The R Protein of SARS-CoV: Analyses of Structure and Function Based on Four Complete Genome Sequences of Isolates BJ01-BJ04

Zuyuan Xu[1,2*], Haiqing Zhang[1*], Xiangjun Tian[2,1*], Jia Ji[1*], Wei Li[1], Yan Li[1], Wei Tian[1,2,3], Yujun Han[1], Lili Wang[1], Zizhang Zhang[2], Jing Xu[1], Wei Wei[2], Jingui Zhu[1], Haiyan Sun[1], Xiaowei Zhang[1], Jun Zhou[1], Songgang Li[1,4], Jun Wang[1], Jian Wang[1,2], Shengli Bi[5], and Huanming Yang[1,2#]

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

The R (replicase) protein is the uniquely defined non-structural protein (NSP) responsible for RNA replication, mutation rate or fidelity, regulation of transcription in coronaviruses and many other ssRNA viruses. Based on our complete genome sequences of four isolates (BJ01-BJ04) of SARS-CoV from Beijing, China, we analyzed the structure and predicted functions of the R protein in comparison with 13 other isolates of SARS-CoV and 6 other coronaviruses. The entire ORF (open-reading frame) encodes for two major enzyme activities, RNA-dependent RNA polymerase (RdRp) and proteinase activities. The R polyprotein undergoes a complex proteolytic process to produce 15 function-related peptides. A hydrophobic domain (HOD) and a hydrophilic domain (HID) are newly identified within NSP1. The substitution rate of the R protein is close to the average of the SARS-CoV genome. The functional domains in all NSPs of the R protein give different phylogenetic results that suggest their different mutation rate under selective pressure. Eleven highly conserved regions in RdRp and twelve cleavage sites by 3CLP (chymotrypsin-like protein) have been identified as potential drug targets. Findings suggest that it is possible to obtain information about the phylogeny of SARS-CoV, as well as potential tools for drug design, genotyping and diagnostics of SARS.

Key words: SARS, SARS-CoV, RNA-dependent RNA polymerase, RNA viruses, proteolysis

Introduction

In the life cycle of coronaviruses, the R (replicase) protein, the largest protein of the virus, is the first translated product following the infection of host cells by the virus. It is immediately translated by host ribosomes into a large polyprotein. This protein is then post-translationally modified to generate structurally independent or associated functioning components, thus initiating RNA replication of the viral genome.

The R protein mainly harbors the RdRp (RNA-dependent RNA polymerase) activity for replication of the genomic RNA by producing the (-) and (+) stranded RNA molecules, generating the subgenomic (+) transcripts required for the function of all the viral structural proteins and other uncharacterized proteins (*1*, *2*). In addition, it also supports the proteinase activities, namely, the main proteinase 3CLP that primarily mediates cleavage of RdRp and HEL (helicase), and accessory proteinases, such as PLP (papain-like protein), that are involved in the post-translationally proteolytic processes for other structural or non-structural proteins (*3*).

Analyses of the structure of the R protein are important because the information derived from the primary structure can be directly applied to drug design. All of the viral structural proteins are believed to be individually expressed from a nested set of co-

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

terminal subgenomic mRNAs. This expression arises through a unique discontinuous transcription mechanism, mainly involving RdRp, making it a potential target for drug inhibition. The endogenous proteolytic processing plays a prominent role in the production of the functional polypeptides. The predicted cleavage sites may be a potential peptido-mimetic substrate analogue for the protease inhibitors.

We report here an analysis of the R protein involving its structure, correlated enzymatic activities and other possible functions, as well as its evolution and potential medical implications, based on four complete genome sequences of the SARS-CoV isolates, BJ01-BJ04, and in comparison with 13 other published isolate genomes.

Results

Genomic structure of the R protein

The whole ORF (open-reading frame) for the R polyprotein accounts for approximately two thirds of the viral genome at the 5′ end, nucleotide (nt) position 246 to 21,466 (*4*). It theoretically encodes for a predicted protein of 7,073 residues with an estimated molecular weight of 790.28 KD. Two ORFs, ORF1ab (nt position 246-21,446) and ORF1a (nt position 246-13,394), are overlapped by a single nucleotide (cytosine) at nt position 13,379 that is the proposed site for (-1) ribosomal frameshift (*4*). A putative pseudoknot structure, the main signal for the (-1) frameshift, is located immediately downstream of the conserved slippery site (UUUAAAC at nt position 13,380 to 13,457).

The ORF1ab for the R protein has an average GC content of 40.8% (A: U: C: G = 28%: 31%: 19%: 21%). The distribution of GC appears relatively even except for the most 5′ end GC-rich region which corresponds to the putative leader protein sequence that locates at ~1-800 nt of the R protein (Figure 1A) (*5*).

A substantial fraction (41.95%, 2,967/7,073) of the ORF1ab for the R protein is composed of nonpolar hydrophobic residues, such as Leu (9.54%), Val (8.19%) and Ala (7.22%) (Table 1), and thus it is overall weakly acidic (pI 6.3) (*5*). In comparison with other viral proteins, the R protein has a relatively even distribution (1.09-9.54%) of the 20 natural amino acids (*5*). An obvious codon usage preference was identified, that is, codon CUU accounts for approximately 30% of Leu, GUU for 40.9% of Val, and GCU for 52.4% of Ala. However, the codon usage preference shows high similarity with the R protein of five other coronaviruses we have analyzed, representing the three groups in *Coronaviridae* (Table S1).

