Identification of probable genomic packaging signal sequence from SARS-CoV genome by bioinformatics analysis[1]

QIN Lei[2], XIONG Bin[3], LUO Cheng[3], GUO Zong-Ming[2], HAO Pei[4], SU Jiong[2], NAN Peng[2], FENG Ying[2], SHI Yi-Xiang[2], YU Xiao-Jing[4], LUO Xiao-Min[3], CHEN Kai-Xian[3], SHEN Xu[3*], SHEN Jian-Hua[3*], ZOU Jian-Ping[4*], ZHAO Guo-Ping[4*], SHI Tie-Liu[4*], HE Wei-Zhong[2*], ZHONG Yang[2,5*], JIAGN Hua-Liang[3*], LI Yi-Xue[2,4*]

[2]Shanghai Center for Bioinformation Technology, Shanghai 201203; [3]Drug Discovery and Design Center, State Key Laboratory of Drug Research, Shanghai Institute of Materia Medica, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 201203; [4]Shanghai Institutes for Biological Sciences, CAS, Shanghai 200031; [5]School of Life Sciences, Fudan University, Shanghai 200433, China.

KEY WORDS severe acute respiratory syndrome (SARS); packaging signal; stem-loop structure; N protein; M protein; antisense RNA; interfere RNA

ABSTRACT

AIM: To predict the probable genomic packaging signal of SARS-CoV by bioinformatics analysis. The derived packaging signal may be used to design antisense RNA and RNA interfere (RNAi) drugs treating SARS. **METHODS:** Based on the studies about the genomic packaging signals of MHV and BCoV, especially the information about primary and secondary structures, the putative genomic packaging signal of SARS_CoV were analyzed by using bioinformatic tools. Multi-alignment for the genomic sequences was performed among SARS-CoV, MHV, BCoV, PEDV and HCoV 229E. Secondary structures of RNA sequences were also predicted for the identification of the possible genomic packaging signals. Meanwhile, the N and M proteins of all five viruses were analyzed to study the evolutionary relationship with genomic packaging signals. **RESULTS:** The putative genomic packaging signal of SARS-CoV locates at the 3' end of ORF1b near that of MHV and BCoV, where is the most variable region of this gene. The RNA secondary structure of SARS-CoV genomic packaging signal is very similar to that of MHV and BCoV. The same result was also obtained in studying the genomic packaging signals of PEDV and HCoV 229E. Further more, the genomic sequence multi-alignment indicated that the locations of packaging signals of SARS-CoV, PEDV, and HCoV overlaped each other. It seems that the mutation rate of packaging signal sequences is much higher than the N protein, while only subtle variations for the M protein. **CONCLUSIONS:** The probable genomic packaging signal of SARS-CoV is analogous to that of MHV and BCoV, with the corresponding secondary RNA structure locating at the similar region of ORF1b. The positions where genomic packaging signals exist have suffered rounds of mutations, which may influence the primary structures of the N and M proteins consequently.

[1] Project supported by the 863 Hi-Tech Program, (No 2001AA235051, 2001AA235071, 2001AA233031), the National Natural Science Foundation of China, (No 29725203, 20072042), the State Key Program of Basic Research of China, (No 2002CB512801, 2002CB512802, 2002CB512801, 110CB510209), the special programs of oppugning SARS from the Ministry of Science and Technology, Chinese Academy of Sciences and the Shanghai Science and Technology Commission, Key Program of Basic Research of Science & Technology Commission of Shanghai Municipality, (No 02DJ14013, 02DJ14014), Shanghai Basic Research Project from the Shanghai Science and Technology Commission, (No 03DZ14102, 02DJ14070), and the One Hundred People Program of Chinese Academy of Sciences (1999), and the Knowledge Innovation Program of Chinese Academy of Sciences, No KSCX2-SW-202.

