Chinese Science Bulletin 2003 Vol. 48 No. 13 1277—1287

Bioinformatical study on the proteomics and evolution of SARS-CoV

LIU Shuqun*, GUO Tao*, JI Xinglai & SUN Zhirong

Institute of Bioinformatics, Department of Biological Sciences and Biotechnology, Tsinghua University, Beijing 100084, China
* The two authors contributed equally to this work.
Correspondence should be addressed to Sun Zhirong (e-mail: sunzhr@mail.tsinghua.edu.cn)

Abstract **A novel coronavirus has been identified as the causative agent of the severe acute respiratory syndrome (SARS). For all the SARS-CoV associated proteins derivated from the SARS-CoV genome, the physiochemical properties such as the molecular weight, isoelectric point and extinction coefficient of each protein were calculated. The transmembrane segments and subcellular localization (SubLocation) prediction and conserved protein motifs search against database were employed to analyze the function of SARS-CoV proteins. Also, the homology protein sequence alignment and evolutionary distance matrix calculation between SARS-CoV associated proteins and the corresponding proteins of other coronaviruses were employed to identify the classification and phylogenetic relationship between SARS-CoV and other coronaviruses. The results showed that SARS-CoV is a novel coronavirus which is different from any of the three previously known groups of coronviruses, but it is closer to BoCoV and MHV than to other coronaviruses. This study is in aid of experimental determination of SARS-CoV proteomics and the development of antiviral vaccine.**

Keywords: SARS, SARS-CoV, conserved protein motif, sequence alignment, evolution.

DOI: 10.1360/03wc0276

The first case of atypical pneumonia, referred to as the severe acute respiratory syndrome (SARS) was identified in Guangdong Province in China in November, 2002. Then, it has spread to several countries and regions such as Hong Kong, Vietnam, Singapore, Canada and Taiwan. Cumulative number of reported cases of SARS has been 8221, and the number of deaths has been 735 in the world up to May 28, 2003. Among them there are 5332 reported cases of SARS and 321 deaths in China, and the mortality rate has increased to 15% from the early 4% estimated by the World Health Organization (WHO)[1)].

WHO stated that under the efforts of 13 laboratories in China, Germany, Canada, etc., a novel coronavirus was identified as the causative agent of the SARS on April 16, 2003, and this virus was named SARS Coronavirus (SARS-CoV)[1—3]. The first completed genome and gene map of SARS-CoV were released by scientists in Canada[3] on April 12. Subsequently, scientists in China, Hong Kong, USA and Singapore released in succession the completed SARS-CoV genomes determined by themselves[4—7]. There were 27 items of SARS-CoV associated nucleotide sequences deposited in GenBank database (http://www.ncbi.nlm.nih.gov/) and 14 of them were completed SARS-CoV genomes up to May 30, 2003.

The classification status of coronavirus is order *Nidovirales*, family *Coronaviridae*, gnenus *Coronavirus*. It replicates in the cytoplasm of human and animal host cells and causes respiratory and enteric diseases in host. Coronaviruses contain a positive-stranded RNA with approximately 30000 nucleotides, and feature the largest viral RNA genome known to date. The coronaviruses known to date are divided into three groups (groups 1, 2 and 3) according to their serotype, groups 1 and 2 contain mammalian viruses, while group 3 contains only avian viruses. Within each group, coronaviruses are classified into distinct species by host ranges, antigenic relationships and genomic organization. Each species of coronavirus typically infects only one host species and is fastidious in cell culture. The viruses can cause severe disease in many animals, including avian infectious bronchitis, feline infectious peritonitis, porcine gastroenteritis, etc., and often cause significant lose in stock raising[8]. The two known human coronavirus HCoV-229E and HCoV-OC43 belong to respectively group 1 and group 2, and are responsible for about 30% of influenza and mild upper respiratory tract illness[9]. Upon coronavirus entry into an appropriate host cell, the 5′ most open reading frame ORF1ab of virus genome is translated into a large polyprotein that is cleaved by viral-encoded proteases to release several protein products[5], including a 3C-like proteinase ($3CL^{pro}$), an RNA-dependent RNA polymerase (Rep), an ATPase helicase (Hel), etc. Out of them Rep and Hel are responsible for replicating the viral genome and generating nested transcripts that are used in the synthesis of the viral proteins by host cell ribosome. There are four primary structural proteins including the nucleocapsid protein N, spike protein S, membrane protein M, and small envelope protein E. All of them are glycoproteins. The full-length replicated RNA plus strand first assembles with the N protein. This RNA-protein complex then associates with M and N proteins embedded in the membranes of the ER and new virus particles are formed as the nucleocapsid complexes which bud into the ER. The virus then migrates through the Golgi and eventually exits the cell likely by exocytosis[10]. It is worth pointing out that the S membrane glycoprotein is functionally important for defining host range and tissue tropism because the virus entry into cell requires the interaction between S protein and its special

1) Cumulative number of reported cases of SARS is published at http://www.who.int/csr/sars/

cellular receptor[8].

The sequence lengths of the 14 completed genomes of SARS-CoV deposited in NCBI GenBank range from 29712 to 29751, and the nucleotide difference between these genomes is about 0.1%, the GC content is about 41%. The genomic organization is typical of coronaviruses, with the characteristic gene order of 5′ *rep*, *S*, *E*, *M* and *N* genes[5,6], of which the SARS-CoV *rep* gene comprises approximately two-thirds of the genome, it includes two open reading frames, ORF1a and ORF1b, which are predicted to encode two polyproteins undergoing co-translational proteolytic processing and generate many protein products. It also predicted that the regions located between S and E and between M and N encode a number of non-structural proteins that vary widely among different coronaviruses with unknown function[4—6].

In this study, all the SARS-CoV associated proteins were subject to bioinformatical analysis. The physiochemical properties of each protein such as molecular weight, isoelectric point and extinction coefficient were calculated. The transmembrane segments and subcellular localization prediction and conserved protein motifs search against databases were employed to analyze and predict the function of each SARS-CoV protein. Furthermore, the evolution distance between SARS-CoV proteins and their corresponding homology proteins of other coronaviruses were calculated and analyzed.