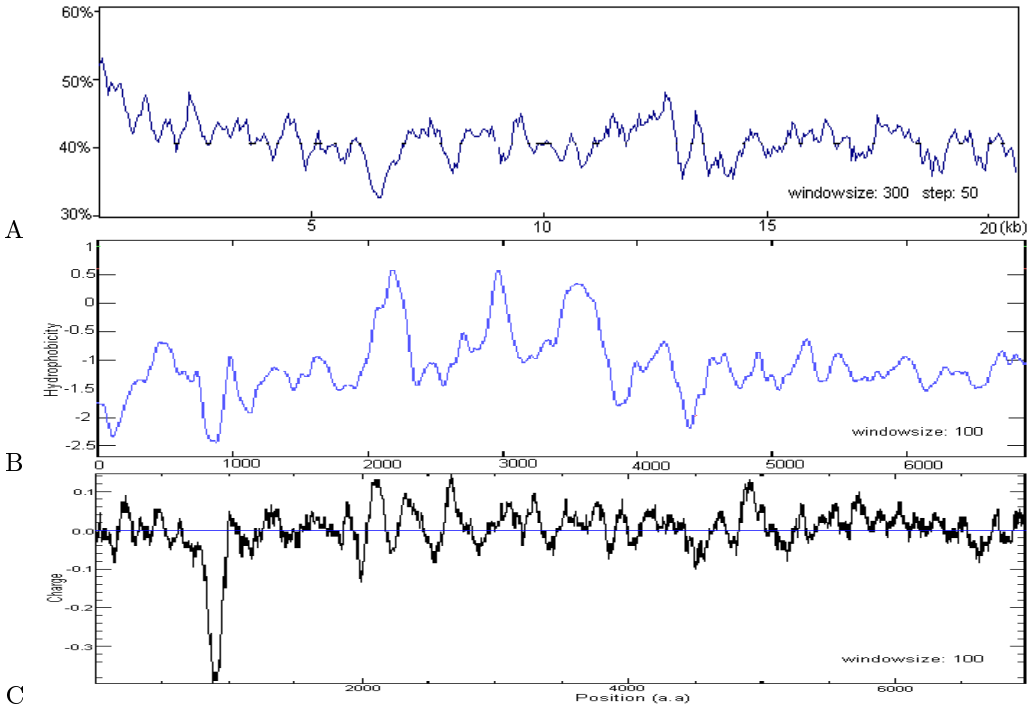

Fig. 1 Diagrams of the GC content (A), hydrophobicity (B) and charge distribution (C) of the R protein. The *X*-axes stand respectively for GC-content (A), hydrophobicity score (B) and charge score (C), generated by corresponding

algorithms (see materials and methods for details). The corresponding *Y*-axes stand for nt position (A) or amino acid (a.a.) position (B, C) of the R protein. The window sizes are 300 nt (A) and 100 a.a. (B, C).

Table 1 General Biochemical Features of the R Protein

	a.a.	No.*	F# (%)		a.a.	No.*	F#(%)
Non-polar,	Ala	511	7.22	Polar,	Ser	458	6.48
neutral	Val	579	8.19	neutral	Thr	495	7.00
	Leu	675	9.54		Cys	233	3.29
	Ile	343	4.85		Tyr	324	4.58
	Pro	274	3.87		Asn	366	5.17
	Phe	331	4.68		Gln	234	3.31
	Trp	77	1.09				
	Met	177	2.50				
	Gly	419	5.92				
	Total	2,967	41.95		Total	2,529	35.76
Charged,	Asp	395	5.58	Charged,	Lys	415	5.87
negative	Glu	348	4.92	positive	Arg	259	3.66
					His	160	2.26
	Total	743	10.50		Total	834	11.79

*No.: Number of the amino acid.

#F: Frequency in percentage of the amino acid in the R protein.

The distribution of GC content and hydrophobicity revealed three highly hydrophobic but AT-rich subregions close to the middle of the ORF. The 5′-end one (nt position ~6,500-7,100) was located immediately downstream of PLP (papain-like protein, nt position $4896^{\pm 15}$-$5535^{\pm 15}$), but the other two (nt position ~9,000-9,500 and ~10,700-11,600) corresponded to the known HODs (hydrophobic domains) (Figure 1A and 1B). An obvious negatively charged and highly hydrophilic region of Asp- and Glu-rich, named BGI Hydrophilic Domain (BGI-HID), was identified at nt position ~2,600-3,200 (Figure 1 B and 1C).

Localization of the function-related regions in the R protein

The whole ORF is composed of fifteen regions, conventionally named NSPs (non-structural proteins), which are defined by the putative cleavage sites by 3CLP and PLP, including the four known functional peptides (PLP, 3CLP, RdRp and HEL) (Figure 2). Three out of the eleven uncharacterized NSPs have predicted functions derived from their similarity to known or putative counterparts in other coronaviruses, while the function of the remaining eight has yet been totally unknown.