[*] Correspondence to Prof SHEN Xu, SHEN Jian-Hua, ZOU Jian-Ping, ZHAO Guo-Ping, SHI Tie-Liu, HE Wei-Zhong, ZHONG Yang, JIAGN Hua-Liang, LI Yi-Xue

Received 2003-05-16 Accepted 2003-05-18

INTRODUCTION

The severe acute respiratory syndrome (SARS) associated coronavirus (SARS-CoV) has been recognized as the causative agent for SARS. Since it broke out early this year, SARS has infected thousands of people and caused hundreds of death in many countries and places of the world[1,2]. It is urgent to develop efficient therapeutic means for the treatment of SARS patients. Antisense RNA and RNA interference (RNAi) technologies have shown potential prospect in treating some severe diseases[3,4] and should be considered as important candidate medicines in oppugning SARS. Many potential targets for designing antisense RNA or RNAi have been considered. Much more attention should be paid to the genomic packaging signal sequences of SARS-CoV in designing anti-SARS antisense RNA and RNAi segments.

Within the "soup" of a virally infected cell, viral genome and viral proteins specifically and selectively coalesce into progeny viruses. This process called packaging is an essential step for a productive viral replication cycle, involving recognition and interaction of various molecules[5,6]. RNA packaging signals necessary for virus RNA packaging have been described for several RNA viruses[7-9]. Among them, there are some examples in which the identified packaging signal is necessary and sufficient for viral RNA packaging[10].

It has been reported that coronaviruses selectively package their genomic RNAs exclusively or much more efficiently than those subgenomic RNAs[11,12]. Only few have been understood about how coronaviruses recognize the viral genomic RNA to ensure specific encapsidation and packaging. Thus far, packaging signals were only well studied in the murine hepatitis virus (MHV) and bovine coronavirus (BCoV). By using MHV defective interfering (DI) RNAs, a 69-nucleotide packaging signal that maps approximately 20 kb from the 5' end of the genome within the gene ORF1b was identified[11,13-16]. Inclusion of the packaging signal in a none-MHV RNA is sufficient to allow the RNA to be packaged into MHV virons[11]. Mutagenesis of the predicted bulged stem-loop structure of the MHV packaging signal disrupts the ability of the sequences to function as a packaging signal[11]. A similar packaging signal has also been identified in BCoV, which could package the noncoronavirus RNA into BCoV virions when the packaging signal was introduced. Comparative research of the genomic packaging signals of MHV and BCoV revealed that the conservation of RNA secondary structure was essential for their interaction with the N proteins[17,18]. All these support the idea that an essential packaging signal exists in SARS-CoV packaging process. In this study, bioinformatics analysis tools such as multi-alignment, RNA secondary structure prediction, and mutation analysis methods were employed to predict the putative genomic packaging signal of SARS-CoV.

MATERIALS AND METHODS

Materials The genomic sequences, protein sequences of N proteins and M proteins of MHV, BCoV, porcine epidemic diarrhea (PEDV), human coronavirus (HCoV) 229E and SARS-CoV were retrieved from the GenBank in NCBI. Genomic sequences and protein sequences listed in Tab 1 were used in this study.

Bioinformatics analysis Multi-alignment of genomic sequences and protein sequences in this study were carried out with vector NTI software from the Informax Inc[19]. The secondary structures of genomic RNA sequences were predicted by using the RNA Structure 3.71 program[20]. Transmembrane analysis of the M protein was carried out with TMAP software online[21].

Tab.1 Genome and protein sequences used in this study.

Virus strain	Group	Genome ID	Genome size	N protein ID	M protein ID
MHV-1	2	AF029248	31.2 kb	gi_2641131	gi_2641132
MHV-2	2	NC_001846	31.2 kb		
MHV-3	2	AF208066	31.1 kb		
BCoV	2	NC_003045	31 kb	gi_15081553	gi_15081552
PEDV	1	NC_003436	28 kb	gi_19387581	gi_19387580
HCoV 229E	1	NC_002745	27.4 kb	gi_12175753	gi_12175752
SARS-CoV	1or 2	NC_004718	29.7 kb	gi_29836503	gi_29836504

Mapping the derived genomic packaging signals from secondary structures to primary sequences in the result of multi-alignment was performed manually.