1 Materials and methods

The SARS-CoV associated protein sequences were searched in SWISS-PROT database (http://www.ebi.ac.uk/swissprot/) with the keyword "SARS". The protein sequences deposited in SWISS-PROT refer to almost all the corresponding nucleotide sequences in GenBank, so the protein sequences obtained from SWISS-PROT is more typical than those from GenBank[11]. There are 10 SARS-CoV associated protein sequences in SWISS-PROT up to May 28, 2003, including structural proteins S, E, M and N, and function unknown proteins X1, X2, X3 and X4. Through comparing each of these proteins with the corresponding ORF of the completed SARS-CoV genome AY274119 in GenBank, we found that the proteins encoded by ORF9-11 and ORF14 have not been deposited into SWISS-PROT. All the SARS-CoV associated proteins including the proteins from SWISS-PROT and the putative proteins from ORF are listed in Table 1. The studies on other coronaviruses have revealed that the ORF1ab occupies about two-thirds of virus genome, and it can be divided into two ORFs: ORF1a and ORF1b[12]. For SARS-CoV, the polyprotein is synthesized by a −1 ribosomal frameshift at the predicted slippery site (13392—13398, 5′-UUUAAAC-3′) where the ORF1a and ORF1b are overlapped. During the translational process of ORF1a and ORF1b, the translated polyprotein is autocatalytically precessed to initially yield mature viral proteases PLP^{pro} (papain-like proteinase) and $3CL^{pro}$ (chymotrypsin-like proteinase), then generated Rep (RNA-dependent RNA polymerase), Hel (ATPase-Helicase) and other protein products whose functions have not been well characterized. Through analysis of the cleavage sites of ORF1ab, we got 14 putative protein products (Table 2), and because these proteins may exist in the actual life cycle of SARS-CoV, these 14 proteins but not the polyprotein corresponding to ORF1ab are subject to bioinformatical calculation and analysis.

First, the amino acid composition, charge distribution, and the possible repetitive sequence motif of each SARS-CoV protein were analyzed by program SAPS[13] (Statistical Analysis of Protein Sequence). Then, the isoelectric point (pI) and extinction coefficient at 280 nm

Table 1 Physiochemical properties of SARS-CoV associated proteins

Protein	Database ID [b]	Length/aa	MW/kD	Extcoef [c]/mol • L^{-1} • cm^{-1}	pI [d]	Functional classification
ORF1AB	P59641	7073	790.3	866830	6.2	polyprotein
S	P59594	1255	139.1	134050	5.4	structural
X1	P59632	274	30.9	50690	5.7	structural [e]
X2	P59633	154	17.8	9770	11.2	non-structural [e]
E	P59637	76	8.3	5300	6.3	structural
M	P59596	221	25.1	51530	9.8	structural
X3	P59634	63	7.5	8250	4.4	N/A [f]
X4	P59635	122	13.9	6760	8.0	structural [e]
ORF9	AY274119	44	5.3	7079	3.4	structural [e]
ORF10	AY274119	39	4.3	360	8.0	N/A [f]
X5	AY274119	84	9.6	21210	9.5	non-structural [e]
N	P59595	422	46.0	42530	10.5	structural
HP5 [a]	P59636	98	10.8	1280	4.7	N/A [f]
ORF14	AY274119	70	7.8	8490	6.3	N/A [f]

a) Hypothetical protein 5; b) P596xx is Swiss-Prot ID; AY274119 is GenBank ID; c) Extcoef denotes molecular extinction coefficient at 280 nm; d) pI denotes isoelectric point; e) the putative function of proteins in this study (for details see the text); f) protein functions are not available.

Table 2 Physiochemical properties of the putative proteins products encoded by ORF1ab

ORF product	ORF	Start-END	Length/aa	MW/kD	Extcoef [f]/mol • L^{-1} • cm^{-1}	pI [g]
LP [a]	ORF1A	1—179	179	19.6	13430 13370	5.2
P65 [b]	ORF1A	180—818	639	70.7	63150	6.0
PLP [c]	ORF1A	819—3239	2421	269.7	28110	5.4
$3CL^{pro}$	ORF1A	3240—3547	308	33.8	31870	6.2
HD2/NSP3	ORF1A	3548—3836	289	33.0	59180	8.9
NSP4 [d]	ORF1A	3837—3919	83	9.3	5870	5.0
NSP5	ORF1A	3920—4117	198	21.9	20460	6.7
NSP6	ORF1A	4118—4229	112	12.4	12270	9.2
GFLP [e]	ORF1A	4230—4369	140	14.8	12870	6.3
Rep	ORF1A, ORF1B	4370—5301	932	106.5	127310	6.0
Hel	ORF1B	5302—5902	601	66.9	62150	8.2
NSP11	ORF1B	5903—6429	527	59.9	89000	7.3
NSP12	ORF1B	6430—6775	346	38.5	31450	4.9
NSP13	ORF1B	6776—7073	298	33.0	56710	7.8

a) Leader protein; b) this protein is corresponding to the P65 of MHV; c) Papain-like proteinase; d) NSP denotes function unknown protein; e) growth factor like protein; f) Extcoef denotes molecular extinction coefficient at 280 nm; g) pI denotes isoelectric point.

of each protein were calculated by PI[1] and EXTCOEF[14] program, respectively. TMAP[15] and TMHMM 2.0[16] were employed to predict the transmembrane segments or transmembrane helical regions of the proteins, from which the subcellular localizations of these proteins were predicted. The HMMPFAM[17] and BLIMPS[18] were used to compare each of SARS-CoV protein sequence against the HMM and BLOCKS databases to search for possible matching conserved protein motifs. Finally, the PSIBLAST program[19] at bioinformatics platform of Workbench[14] was used to search for homology protein sequences of each SARS-CoV protein through comparing against databases at non-redundant protein database (SDSC) (the matrix is BLOSUM80 and the maximum number of rounds is 6). The CLUSTALW program[20] was used to performs multiple sequence alignments of the homology sequences, then the unrooted phylogenetic trees were drawn by DRAWTREE and DRAWGRAM program[21], and the evolutionary distance matrixes based on the above protein alignments were computed by PROTDIST program[2].

2 Results and discussions

Like the proteins of other coronaviruses, the SARS-CoV associated proteins are divided into structural protein and non-structural protein. The former includes proteins S, E, M and N that are responsible for making up of protein envelope of the virus, the latter includes $3CL^{pro}$, Rep, Hel, etc. that are responsible for the proteolytic processing of polyproteins ($3CL^{pro}$) and replication and translation of viral RNA (Rep and Hel). However, there are many putative proteins derivated from ORFs of SARS-CoV genome, for example, proteins X1, X2, X3, X4 and X5, and the functions of these proteins are unknown. All the SARS-CoV associated proteins were subject to analysis and calculation by bioinformatics tools. The sequence length, molecular weight, isoelectric point and extinction coefficient of each protein are listed in Tables 1 and 2.