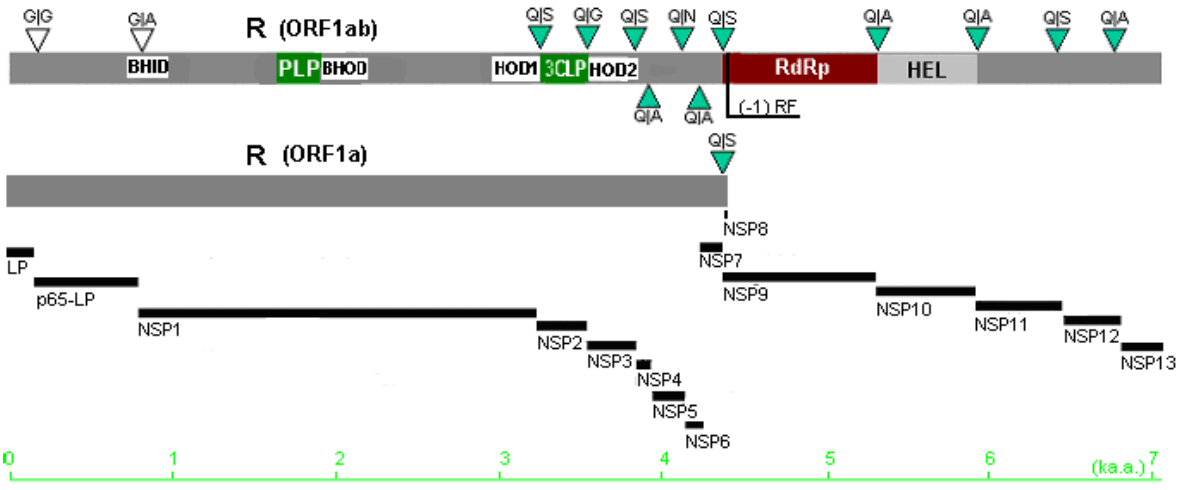

Fig. 2 Diagram of the putative function-related regions in the R protein (ORF1ab and ORF1a). Based on sequence analysis, we speculated and defined 15 regions that potentially function in SARS-CoV. 3CLP and PLP function as proteinase in the R protein. The blank triangles indicate the cleavage sites by PLP, and the solid triangles by 3CLP. The narrow black rectangles indicate the functional regions. The bottom ruler stand for the position of the amino acid of the R protein with a unit of kilo-amino acids (ka.a.). LP: leader protein. p65-LP: MHV p65 like protein. (-1) RF: (-1) ribosome frameshift. BHID: hydrophilic domain identified by BGI. BHOD: hydrophobic domain identified by BGI.

The RdRp activity

The region (NSP9) responsible for the RdRp activity is located between Codons 4,370 and 5,301 of the ORF1ab for the R protein (Figure 2; Table 2). At least 11 highly homologous subregions were identified in RdRp by similarity analysis (Figure 3). The DD (double Asp) domain (Subregion H in Figure S1), which consists of two conserved Asp residues flanked by at least five uncharged, mainly polar residues, is the most conserved region present in viral RdRp. It has been postulated to be involved in RNA binding (*6*).

Table 2 The Location and Size of the Putative Regions of the R Protein

Region	Location[§]	Size (a.a.)
LP[†]	246-782	179
p65-LP[‡]	783-2,669	639
PLP (NSP1)	2,670-9,965	2,422
3CLP (NSP2)	9,966-10,883	306
NSP3	10,884-11,753	290
NSP4	11,754-12,002	83
NSP5	12,003-12,596	198
NSP6	12,597-12,935	113
NSP7	12,936-13,352	139
NSP8	13,353-13,394	13
RdRp (NSP9)	13,353-13,379 13,379-16,147	932
HEL (NSP10)	16,148-17,950	601
NSP11	17,951-19,531	527
NSP12	19,532-20,569	346
NSP13	20,570-21,466	298

[§]nucleotide position of the ORF for the R protein.

[†]LP: leader protein.

[‡]p65-LP: MHV p65 like protein.

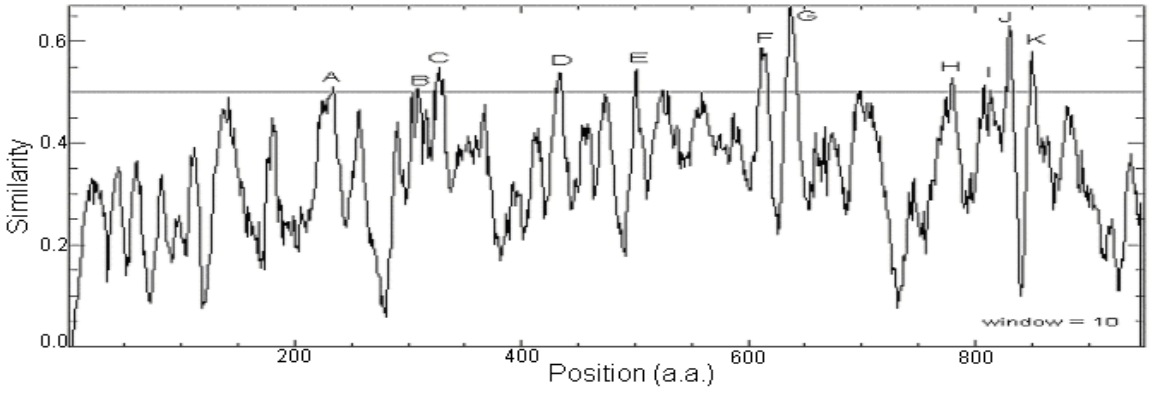

Fig. 3 Similarity analysis of the region for RdRp (NSP9) in the R protein. The X-axis stands for the similarity score of the multiple-alignment, and the Y-axis stands for the amino acid position of the consensus sequence of RdRps.

We used the sequences of RdRp from 7 coronaviruses, including SARS-CoV, to do multiple-alignment. The other 6 coronaviruses are avian infectious bronchitis virus (AIBV), bovine coronavirus (BCoV), human coronavirus 229E (HCoV-229E), murine hepatitis virus (MHV), porcine epidemic diarrhea virus (PEDV), and transmissible gastroenteritis virus (TGEV). Based on the graphic show (generated by EMBOSS-ploycon, window size = 10, see materials and methods for details) of the multiple-alignment, we highlighted 11 high-conserved subregions of the R protein, which might contribute to some important functions and can be potentially used as the target for anti-SARS drug design.

The Proteinase activity

The region for 3CLP is located between amino acid 3,241 and 3,546 of the ORF1a for the R protein, flanked by two known HODs, HD1 and HD2 (renamed as HOD1 and HOD2 to be different from the HID) (Figure 2; Table 2). EMBOSS polydot and polycon demonstrated high-conserved peaks at the N-terminus. Multiple-alignment located the peaks at two segments, "*LNGLWLDD*" (Codons 27-34) and "*CPRHVI*" (Codons 38-43), and defined the putative catalytic sites, His^{41} and Cys^{147} (Figures 4 and 5).

