software/EMBOSS). The structural domains were predicted with InterProScan (http://www.ebi.ac.uk/interpro/scan.html) and RPS_BLAST (http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi). The signal peptide and TM region were predicted by SignalP (http://www.cbs.dtu.dk/services/SignalP/) and TMHMM (http://www.cbs.dtu.dk/services/TMHMM/). The secondary structure was predicted by PSIPRED (http://bioinf.cs.ucl.ac.uk/psipred/ psiform.html) and COILS (http://www.ch.embnet.org/software). The phylogenetic analysis was performed with ClustalW1.8 (ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalW). The similarity among the coronaviruses was calculated and graphically represented by plotcon (http://www.hgmp.mrc.ac.uk/Software/EMBOSS) and dotmatcher (http://www.hgmp.mrc.ac.uk/ Software/EMBOSS). All the bioinformatic analyses were performed on supercomputer SUN 10K (SUN Microsystems Inc., California, USA).

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 are indebted to collaborators and clinicians from Peking Union Medical College Hospital, Beijing Plastic Surgery Hospital, Beijing Peoples' Hospital, Beijing Tiantan Hospital, National Center of Disease Control of China, and the Municipal Governments of Beijing and Hangzhou.

References

1. Jackwood, M.W., *et al.* 2001. Spike glycoprotein cleavage recognition site analysis of infectious bronchitis virus. *Avian Dis.* 45: 366-372.
2. Frana, M.F., *et al.* 1985. Proteolytic cleavage of the E2 glycoprotein of murine coronavirus: host-dependent differences in proteolytic cleavage and cell fusion. *J. Virol.* 56: 912-920.
3. Cavanagh, D., *et al.* 1986. Coronavirus IBV: partial amino terminal sequencing of spike polypeptide S2 identifies the sequence Arg-Arg-Phe-Arg-Arg at the cleavage site of the spike precursor propolypeptide of IBV strains Beaudette and M41. *Virus Res.* 4: 133-143.
4. Abraham, S., *et al.* 1990. Deduced sequence of the bovine coronavirus spike protein and identification of the internal proteolytic cleavage site. *Virology* 176: 296-301.
5. Garoff, H., *et al.* 1998. Virus maturation by budding. *Microbiol. Mol. Biol. Rev.* 62: 1171-1190.
6. Gallagher, T.M. and Buchmeier, M.J. 2001. Coronavirus spike proteins in viral entry and pathogenesis. *Virology* 279: 371-374.
7. Bos, E.C., *et al.* 1995. Mutational analysis of the murine coronavirus spike protein: effect on cell-to-cell

fusion. *Virology* 214: 453-463.

8. 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.
9. Marra, M.A., *et al.* 2003. The Genome sequence of the SARS-associated coronavirus. *Science* 300: 1399-1404.
10. Ruan, Y.J., *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.
11. Qin, E.D, *et al.* 2003. A complete sequence and comparative analysis of a SARS-associated virus (Isolate BJ01). *Chin. Sci. Bull.* 48: 941-48.
12. Rota, P.A., *et al.* 2003. Characterization of a novel coronavirus associated with severe acute respiratory syndrome. *Science* 300: 1394-1399.
13. Ignjatovic, J. and Galli, L. 1993. Structural proteins of avian infectious bronchitis virus: role in immunity and protection. *Adv. Exp. Med. Biol.* 342: 449-453.
14. Leparc-Goffart, I., *et al.* 1997. Altered pathogenesis of a mutant of the murine coronavirus MHV-A59 is associated with a Q159L amino acid substitution in the spike protein. *Virology* 239: 1-10.
15. Kusters, J.G., *et al.* 1987. Molecular epidemiology of infectious bronchitis virus in The Netherlands. *J. Gen. Virol.* 68: 343-352.
16. de Groot, R.J., *et al.* 1987. Evidence for a coiled-coil structure in the spike proteins of coronaviruses. *J. Mol. Biol.* 196: 963-966.
17. Rapaport, D., *et al.* 1995. A synthetic peptide corresponding to a conserved heptad repeat domain is a potent inhibitor of Sendai virus-cell fusion: an emerging similarity with functional domains of other viruses. *Embo. J.* 14: 5524-5531.
18. Opstelten, D.J., *et al.* 1993. Complex formation between the spike protein and the membrane protein during mouse hepatitis virus assembly. *Adv. Exp. Med. Biol.* 342: 189-195.