(ⅰ) Analysis of the protein products encoded by ORF1ab. The polyprotein encoded by ORF1ab generates about 14 protein products (Table 2) through cotranslational proteolytic processing. Among them the leader protein is an acidic protein with molecular weight of 19.6 kD. The positive-charged amino acids (K+R) content is 10.1%, and the negative-charged amino acids (E+D) content 14.5%. Both TMAP and TMHMM predict that there is no transmembrane region in this protein. BLIMPS analysis of the leader protein reveals that the segment 25—49 matches to the conserved sequence motif IPB001407D that is from Influenza RNA-dependent RNA polymerase subunit PB1 in BLOCKS database. PSIBLAST search reveals that partial segments of this protein possess about 30% identities with partial segments of intrinsic factor-vitamin B12 receptor, cubilin, Cation transporter/ATPase and Aminopeptidase N, respectively. P65 is a weakly acidic protein with molecular weight of 70.7 kD. TMAP predicts the existence of two potential transmembrane regions spanning residues 400—424 and 467—486, but TMHMM predicts no transmembrane helical segment. BLIMPS search reveals that the segment 508—545

1) Program by Dr. Luca Toldo, developed at http://www.embl-heidelberg.de

2) Felsenstein, J. PHYLIP (Phylogeny Inference Package) version 3.5c, 1993, Distributed by the author. Department of Genetics, University of Washington, Seattle.

matches to conserved protein motif IPB001156 (Transferrin) in BLOCKS database. HMMPFAM search reveals that segments 142—185 and 291—351 match to PF05256 (Uncharacterised protein family) and PF00087 (Snake toxin) in HMM database, respectively. PSIBLAST analysis of this protein exhibits that some partial segments of P65 share about 30% sequence identities with proteins Drosophila melanogaster, extracellular matrix protein B, mitochondrial ribosomal protein, Hypothetical 12.7 kD protein in 16S-23SDNA spacer and Listeria innocua. PLP^{pro} is a strongly acidic protein with molecular weight of 269.7 kD and comprises abundant Cys (3%), the sequence region 111—183 (DEEEEDDAECEEEEIDETCEHEYGTEDDYQGLPLEFGASAETVRVEEEEEEDWLDDTTEQSEIEPEPEPTPEE) is a high scoring negative charge segments. TMAP and TMHMM analyses both indicate the presence of seven transmembrane helical segments, while the large N-terminus 1—1390 region is oriented on the surface of the virus particle or inside the lumen of the ER/Golgi, and C-terminus 112 residues are oriented inside the virus particle or cytoplasm. HMMPFAM analysis of PLP^{pro} reveals that segments 44—182, 218—326, 790—1065, 1216—1416 and 2152—2312 regions match separately to conserved protein motifs of PF05066 (RNA_pol_delta), PF01661(A1pp), PF01831 (Peptidase_C16), PF03649 (UPF0014), PF01891(CbiM) in HMM database. BLIMPS search results show that some partial segments of PLP^{pro} match to conserved protein sequences IPB002589 (Domain of unknown function DUF27), IPB001509 (NAD dependent epimerase/dehydratase family), IPB001394 (Ubiquitin carboxyl-terminal hydrolase family 2), IPB000135 (High mobility group proteins HMG1 and HMG2), PR00375 (Huntingtin signature) in BLOCKS database, respectively. PSIBLAST search indicates that the sequence identities between SARS-CoV PLP^{pro} and the corresponding proteins of Murine hepatitis virus (MHV), Bovine coronavirus (BoCoV), Avian infectious bronchitis virus (AIBV), Human coronavirus 229E (HCoV-229E) and Transmissible gastroenteritis virus (TGEV) range from 30% to 40%. These analytical results of PLP^{pro} indicate that it most likely functions as hydrolase. $3CL^{pro}$ is an acidic protein with molecular weight of 33.8 kD. It contains two repetitive segments that repeat two times: TTTLN located at positions 24—28, 224—228, and AGTD located at 173—176 and 194—197. TMHHMM fails to predict any existence of transmembrane helical region, and BLIMPS and HMMPFAM search yield no matching conserved protein motifs. PSIBLAST analysis of $3CL^{pro}$ shows that the sequence identities between $3CL^{pro}$ of SARS-CoV and the corresponding proteins of MHV, BoCoV, AIBV, HCoV-229E and AIBV range from 40% to 50%, and $3CL^{pro}$ also shares 43.8% sequence identity with the segment 2879—3180 of Hydrolase, indicating that $3CL^{pro}$ can carry out the function of hydrolysis of protein. Rep is an acidic protein with molecular weight of 106.5 kD. TMAP and TMHMM predict no transmembrane region. HMMPFAM analysis of Rep yields a matching conserved sequence family PF00680 in HMM database, and PF00680 is derived from RNA dependent RNA polymerase. PSIBLAST search results show that the sequence of Rep share about 60% identities with the sequences of corresponding Reps of other coronaviruses. Hel is a basic protein with molecular weight 66.9 kD. TMAP and TMHMM predict no existence of transmembrane segment. HMMPFAM search yields a match to PF01443 (Viral_helicase1 Viral (Superfamily 1) RNA helicase). PSIBLAST search results indicate that the sequence identities between SARS-CoV Hel and the corresponding helicases of other coronaviruses are about 60%. The best-conserved motif is a glycine-rich region located at 282—288, which typically forms a flexible loop between a β-strand and an α-helix. This loop interacts with one of the phosphate groups of the nucleotide. This sequence motif is generally referred to as the "P-loop"[22]. The analyses above reveal the strongly conserved property of $3CL^{pro}$, Rep and Hel. HD2/NSP3 is a basic protein with molecular weight of 33.0 kD. TMHMM analysis predicts the presence of seven-transmembrane α-helices, with the N-terminus residues 1—11 oriented inside virus particle, and the C-terminus residues 236—290 located outside the virus particle, in which the segment 209—219 (CIMLVYCFLGYCCCCYFGLFC) is a sequence containing abundant Cys. PSIBLAST analysis reveals that HD2/NSP3 of SARS-CoV shares about 30% sequence identities with the corresponding proteins of other coronaviruses. HMMPFAM search reveals that the segment 43—140 of HD2/NSP3 matches to the conserved protein motif PF00420 (NADH-ubiquinone/plastoquinone oxidore) in HMM database. BLIMPS search reveals that some partial segments of HD2/NSP3 match to the conserved sequence motifs IPB001169 (Integrin beta, C-terminus), IPB002561 (Filovirus glycoprotein), IPB000832 (G-protein coupled receptors family) and IPB002091 (Aromatic amino acids permease) in BLOCKS database. The results above imply that HD2/NSP3 might be a member of G-protein coupled receptors family that is inserted into the envelope of the virus and functions as signal transduction. NSP4 is an acidic protein with molecular weight of 9.3 kD, and it contains two segments that repeat two times: VLLS at positions 12—15 and 58—61, LLSVL at positions 13—17 and 55—59. NSP4 shares about 40% sequence identities with the corresponding proteins of other coronaviruses. TMHMM prediction yields no transmembrane helical region. BLIMPS and HMMPFAM searches fail to identify any matching conserved sequence motifs. NSP5 is a neutral protein with molecular weight of 21.9 kD, and