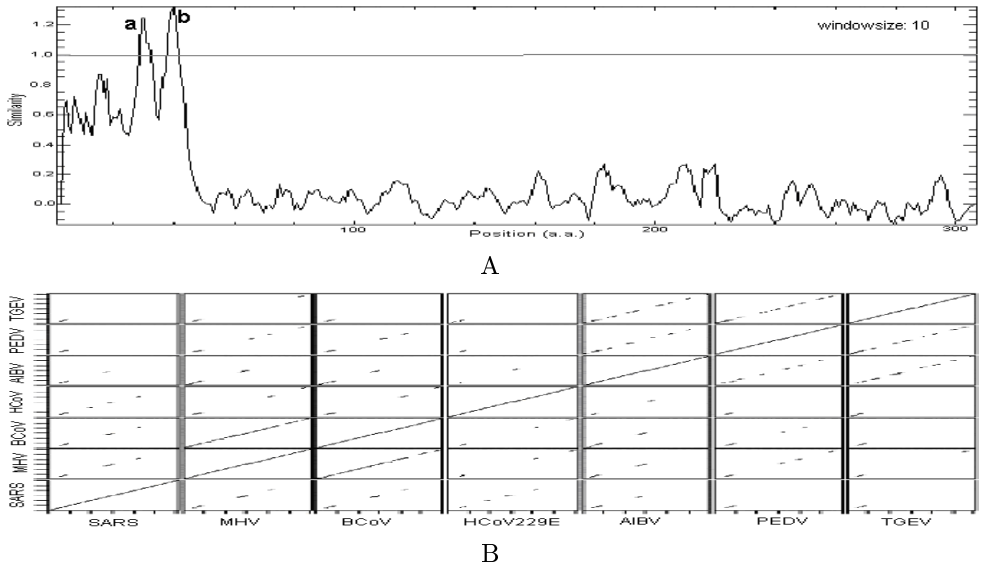

Fig. 4 Similarity analysis and conserved subregion in 3CLP (NSP2). Based on the multiple-alignment of 3CLP from seven coronaviruses that are similar to the samples used in Fig. 3, the diagram A (generated by EMBOSS-polycon, window size = 10, see materials and methods for details) shows the most similar subregions *a* and *b* of 3CLP. Based on the global pair-wise alignment of the seven 3CLPs, polydot diagram B (generated by EMBOSS-polydot, see materials and methods for details) shows the conserved regions of every pair.

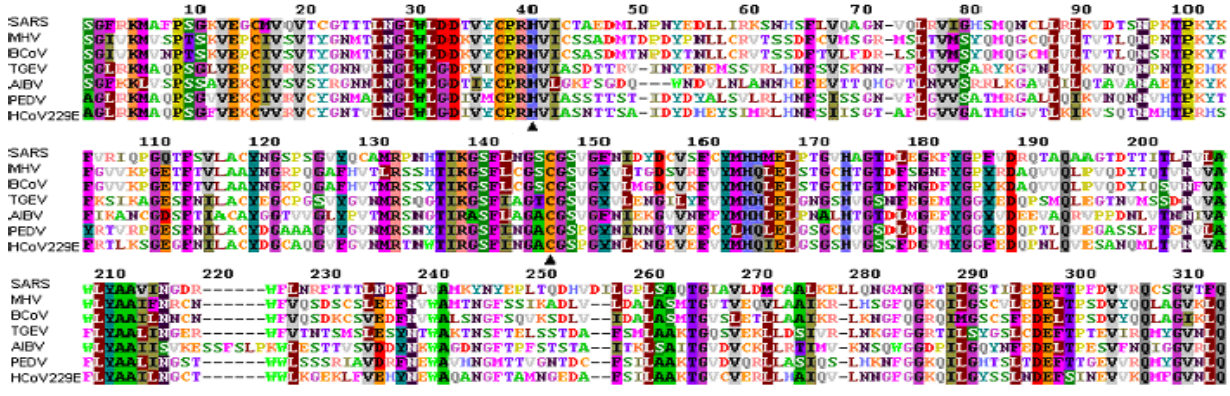

Fig. 5 Multiple alignment of the region for 3CLP (NSP2) among seven coronaviruses. 3CLP is the main proteinase of coronaviruses, with the catalytic sites His^{41} and Cys^{147}. The black triangles indicate the putative catalytic sites of

3CLP. The numbers above the sequences indicate the amino acid position of 3CLP. The amino acid was highlighted in different colors.

The region for PLP is located between Codons $1{,}632^{\pm 5}$ and $1{,}845^{\pm 5}$, and also between the newly identified BHID and BHOD in ORF1a for the R protein. It appears to be a single domain with moderate conservation by our analysis (Figure 2).

HEL and other NSPs

HEL is located on NSP10 (Codons 5,302-5,902), immediately downstream of RdRp that is postulated to be associated with HEL and ATPase activities of HEL both structurally and functionally. Besides, the N-terminal encoded coronavirus-specific LP (leader protein) region appears to be homologous to the tetrahydrofolate dehydrogenase/cyclohydrolase in Codons 76-160. NSP1, in addition to PLP, is similar to the appr-1′-p processing enzyme family and zinc carboxypeptidase A metalloprotease (M14). NSP2 is found to encompass a YejX-family in DUF463. NSP7 is similar to a common growth-factor-like protein (GFL). An FtsJ-like methyltransferase, which was thought to be involved in mRNA-Cap, was identified in NSP13 at the C-terminus with low identity (Figure 2; Table 2).

Sequence variations in coronaviruses

With the complete sequence of BJ01 as reference, BJ02, BJ03, and BJ04 have 8, 10, and 10 substitutions, respectively, in the R protein, and 5 (62.5%), 8 (80%), 9 (90%) of them are non-synonymous substitutions (Table 3).

Overall, we have detected 92 substitutions in the R protein in comparison with 17 published full-length genome sequences. Approximately three quarters (70.65%, 65/92) are non-synonymous. Using the complete sequence of BJ01 as the reference, Isolate GD01 has the biggest number of variations, 29, and TW1 the smallest, 5.