no transmembrane segment is predicted. HMMPFAM search indicates the segment 16—117 of NSP5 matches to PF04233 (Phage Mu protein F like protein) and segment 29—140 matches to PF04696 (pinin/SDK/memA/ protein conserved region) in HMM database. BLIMPS search reveals some partial segments of NSP5 match to conserved sequence motifs IPB003660 (HAMP domain), PR00331 (Haemagglutinin HA2 chain signature) and PR00329 (Haemagglutinin HA1/HA2 chain signature) in BLOCKS database. Sequence alignments of SARS-CoV NSP5 and the corresponding proteins of other coronaviruses reveal 40% sequence identities between them. NSP6 is a strongly basic protein with molecular weight of 12.4 kD. TMAP and TMHMM predict no transmembrane segment. BLIMPS and HMMPFAM searches fail to identify matching conserved sequence motif. PSIBLAST analysis indicates that the sequence identities between NSP6 of SARS-CoV and the corresponding proteins of other coronaviruses are about 45%. GFLP (Growth factor like protein) is an acidic protein with molecular weight of 14.8 kD, and no transmembrane region is predicted. BLIMPS search reveals that segment 73—86 of GFLP matches to IPB000315 (B-box zinc finger superfamily) and the segment 36—74 matches to IPB003854 (Gibberellin regulated protein) in BLOCKS database. The sequence identities of GFLPs between various coronaviruses are about 55%. No transmembrane regions are predicted in NSP11, NSP12 and NSP13. BLIMPS search reveals that the segments 101—117 and 247—257 of NSP11 match to conserved sequence motifs PR00059 (Ribosomal protein L6 signature) and IPB001608 (Uncharacterized pyridoxal-5'-phosphate dependent enzyme family) in BLOCKS database, respectively, and the segment 179—215 of NSP13 matches to IPB000903 (Myristoyl-CoA: protein N-myristoyltransfe- rase). HMMPFAM search reveals that the segment 50—213 of NSP13 matches to PF01728 (FtsJ-like methyltransferase) in HMM database. For NSP12, no matching conserved sequence motif is found in HMM and BLOCKS databases. The sequence identities of NSP11, NSP12 and NSP13 between various coronaviruses are about 50%, 40% and 60%.

The results above indicate that the protein products encoded by SARS-CoV ORF1ab are very conservative among various coronaviruses. The matching conserved sequence families searched by BLIMPS, HMMPFAM and PSIBLAST are mainly from protease, which implies that these proteins encoded by ORF1ab play various catalytic function roles during the virus life cycle.

(ii) Analysis of structural proteins of SARS-CoV. The S protein of coronaviruses is a large membrane glycoprotein. The mature S protein is inserted in the viral envelope with the majority of the protein exposed on the surface of viral particles and forms the primary viral surface projection. It is believed that the S trimer makes this virus family look like a corona structure under an electron microscope. The S proteins of majority of coronaviruses are cleaved into S1 and S2 subunits after synthesis. The S1 subunit is responsible for recognizing and binding of cellular receptors while the integral membrane S2 subunit is required to mediate fusion of viral and cellular membrane. For these reasons, the virulence and host cellular specificity of coronavirus are defined by S protein[8]. The S protein of SARS-CoV comprises 1255 amino acids and it is a strong acidic protein with molecular weight of 139.1 kD, pI 5.4, and there is a repetitive segment FNGLT at positions 529—533 and 837—841. TMHMM analysis of S protein predicts a transmembrane helical region near the C-terminus at residues 1196—1218, with the N-terminus segment 1—1195 oriented on the surface of viral particle, and C-terminus hydrophilic region 1219—1255 oriented inside the viral particle, in which the segment 1217—1236 (CCMTSCCSCLKGACSCGSCC) is rich in Cys. HMMPFAM search reveals that the region 75—609 of SARS-CoV S protein matches to the conserved coronavirus S1 domain PF01600 in HMM database, and the region 641—1247 matches to conserved coronavirus S2 domain PF016001 in HMM database. The feature of the cleavage site of coronavirus S protein usually is a basic amino acid sequence RRXRR or RXRR; however, the S protein of SARS-CoV lacks such basic cleavage site, implying that it is probably not cleaved into S1 and S2 subunits and just forms S1 and S2 domains. Through searching for the homology sequences of coronavirus S proteins with PSIBLAST and aligning them with CLUSTALW, we found that the S2 domain of SARS-CoV is more conservative than the S1 domain because the former shares 18%—21% sequence identities with the S1 subunits of other coronaviruses, and the latter shares 34%—35% sequence identities with S2 subunits of other coronaviruses. This result can interpret well the functions of S protein: S1 domains are responsible for identification and binding of special receptors, various coronaviruses use different receptors and infect different host cells, which can explain the non-conserved property of S1 domains; while S2 domains are responsible for mediating fusion of viral and cellular membrane, such unique function explains the conserved property of S2 domains. The current mechanism of protein-mediated membrane fusion proposes the collapse model: first, the interaction of S1 subunit with the special cellular membrane receptor induces conformational change of S1, which in succession induces the structural changes of S2——some of the amphipathic α-helices in the carboxyl half of S2 collapse into coiled-coils, thus bringing a fusion peptide toward the transmembrane domain, resulting in cellular and viral membrane fusion[8]. We predicted the secondary structure of SARS-CoV S protein using eight different secondary

structure prediction methods provided by the bioinfomatics platform of Workbench. The prediction results reveal nine successive α-helices at residue positions 750—1010 (Fig. 1). Furthermore, the relative hydrophobic and hydrophilic character of amino acids in S protein is calculated by GREASE program[23]. The hydropathicity profile in 885—950 and 973—990 regions shows a distinct property that hydrophilic and hydrophobic amino acids present alternately, suggesting that such regions may be the amphipathic α-helices whose conformation could collapse into coiled-coils. It has recently been shown that for HCoV-229E virons, residues 417—546 are required for