Table 3 Substitutions in Different Regions of the R Protein

Regions	Sn	F (%)*	Syn	nSyn	F (nSyn) (%)#
LP§	5	0.93	0	5	100
p65-LP†	4	0.21	2	2	50
NSP1 (PLP)	36	0.50	8	28	77.78
NSP2 (3CLP)	5	0.54	3	2	40
NSP3	2	0.23	0	2	100
NSP4	2	0.80	0	2	100
NSP5	3	0.51	1	2	66.67
NSP6	0	0.00	0	0	0
NSP7	4	0.96	0	4	100
NSP8	0	0.00	0	0	0
NSP9 (RdRp)	4	0.14	0	4	100
NSP10 (HEL)	7	0.39	4	3	42.86
NSP11	6	0.38	4	2	33.33
NSP12	4	0.39	2	2	50
NSP13	10	1.11	3	7	70
Total	92	0.43	27	65	70.65

Sn: Substitutions No.; Syn:Synonymous No.; nSyn: non-Synonymous No..

*F: Frequency in percentage of the number of substitutions in the corresponding region vs. its size in nucleotide.

#F (nSyn): Frequency in percentage of non-synonymous substitutions.

§LP: leader protein. An L→STOP substitution was found in the leader protein.

†p65-LP: MHV p65 like protein

Comparative genomics of the R protein

Besides the 14 coronaviruses, we have only found 4 matches in GenBank (Venezuelan equine encephalitis virus, Gill-associated virus, Yellow head virus and Simian hemorrhagic fever virus), either complete or partial, with the overall ORF of the R protein. It should be noticed that the identified conserved regions of the R protein mainly contributed to the matches.

The essentially complete sequence of the R protein has been identified in another 6 species of *Coronaviridae*. Our comparative analysis demonstrated that the one-third N-terminal region is highly variable, in a sharp contrast with the other two-thirds, constituting a relatively stable region (Figure 6). Pair-wise global alignment was also made among 7 coronaviruses, which indicated that BCoV (Bovine Coronavirus) and MHV (Murine Hepatitis Virus) would have a relatively higher similarity index (Figure 7).

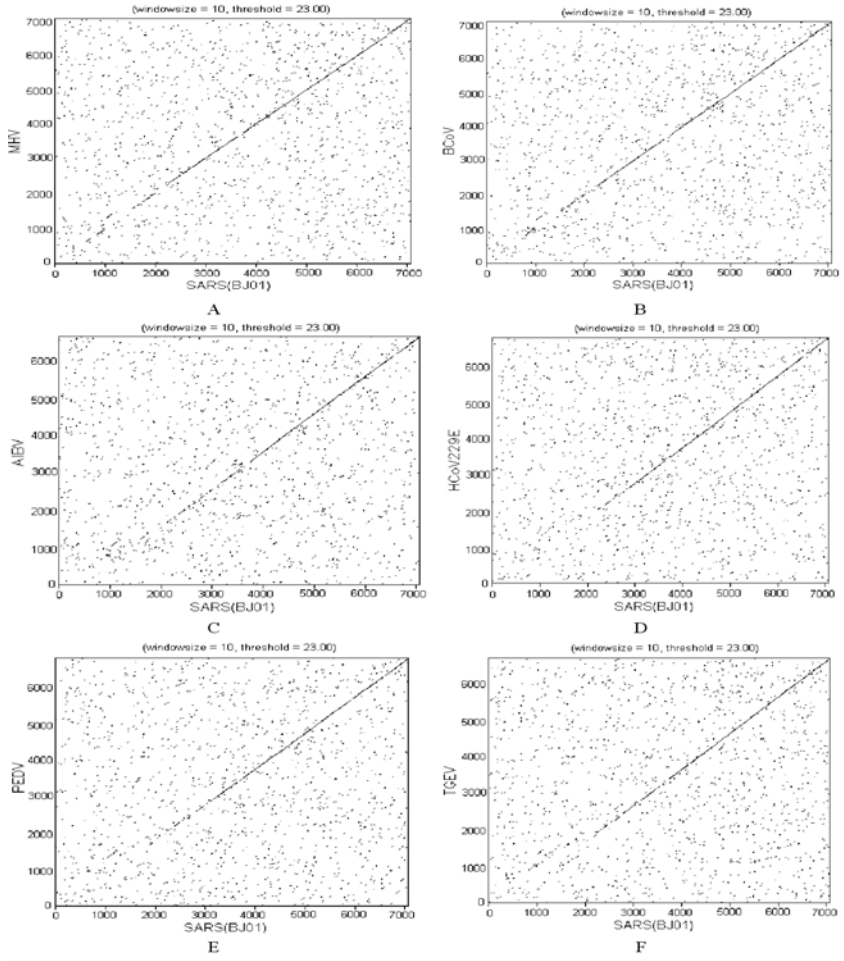

Fig. 6 Dotplot diagram (generated by EMBOSS-doplot, window size=10, threshold=23, see materials and methods for details) of the similarity in the R protein between SARS-CoV and other 6 coronaviruses. The X- and Y-axes stand for the amino acid position of corresponding R protein. (A) SARS vs MHV; (B) SARS vs BcoV; (C) SARS vs AIBV; (D) SARS vs HCoV229E; (E) SARS vs PEDV; (F) SARS vs TGEV. It is suggested that MHV and BCoV are more homologous to the SARS-CoV.

	SARS	BCoV	MHV	PEDV	HCoV	TGEV	AIBV
SARS	**100**/100						
BCoV	**42.3**/58.7	**100**/100					
MHV	**42.2**/57.9	**74.1**/84.4	**100**/100				
PEDV	**36.9**/52.8	**36.7**/52.5	**36.4**/52.3	**100**/100			
HCoV	**36.5**/53.2	**35.8**/52.3	**35.6**/51.7	**59.6**/74.1	**100**/100		
TGEV	**36.5**/52.5	**36.1**/52.3	**36.0**/51.8	**51.9**/66.3	**52.0**/67.4	**100**/100	
AIBV	**36.4**/52.7	**36.5**/51.8	**36.0**/51.1	**36.4**/52.4	**35.7**/51.7	**35.6**/52.5	**100**/100

Fig. 7 Pair-wise alignment based on amino acid sequences of the R protein among SARS-CoV and the other 6 coronaviruses. The alignment was performed by EMBOSS-stretcher (see materials and methods for details), in which

Myers and Miller algorithm (*7*) was used instead of the standard sequence global alignment, Needleman and Wunsch algorithm, only to save time and disk memory. The bold number and the normal number indicate the identity and the similarity score, respectively.