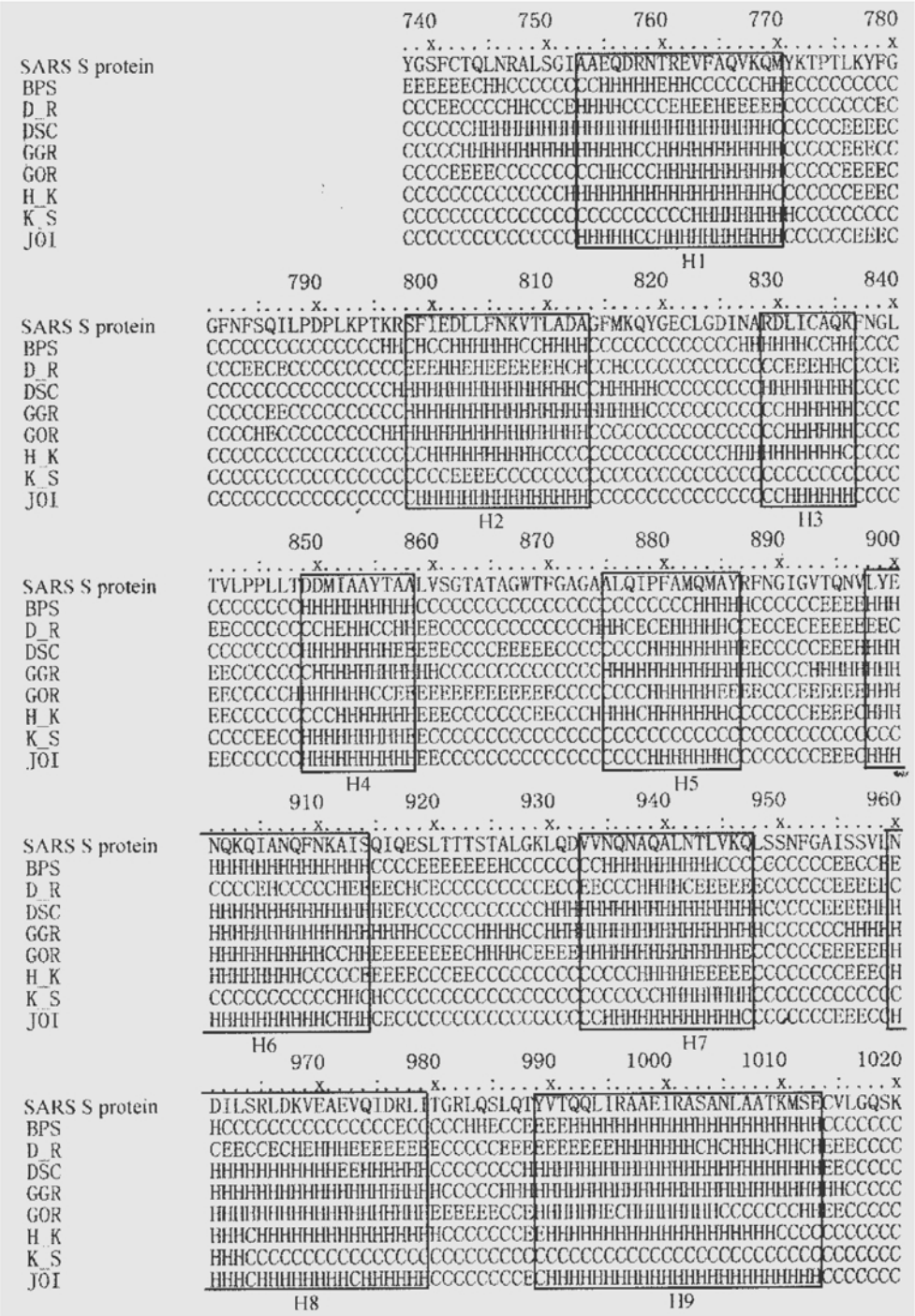

Fig. 1. The secondary structure of S protein predicted by eight protein secondary structure prediction methods. The segments in the boxes indicate the α-helical regions H1—H9 predicted by more than three prediction methods.

binding to the cellular receptor aminopeptidase N[24]. Through sequence alignment between S proteins of SARS-CoV and HCoV-229E, we found that the residues 492—651 of SARS-CoV S protein correspond to residues 417—547 of HCoV-229E S protein. However, further experiments are required for identifying if residues 492—651 of SARS-CoV S protein are necessary for receptor binding.

E protein of SARS-CoV is a small acidic protein comprising 76 amino acids with molecular weight of 8.3 kD. Both TMAP and TMHMM predict a hydrophobic transmembrane helix at residues 12—34, with residues 1—11 oriented inside viral particle and hydrophilic segment 35—76 on the surface of viral particle. HMMPFAM search reveals that the complete sequence of E protein matches to the conserved sequence motif PF02723 (NS3/small envelope protein E) in HMM database. BLIMPS search also reveals that the segment 36—67 of E protein matches to IPB003873 (conserved coronavirus small envelope protein E) in BLOCKS. However, PSIBLAST search reveals that there are only two coronaviruses whose E proteins are similar to that of SARS-CoV in sequence: the murine hepatitis virus MHV and Rat sialodacryoadenitis coronavirus RSCoV, the sequence identities between E protein of SARS-CoV and those of the two coronaviruses are 23% and 24%, respectively, suggesting that E protein is not conservative in evolution. Furthermore, studies have revealed that the E protein in porcine transmissible gastroenteritis coronavirus is essential for virus replication[25], while in MHV it has been shown that although deletion of gene E reduces virus replication by more than ten thousand fold, the virus still can replicate[26]. Further studies are required for ascertaining if E protein of SARS-CoV plays a functional role in virus replication in addition to attaching itself to the virus envelope.

M protein is a basic protein comprising 221 amino acids with molecular weight of 25.1 kD. The residues 44—97 near the N-terminus make up of a strongly hydrophobic region, and a repetitive segment LVIG presents at positions 21—24 and 137—140. PSIBLAST and alignment results reveal a strongly conserved sequence (WSFNPE) of SARS-CoV M protein at positions 109—114, while the sequence identities between M proteins of various coronaviruses are about 30%. TMHMM predicts three transmembrane α-helical segments located at residues 15—37, 50—72 and 77—99, with N-terminus residues 1—14 on the surface of virus particle and C-terminus hydrophilic domain 100—221 on the inside of the virus particle. The segment inside the virus particle is rich in positive charged amino acids, and is believed to interact with the nucleocapsid or virus RNA. HMMPFAM and BLIMPS searches reveal that the complete sequence of E protein of SARS-CoV matches to conserved protein family of coronavirus M matrix/glycoprotein PF01635 and IPB002574, indicating that the M protein of SARS-CoV is a typical coronavirus M protein. BLIMPS search yields matches of some partial segments of M protein to IPB001717 (Anion exchanger family), IPB001425 (Bacterial rhodopsin) and IPB001712 (Bacterial export FHIPEP family) in BLOCKS database, implying that M protein may play function roles in anion exchange and signal transduction. Studies on other coronaviruses have revealed that the association of the S protein with the M protein is an essential step in the formation of the viral envelope and in the accumulation of both proteins at the site of virus assembly[10]. The pI of S protein of SARS-CoV S is 5.4, that of M protein is 9.8. It is believed that the electrostatic interaction between S and M contributes to virus assembly.