An unrooted phylogenetic tree based on multiple-alignment of amino acid sequences is proposed (Figure 8A). This proposed tree places SARS-CoV outside the three known groups, between Group 1 and Group 3, with genetic distance almost similar to Group 2.

The unrooted phylogenetic trees were also constructed with amino acid sequences of NSP1, PLP, 3CLP, RdRp, and HEL. They demonstrated that the evolution of different regions is non-synchronous (Figure 8B, C, D, E, F).

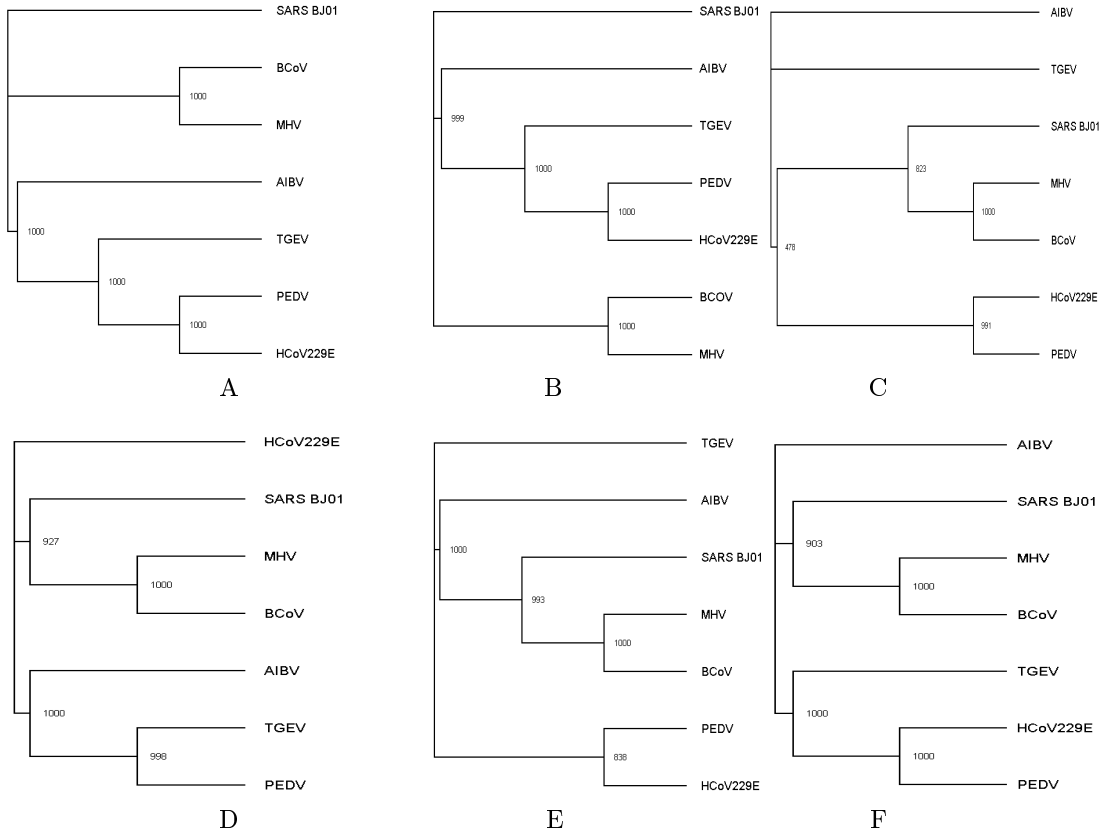

Fig. 8 Proposed phylogenetic trees based on amino acid sequences of the R protein (A), and that of NSP1 (B), PLP (C), 3CLP (D), RdRp (E), and NSP10 (HEL) (F). All the bootstrap trees are generated by ClustlW (see materials and methods for details). The numerical value near the node of branches is the trial for bootstrap.

Discussion

Proteinase activities and drug design

Two regions have been identified to be responsible for the proteinase activities of the R protein (Figure 2) based on our comparative analyses of the four complete genome sequences of the BJ group and other previously published experimental data (*8-11*). The PLP domain, which was named after the similar catalytic dyad arrangement to cellular proteinases related to papain (*3*), appears to be a characteristic functional domain within the putative NSP1. The possibility that it was composed of two interactive peptides as the putative PLP1 and PLP2 in MHV was excluded, based on homologous comparison with PLPs in other coronaviruses. Substrate specificities of the PLP show that it has several preferred cleavage sites in different coronaviruses (*3*). We have found that the structural S (spike), E (envelope), M (membrane), and N (nuleocapsid) proteins, have 18, 1, 4 and 10 possible cleavage sites, respectively. However, amino acids around those cleavage sites might also be the major deter-

minants for recognition (*10*). Therefore, among the above 33 sites, we could only identify, with certainty, one cleavage site (a.a. position 39, RG|V) in the S protein and two (a.a. position 198, RG|N and a.a. position 205, RG|N) in the N protein by the PLP.

The 3CLP domain, which was first reported from picornavirus 3C proteinases ($3C^{pro}$) and thus named (*12*), is located on NSP2. Experimental data from AIBV, MHV and HCoV-229E also suggested that the 3CLP is responsible for proteinase activities demonstrated by the R protein (*9, 13, 14*). The recombinant main proteinase of SARS-CoV can mediate cleavage of a TGEV M^{pro} (main proteinase) substrate (*15*). The putative catalytic sites, His^{41} and Cys^{147} (a.a. position of NSP2) may be associated with the functional performance of 3CLP in the replication complex (*16*, *17*). By searching for the 3CLP conserved cleavage sites, we have found a single site each in the S and N proteins, but none in the E and M proteins. This is consistent with a previous study reporting that M proteins become integrated without involvement of a cleaved signal peptide (*18*).