Protein N is a strongly basic protein comprising 422 amino acids with molecular weight of 46.0 kD, a repetitive segment SRGNS is found at positions 191—195 and 203—207. Both TMAP and TMHMM predict no transmembrane region. PSIBLAST and alignment analysis reveal a strongly conserved segment PRWYFYYLGTGP located at position 107—118, and the function of this segment is unknown. Furthermore, there is a segment KKKKTDEAQPLPQRQKKQ at position 373—390 in SARS-CoV N protein that is rich in Lys, which matches to the bipartite nuclear targeting signal QDOC50079 in PFAM database, and is responsible for nuclear location and interaction. A hydrophilic segment (SRGGSQASSRSSSRSRGNSRNSTPGSSRGNS) that is rich in Ser is located at position 177—207. It is not clear if such segment interacts with nuclear. BLIMPS search indicates that the complete sequence of SARS-CoV N protein matches to IPB001218 (Coronavirus nucleocapsid protein); in addition, some partial sequences of N protein match to IPB002565 (Orbivirus NS3), IPB000096 (Serum amyloid A protein), IPB001677 (Transferrin binding protein), and IPB000689 (UbiH/ COQ6 monooxygenase family) in BLOCKS database. The sequence identities between the N proteins of SARS-CoV and other coronaviruses are about 30%.

(iii) Analysis of function unknown proteins of SARS-CoV. The putative protein X1 from SARS-CoV genome is an acidic protein comprising 274 amino acids with molecular weight of 30.9 kD. The analysis of X1 sequence reveals a strongly hydrophobic region FICNLLLLFVTIYSHLLLVAAGMEAQFLYLYALIYFL located at 79—115. Both TMAP and TMHMM predict three transmembrane helical regions located at residues 34—56, 77—99 and 103—125, with N-terminus residues 1—33 on the surface of virus particle and C-terminus residues 126—274 on the inside of the virus particle. PSIBLAST search yields no corresponding homology sequence of other coronaviruses, and HMMPFAM analysis fails to identify any matching conserved sequence family in

HMM database, while BLIMPS search reveals that the segment located at 231—252 matches to PR01542 (Foot-and- mouth disease virus VP1 coat protein signature), and segment at 67—87 matches to the conserved protein motif PR00699 (*C. elegans* integral membrane protein Srb signature), implying that X1 of SARS-CoV may be a structural protein inserted in the viral envelope.

X2 is a strong basic protein comprising 154 amino acids, and the proportion of positive-charged amino acid (K+R) content to negative charged amino acid (D+E) content is 22∶1. A repetitive small segment SLLK at positions 28—31 and 95—98 is found in X2. PSIBLAST search fails to identify any corresponding homology sequences. A C-terminal segment KKVSTNLCTHSFR-KKQVR located at residues 137—154 is found to match to bipartite nuclear localization signal PS50079|NLS_BP in PFAM database. Both TMAP and TMHMM predict no transmembrane region. The results above suggest that X2 may function like N protein (for details see discussion for N protein), which is oriented inside virus particle, interacts with virus RNA, and should be essential for virus assembly.

X3 is a small acidic protein comprising 63 amino acids with molecular weight of 7.5 kD. PSIBLAST search fails to identify any matching homology sequence, and both HMMPFAM and BLIMPS searches yield no matching conserved sequence family. Although TMHMM fails to predict transmembrane α-helix, TMAP analysis predicts a transmembrane segment located at position 9—37 with N-terminal residues 1—8 on the outside of virus particle. We cannot predict the function of X3 of SARS-CoV based on the analytical results above. X4 is a basic protein comprising 122 amino acids with molecular weight of 13.9 kD, and there is a strongly hydrophobic segment LFLIVAALVFLILCF located at residues 101—115 in X4. Both TMAP and TMHMM predict a transmembrane α-helical region at position 95—117, with N-terminal residues 1—94 on the outside of virus particle and C-terminal residues 118—122 on the inside of virus particle. We also note three uninterrupted basic amino acids KRK at C-terminal tail 118—120 that may interact with virus RNA. BLIMPS search reveals the segment 101—119 of X4 matches to Tetracycline resistance protein TetB signature PR01036 in BLOCKS database, while PSI-BLAST search fails to yield any homology sequence matching to X4. These analyses suggest that X4 may be a structural protein that is inserted in virus envelope and plays a function role in virus assembly. X5 is a basic protein comprising 84 amino acids with molecular weight 9.5 kD. PSIBLAST search fails to identify any corresponding homology sequence of other coronaviruses matching to X5. TMAP and TMHMM predict no transmembrane segment. BLIMPS search reveals that the segment 52—68 of X5 matches to conserved sequence family IPB000001 (Kringle) in BLOCKS database. These analyses suggest that X5 may be oriented inside virus particle, the strongly positive-charged regions located at N- and C-terminuses may interact with virus RNA and facilitate virus assembly.

HP5 (Hypothetical protein 5) is a strongly acidic protein comprising 98 amino acids with molecular weight of 10.8 kD. Both TMAP and TMHMM predict no transmembrane segment. HMMPFAM and BLIMPS searches yield no matching conserved protein motif. PSIBLAST search fails to identify any matching homology sequence. We cannot predict the function of HP5 from these analyses above.

ORF9 encodes a predicted small acidic protein comprising 44 amino acids with molecular weight of 5.3 kD, and the negative-charged amino acids are all located at N- and C-terminuses of this protein. PSIBLAST yields no matching homology protein. Both TMAP and TMHMM predict a transmembrane α-helical region at residues 9—31 with N-terminal residues 1—8 on the inside of virus and C- terminal residues 32—44 on the surface of virus particle. BLIMPS search reveals that the sequence 11—32 of ORF9 matches to conserved sequence family PR00697 (*C. elegans* Sra family integral membrane protein signature) in BLOCKS database, segment 8—32 matches to IPB001898 (Sodium:sulfate symporter family), and some other partial sequences match to IPB003804 (L-lactate permease), PR01434 (NADH-ubiquinone oxidoreductase chain 5 signature), PR01535 (Vomeronasal type 2 receptor family signature), IPB001421 (Mitochondrial ATPase subunit 8), IPB003362 (Bacterial sugar transferase), PR00169 (Potassium channel signature) and PR00701 (60 kD inner membrane protein signature), respectively. We speculate from these results that the protein encoded by ORF9 may be a structural protein oriented in virus envelope whose function role is involved in ion transferring. The protein encoded by ORF10 is a small basic protein comprising 39 amino acids with molecular weight of 4.3 kD. PSIBLAST search fails to get any homology sequence matching to this protein. Although TMHMM predicts no transmembrane helical region, TMAP predicts a transmembrane segment located at residues 3—31. BLIMPS search reveals that the segment 4—29 matches to IPB001010 (Thionin) in BLOCKS database, and segments 2—25 and 6—15 match to PR01520 (Zeta-tubulin signature) and IPB003606 (N-terminal to some SET domains), respectively. The protein encoded by ORF14 is an acidic protein comprising 70 amino acids with molecular weight of 7.8 kD. PSIBLAST search fails to identify any matching homology sequence. TMHMM predicts no transmembrane helical region, while TMAP predicts a transmembrane segment at position 34—61. Both HMMPFAM and BLIMPS searches yield no any matching conserved protein motif. We cannot predict the func-

tions of OFR10 and ORF14 from the results above.

(iv) Phylogenetic analysis of the proteins of SARS-CoV. We searched the non-redundant protein database (SDSC) for the homology sequences of each protein of SARS-CoV using PSIBLAST program, then the homology sequences of these proteins were subject to evolutionary distance matrix calculations (Table 3) and the unrooted phylogenetic trees were drawn (data not shown) to ascertain the classification status and evolutional relationship between SARS-CoV and other coronaviruses.

The topologies of the resulting phylograms are remarkably similar, the species formed monophyletic clusters consistent with the established three taxonomic groups. For each protein of SARS-CoV that possesses of the homology sequences, the SARS-CoV sequence segregated into a fourth, well-resolved branch, which indicates that SARS-CoV is not related to any of the three previously characterized groups of coronaviruses and forms a distinct group within the genus *Coronavirus*, this result is consistent with that of refs. [5,6]. The evolutionary distance matrix analyses reveal that most of SARS-CoV proteins (except PLP and N) are approximately equidistant from the corresponding proteins of previously characterized coronaviruses. For example, the evolutional distance of SARS-3CL^{pro} against the corresponding proteins of other coronaviruses ranges from 2.61 to 3.52, the average value is 3.09 and the standard deviation is 0.36; Rep ranges from 1.40 to 1.84, the average value is 1.66 and the standard deviation is 0.21; Hel ranges from 1.26 to 1.78, the average value is 1.52 and the standard deviation is 0.22, etc. (see Table 3). These results support the conclusion that SARS-CoV does not belong to any of the three previously characterized groups and should be within the fourth group reported by refs. [5, 6]. However, the evolutional distance analytical results indicate that the Hel and Rep related more closely to the corresponding proteins of other coronaviruses, the average distances between them are less than 1.7 (Table 3). The proteins Hel and Rep are responsible for virus RNA replication and transcription, and undergo a huge evolutional pressure because of their conserved functions, so the sequences and structures of them are more conservative than those of other SARS-CoV proteins. Furthermore, it is interesting that most of proteins of BoCoV and MHV are more close to the corresponding proteins of SARS-CoV in evolutional distance. For example, there are seven proteins PLP, NSP4, Rep, Hel, NSP11, NSP12 and S in BoCoV possessing of the closest evolutional distances against the corresponding proteins of SARS-CoV, and six proteins 3CL^{pro}, HD2, GFLP, Rep, M and N in MHV possessing of the closest evolutional distances against the corresponding proteins of SARS-CoV. However, there is no evidence to prove that the proteins of SARS-CoV come from the recombination of BoCoV and MHV or other coronaviruses.

3 Conclusions

The phylogenetic and evolutional distance analyses

Table 3 The evolutional distances between SARS-CoV associated proteins and the corresponding proteins of other coronaviruses

SARS-CoV	MHV [a)]	BoCoV	PEDV	HCoV-229E	TGEV	AIBV	FIPV	PRCoV	CCoV	RSCoV	HCoV-OC43	PHEV	TuCoV	ECoV	PuV	Average [b)]	StDev [c)]
PLP	7.22	7.09	9.26	9.49	9.86	9.84										8.79	1.29
3CL^{pro}	2.61	2.67	2.99	3.52	3.17	3.51	3.17									3.09	0.36
HD2	5.81	6.05	6.34	7.15	6.90	8.41										6.78	0.95
NSP4	2.73	2.54	3.42	3.53	2.94	3.05										3.04	0.38
NSP5	2.56	2.64	2.59	2.22	2.50	3.09										2.60	0.28
NSP6	3.44	2.96	2.85	3.09	3.43	3.41										3.20	0.26
GFLP	2.14	2.25	2.18	2.69	2.49	2.16										2.31	0.22
Rep	1.40	1.40	1.78	1.84	1.82	1.74										1.66	0.21
Hel	1.26	1.22	1.60	1.65	1.60	1.78										1.52	0.22
NSP11	1.71	1.70	2.24	2.20	2.21	2.24										2.05	0.27
NSP12	2.53	2.49	2.81	3.16	2.99	3.32		12.18								2.88*	0.34
S	5.63	5.33	8.06	6.62	7.53	7.19	7.11	7.56	7.31	5.65	5.4	5.60				6.58	0.99
M	3.35	3.37	4.59	5.13	4.90	5.85	4.93		4.93	3.54		3.36	16.43			4.40*	0.91
N	4.02	4.05	7.13	6.14	6.27	5.71	6.14		6.18	4.07	4.13	4.08	5.69	4.03	4.08	5.12	1.44

a) MHV, Murine hepatitis virus; BoCoV, bovine coronavirus; PEDV, porcine epidemic diarrhea virus; HCoV-229E, human coronavirus 229E; TGEV, transmissible gastroenteritis virus; AIBV, avian infectious bronchitis virus; FIPV, feline infectious peritonitis virus; PRCoV, porcine respiratory coronavirus; CCoV, canine coronavirus; RSCoV, rat sialodacryoadenitis coronavirus; HCoV-OC43, human coronavirus OC43; PHEV, porcine hemagglutinating encephalomyelitis virus; TuCoV, turkey coronavirus; EcoV, equine coronavirus; PuV, puffinosis virus; b) average value of evolutional distance; c) standard deviation.

of SARS-CoV associated proteins indicate that it is a previously unknown coronavirus and neither come from the recombination of the previously characterized coronaviruses nor arise as a mutant of known human or animal coronaviruses. We conjecture that there are two possible sources for SARS-CoV: One possibility is that it arises as a mutant of an unknown animal coronavirus or as a recombinant of two unknown coronaviruses, and acquired new virulence factors and the ability to infect humans; the other one is that it arises as a mutant of unknown human coronavirus that acquired new virulence factors. However, antibodies to the SARS-CoV werefound in serum samples obtained from patients with SARS during convalescence but not in human serum samples from people without SARS[3], which indicates no existence of such "mild" unknown human coronavirus. The large-scale serologic tests of wild and domestic animals and birds in the region where the SARS outbreak first appeared may identify the usual host or "sources" of SARS-CoV. When this manuscript will be completed, a research team from the University of Hong Kong stated that a coronavirus resembling the SARS virus had been isolated from six masked palm civets (*Paguma larvata*) in a wild-animal market in Shenzhen, China, and five out of the ten civet handlers at the market had antibodies against the SARS virus in their blood[27]. However, how does the coronavirus in palm civet acquire the ability to infect human (mutant or recombinant)? Does the SARS-CoV arisen from palm civets coronavirus still possess the ability to infect palm civets? Does it cause the lower respiratory tract infection? Is there any other animal hosts for SARS-CoV? The resolvents of these questions are essential for annihilation of SARS completely.