The three HODs we have identified by overall analysis of the entire ORF are significantly hydrophobic (Figure 1B). The downstream two, HOD1 and HOD2, have been postulated to mediate the microsomal membrane association of the replication complex and to alter dramatically the architecture of host cell membranes, resulting in the optimal construction of the reaction complex for replication and translation (*9*, *16*). The function of BGI-HOD, as well as the BGI-HID, requires further experimental data.

It has been established that HIV protease inhibitors have been designed to mimic the peptidic linkages that are cleaved by the protease (*19*). More detailed understanding of the proteinase activities and their localization in the R protein would provide essential information for drug screening and design.

RdRp as a potential target for drug design

Our effort has been devoted to searching for homologous domains in the RdRp that might be candidate targets for future inhibitors for RdRp and potentially for drugs designed against SARS. RdRp inhibitors have been previously reported for HCV (hepatitis C virus) (*20*). It is noteworthy that the R protein does not have any homology in human. Therefore, to minimize the possible toxicity, the R protein may be a candidate target for a drug development.

We have defined 11 domains in the RdRp subregion (Figure S1). Among them, the F, G, J and K domains are the most conservative. Analyses of hydrophobicity and electric charge show that F and G are hydrophilic with positive charge, while J and K are neutral. The conserved amino acid in these four domains and that of the other six could all contribute to drug design, based on the assumption that the conserved region is the most likely to be essential to the function of protein.

We also note that Ribavirin, which has recently been found to inhibit RdRp in Influenza virus, is promising for clinical treatment of SARS (*21*). A comparison of RdRp between Influenza virus and SARS-CoV has been performed, but no obvious sequence similarity has been identified, even three-dimensionally or functionally similar motifs or domains cannot be excluded. However, it should be noted that homologous primary structures might not share the same tertiary structure.

Evolution of the R protein

The RdRp probably evolved very early because it is one of the essential proteins in all RNA viruses (*22*). The R protein may be the only protein that is rooted in a common ancestry of many ssRNA viruses, and may have its orthologues in virus genomes outside family *Coronaviridae* on comparative analysis. The phylogenetic relationship with the R protein has been established mainly on the basis of the conservation of the homologous RdRp domains, together with the similar polycistronic genome organization, and the use of common transcriptional and post-translational strategies of the viruses (*23*).

However, the homology analysis yields rather disappointing results. No significant homology has been revealed thus far regarding the R protein of coronaviruses. The hypothesis of the divergent evolution from a common *Nidovirales* ancestor, containing a replicase gene with an organization resembling that of the contemporary subsets of *Nidovirales*, provides the most parsimonious explanation for the observed diversity of the proteolytic enzymes (*3*).

The results we have achieved for PLP, RdRp, and HEL show they have different mutation rates, indicating that the R protein might not be an intact element in evolution. An alternative interpretation would be the non-even distribution of mutations in different regions, without the involvement of selective pressure.

The R protein itself does not have a high rate of mutation

It is well known that RNA viruses have relatively higher detectable rate of mutation than DNA viruses (*5*). It is postulated that a dysfunction in proof reading of the RNA polymerase is responsible for the higher mution rate. The infidelity of transcription would affect both the R protein per se and other protein or functional elements.

It would be misleading to suggest that the substitutions of the R protein represent a large proportion of the variation detected in sequences of various isolates of SARS-CoV. As seen in our comparative analysis, it appears that the R protein accounts for 64.8% of the total number of substitutions. However, if the large size of the protein, accounting for a substantial proportion, is taken into consideration, the substitution rate of the R protein is only 0.43%, which is actually lower than the overall substitution rate of the entire SARS-CoV genome.

It can be preliminarily concluded that selective pressure might play a more significant role than the high rate of replication error of the R protein in maintaining the mutations that would be beneficial to the growth rate, host range, and so on. The relatively low substitution rate of the R protein itself may be a reflection of its stability in evolution.

However, it has not escaped from our attention that, in spite of the high rate of substitutions, none or rare indel (insertion or deletion) has been found in various isolates of SARS-CoV so far. This observation would suggest that the RdRp has a high fidelity of frame moving, even though its related region or structure has not been defined yet.

Materials and Methods

Samples and sequences: SARS-CoV samples, Isolates BJ01-BJ04, were taken from the SARS patients diagnosed in February and March 2003 in Beijing, China, according to WHO guidelines (http://www.who.int/csr/sars/guidelines/en/). The processing of tissue samples, inoculation into Vero-6 cell culture, virion preparation and viral RNA extraction, and RT-PCR amplification and cloning into sequencing vectors, was performed according to standard protocols at BGI and the Center of Disease Control and Prevention of China (CDC) (*24*). Sequencing was performed on MegaBACE 1000 (Amersham, New Jersey, USA) and ABI 377 (Applied Biosystems, California, USA).

The updated complete genome sequences of the BJ Group (BJ01-BJ04) have been deposited by BGI in GenBank (accession numbers: AY278488, AY278487, AY278490, AY279354) and are freely available (http://www.genomics.org.cn/bgi/news/zhongxin/news030416-2_fasta.htm). All the experimental materials, including all the cDNA clones representing various segments of the viral genome with known sequences, are available for collaborators.

Thirteen other full-length sequences of SARS-CoV genomes published from May 2003 by BGI or others were used in this study (accession numbers: AY278554, AY297028, AY274119, AY291451, AY283798, AY283797, AY283796, AY283795, AY283794, AY282752, AY278741, AY278491, AY278489). Ten coronavirus genome sequences containing the complete or partial ORF for the R protein were downloaded from GenBank and used for comparative analysis (accession numbers: NC_004718, AY274119, NC_003436, NC_002306, NC_003045, NC_002645, AF220295, NC_001846, NC_001451, M23694). The nucleotide positions of all SARS-CoV referred to the complete genome sequence of Isolate BJ01 (*5*).

We used six presently available software packages for structure and function analysis. ORF Finder (http://ww.ncbi.nlm.nih.gov/gorf/gorf.html) was used to determine ORFs and DNA_GC_Calculator to calculate GC content (http://www.genome.iastate.edu/ftp/share/DNAgcCal/). TopPred2 (http://www.sbc.su.se/~erikw/toppred2/) was selected to predict the hydrophobic region, AnTheProt (http://www.bimcore.emory.edu/home/Software/NPSA/Npsa.html) and the EMBOSS package (http://www.hgmp.mrc.ac.uk/Software/EMBOSS/) to characterize the proteins, and ClustalW to perform multiple-alignment and phylogenetic analysis. All analyses mentioned above were accomplished on supercomputers DOWNING 2000/3000 (DOWNING Computers Inc., Beijing, China), SUN E10K (SUN Microsystems Inc., California, USA), SGI Origin 3800 (Silicon Graphics, Inc., California, USA), and IBM P690 (IBM Corp., New York, USA).

Acknowledgements

The authors thank the 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, National Center of Disease Control of China, the Provincial Government of Zhejiang, the Municipal Governments of Beijing and Hangzhou, and the Library of Chinese Academy of Sciences. Special gratitude is expressed here to the patients and their families for their devotion and cooperation. We appreciate the comments of Dr. Gwendolyn E. P. Zahner, visiting professor at BGI, Dr. Qimin You, Dr. Lin Hu, and other colleagues on drafts of this manuscript.

References

1. Cavanngh, D. and Brown, T.D.K. (ed.) 1997. *Coronaviruses and their diseases.* 327-356. Plenum Press, New York, USA.
2. De Vries, A.F., *et al.* The genome organization of the Nidovirales: similarities and differences between Arteri-, Toro-, and Coronaviurses. *Semin. Virol.* 8: 33-47.
3. Ziebuhr, J., *et al.* 2000. Virus-encoded proteinases and proteolytic processing in the Nidovirales. *J. Gen. Virol.* 81: 853-879.
4. Qin, E.D., *et al.* 2003. A complete sequence and comparative analysis of a SARS-associated virus (Isolate BJ01). *Chin. Sci. Bull.* 48: 941-948.
5. Brierley, I. 1995. Ribosomal frameshifting on viral RNAs. *J. Gen. Virol.* 76: 1885-1892.
6. Norman, M. (ed.) 1988. *Oxford Surveys on Eukaryotic Genes (Vol. 5).* 91-131. Oxford University Press, Oxford, United Kingdom.
7. Myers, E. and Miller, W. 1988. Optimal alignments in linear space. *CABIOS* 4: 11-17.
8. Lim, K.P., *et al.* 2000. Identification of a novel cleavage of the first papain-like proteinase domain encoded by open reading frame 1a of the coronavirus avian infectious bronchitis virus and characterization of the cleavage products. *J. Virol.* 74: 1674-1685.
9. Ziebuhr, J., *et al.* 1995. Characterization of human coronavirus (strain 229E) 3C-Like proteinase activity. *J. Virol.* 69: 4331-4338.
10. Jens, H., *et al.* 1998. Proteolytic processing at the amino terminus of human coronavius 229E gene 1-encoded polyproteins: identification of a papain-like proteinase and its substrate. *J. Virol.* 72: 910-918.
11. Kanjanahaluethai, A. and Baker, S.C. 2000. Identification of Mouse Hepatitis Virus papain-like proteinase 2 activity. *J. Virol.* 74: 7911-7921.
12. Rueckert, R.R. and Wimmer, E. 1984. Systematic nomenclature of picornavirus proteins. *J. Virol.* 50: 957-959.
13. Liu, D.X., *et al.* 1994. A 100-kilodalton polypeptide encoded by open reading frame (ORF) 1b of the coronavirus infectious bronchitis virus is processed by ORF1a products. *J. Virol.* 68: 5772-5780.
14. Lu, Y., *et al.* 1995. Identification and characterization of a serine-like perteinase of the murine coronavirus MHV-A59. *J. Virol.* 69: 3554-3559.
15. Anand, K., *et al.* 2003. Coronavirus main proteinase ($3CL^{pro}$) structure: Basis for design of anti-SARS Drugs. *Science* 300: 1763-1767.
16. van der Meer, *et al.* 1998. ORF1a-encoded replicase subunits are involved in the membrane association of the arterivirus replication complex. *J. Virol.* 72: 6689-6698.
17. Pedersen, K.W., *et al.* 1999. Open reading frame 1a-encoded subunits of the arterivirus replicase induce endoplasmic reticulum-derived double-membrane vesicles which carry the viral replication complex. *J. Virol.* 73: 2016-2026.
18. Raamsman, M.J., *et al.* 2000. Characterization of the coronavirus mouse hepatitis virus strain A59 small membrane protein E. *J. Virol.* 74: 2333-2342.
19. De Clercq, E. 2002. Strategies in the design of antivral drugs. *Nature Rev.* 1: 13-24.
20. Walker, M. P. and Hong, Z. 2002. HCV RNA-dependent RNA polymerase as a target for antiviral development. *Curr. Opin. Pharmacol.* 2:1-7.
21. De Clercq, E. 2001. Antiviral drugs: current state of the art. *J. Clin. Virol.* 22: 7-10.
22. Bruenn, J.A. 2003. A structural and primary sequence comparison of the viral RNA-dependent RNA polymerases. *Nucleic Acids Res.* 31: 1821-1829.
23. Den Boon, J.A., *et al.* 1991. Processing and evolution of the N-terminal region of the arterivirus replicase ORF1a protein: identification of two papain-like cysteine proteases. *J. Virol.* 69: 4500-4505.
24. Zhu, Q. Y., *et al.* 2003. Isolation and identification of a novel coronavirus form patients with SARS. *J. Chin. Biotech.* 23: 106-112.

Supporting Online Material
http://www.gpbjournal.org/journal/pdf/GPB1(2)-08.htm
Table S1
Figure S1