The predictions of the molecular weight, pI and extinction coefficient for all the proteins of SARS-CoV are helpful for their separation, extraction and purification. And the transmembrane region predictions, the conserved sequence family searches, and homology sequence alignments provide significant information for predicting protein function. The analytical results indicate that the protein products encoded by ORF1ab mainly play a function role in polyprotein cleavage, virus RNA replication and RNA transcription. X1 protein is possessed of transmembrane segments and an integral membrane protein Srb signature sequence, suggesting that X1 may be a structural protein that is oriented in the virus envelope. X2 protein is a strongly basic protein. It possesses no transmembrane segment and has a bipartite nuclear localization signal, suggesting that it may interact with and bind to virus RNA and facilitate virus assembly. The secondary structure features and hydrophobic and hydrophilic character of S protein support well the collapse model of coronavirus membrane fusion mechanism. The bioinformatical analysis of S protein provides useful information in search of antigenic determinant aimed at S protein. This study establishes the foundations for construction of the structures of SARS-CoV associated proteins with molecular modeling and for determination of SARS-CoV proteomics with experiments, and in the long run, it provides significant information for design of anti-SARS drugs and for development of antiviral vaccine.

References

1. Peiris, J. S. M., Lai, S. T., Poon, L. L. M., Coronavirus as a possible cause of severe acute respiratory syndrome, Lancet, 2003, 361: 1319—1325.
2. Drosten, C., Günther, S., Preiser, W. et al., Identification of a novel coronavirus in patients with severe acute respiratory syndrome, N. Engl. J. Med., 2003, 348: 1967—1976.
3. Ksiazek, T. G., Erdman, D., Goldsmith, C. S. et al., A novel coronavirus associated with severe acute respiratory syndrome, N. Engl. J. Med., 2003, 348: 1947—1958.
4. Qin, E. D., Zhu, Q. Y., Yu, M. et al., A complete sequence and comparative analysis of strain (BJ01) of SARS-associated virus, Chinese Science Bulletin, 2003, 48: 941—948.
5. Rota, P. A., Oberste, M. S., Monroe, S. S. et al., Characterization of a novel coronavirus associated with severe acute respiratory syndrome, Science, 2003, 300: 1394—1399.
6. Marra, M. A., Jones, S. J., Astell, C. R. et al., The genome sequence of the SARS-associated coronavirus, Science, 2003, 300: 1399—1404.
7. Jun, R. Y., Lin, W. C., Ling, A. E. et al., Comparative full-length genome sequence analysis of 14 SARS coronavirus isolateds and common mutations associated with putative origins of infection, Lancet, 2003, 361: 1779—1790.
8. Gallagher, T. M., Buchmeier, M. J., Coronavirus spike protein in viral entry and pathogenesis, Virology, 2001, 279: 371—374.
9. Bonavia, A., Zelus, B. D., Wentworth, D. E. et al., Identification of a receptor binding domain of the spike glycoprotein of human coronavirus HCoV-229E, J. Virol., 2003, 77: 2530—2538.
10. Garoff, H., Hewson, R., Opstelten, D. J., Virus maturation by budding, Microbiol. Mol. Biol. Rev., 1998, 62: 1171—1190.
11. Boeckmann, B., Bairoch, A., Apweiler, R. et al., The Swiss-Prot protein knowledgebase and its supplement TrEMBL in 2003, Nucleic. Acids. Res., 2003, 31: 365—370.
12. Herold, J., Raabe, T., Schelle-Prinz, B. et al., Nucleotide sequence of the human coronavirus 229E RNA polymerase locus, Virology, 1993, 195: 680—691.

13. Brendel, V., Bucher, P., Nourbakhsh, I. et al., Methods and algorithms for statistical analysis of protein sequences, Proc. Natl. Acad. Sci. USA., 1992, 89: 2002—2006.
14. Subramaniam, S., The biology workbench—A seamless database and analysis environment for the biologist (editorial), Proteins, 1998, 32: 1—2.
15. Persson, B., Argos, P., Prediction of transmembrane segments in proteins utilising multiple sequence alignments, J. Mol. Biol., 1994, 237: 182—192.
16. Sonnhammer, E. L., Heijne, G. Von., Krogh, A., A hidden Markov model for predicting transmembrane helices in protein sequences, Proc. Int. Conf. Intell. Syst. Mol. Biol., 1998, 6: 175—182.
17. Bateman, A., Birney, E., Cerruti, L. et al., The Pfam protein families database, Nucleic Acids Res., 2002, 30: 276—280.
18. Wallace, J. C., Henikoff, S., PATMAT: a searching and extraction program for sequence, pattern and block queries and databases, Comput. Appl. Biosci., 1992, 8: 249—254.
19. Altschul, S. F., Madden, T. L., Schaffer, A. A., Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, Nucleic. Acids. Res., 1997, 25: 3389—3402.
20. Thompson, J. D., Higgins, D. G., Gibson T. J., CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, Nucleic. Acids. Res., 1994, 22: 4673—4680.
21. Felsenstein, J., PHYLIP——Phylogeny Inference Package (Version 3.2), Cladistics, 1989, 5: 164—166.
22. Saraste, M., Sibbald, P. R., Wittinghofer, A., The P-loop——a common motif in ATP- and GTP-binding proteins, Trends. Biochem. Sci., 1990, 15: 430—434.
23. Kyte, J., Doolittle, R. F., A simple method for displaying the hydropathic character of a protein, J. Mol. Biol., 1982, 157: 105—132.
24. Bonavia, A., Zelus, B. D., Wentworth, D. E. et al., Identification of a receptor-binding domain of the spike glycoprotein of human coronavirus HCoV-229E, J. Virol., 2003, 77: 2530—2538.
25. Ortego, J., Escors, D., Laude, H. et al., Generation of a replication-competent, propagation-deficient virus vector based on the transmissible gastroenteritis coronavirus genome, J. Virol., 2002, 76: 11518—11529.
26. Kuo, L., Masters, P. S., The small envelope protein E is not essential for murine coronavirus replication, J Virol., 2003, 77: 4597—4608.
27. Cyranoskiand, D., Abbott, A., Virus detectives seek source of SARS in China's wild animals, Nature, 2003, 423: 467

(Received June 3, 2003; accepted June 26, 2003)