RESULTS AND DISCUSSION

The genomic packaging signals of MHV and BCoV The genomic packaging signal of MHV has been accurately positioned in a 69 bp region, about 1400 bp preceding the 3' end of ORF1b[11,16,22]. This RNA sequence shapes as a typical stem-loop secondary structure in the genome. Recombinant vectors solely containing this region was transfected into DBT cells and was packaged into MHV virions, indicating that this region contains all the signals for packaging. The mutant at the 5' end of this region blocks the packaging signal function completely.

It also has been reported that the 190 bp sequences comprising the 69 bp core sequences are packaged more effectively. This indicates that the sequences around the 69 bp core region may also be recognized by the M protein[5,7]. The sequence around the 69 bp core sequences is variable without affecting the packaging of genomic RNA. Sequences of the BCoV RNA share high homology to that of the MHV RNA. Recent study revealed that a similar packaging signal was addressed near the 3' end of ORF1b. RNA secondary structure analysis showed that the packaging signals of both MHV and BCoV form a long stem-loop structure, sharing an identical shape at the top two loops (Fig 2 and 3). Experiments also showed that RNA transcripts were packaged into both BCoV and MHV virions when the cloned region was appended to a noncoronavirus RNAs[17]. This emphasized that the top two loops in the RNA secondary structure were the most important factor for the N protein-packaging signal interaction. In this research, we mapped this high conserved region to the multi-alignment result of BCoV and MHV genomic sequences. The primary sequences corresponding to these secondary structures were addressed at the same position (Fig 1). The primary sequence of the top two loops of MHV is located at the 5' end of the 69 bp region, less than 30 bp in length. This implies that the key packaging signal might be even shorter than 69 bp (Fig 1). The astonishing similarity of primary sequences and secondary structures between MHV and BCoV packaging signals shows that the secondary structure responsible for genome packaging might constantly exist in a similar region among coronaviruses.

Multi-alignment of SARS-CoV genomic sequence with MHV, BCoV, PEDV, and HCoV 229E Genomic RNA sequences of SARS-CoV, MHV, BCoV, PEDV, and HCoV 229E are put into multi-alignment using Vector NTI[19]. The result indicated that much higher similarity existed in the ORF1ab region among all these coronaviruses except an especially high variable region near the 3' end of ORF1b, where the packaging signal of MHV and BCoV situated. It seems that the sequences of SARS-CoV, PEDV, and HCoV in this region might belong to another group different from MHV and BCoV. The exception of this region shows it might have higher rate of mutation or recombination than other parts of ORF1b. The finding of the packaging signals of MHV and BCoV in this region implies its responsibility for the packaging of coronaviruses genomic RNA (Fig 1).

Secondary structure of the RNA sequences around the putative packaging signal region About 1000 bp genomic sequence around the putative packaging signal region from all the five corovaviruses mentioned above were used to predict the RNA secondary structures by using RNA Structure 3.71[20], resulted that all the five sequences had a similar long stem-loop structure (Fig 2). The top four loops of the long stem-loop structure (marked with rectangles in Fig 2) are the mostly similar region among the coronaviruses as clearly shown in Fig 3. Fig 3A and 3B are the key packaging signals of MHV and BCoV. So it is reasonable to assume that the stem-loop region of SARS-CoV is also the genomic packaging signal (Fig 3B), similar results were found in other two viruses, HCoV and PEDV (Fig 3C and 3D). When the primary sequence corresponding to these stem-loop regions was mapped to the multi-alignment result, the hypothetical packaging signal of SARS-CoV was found to locate at the highly variable region of ORF1b, just before the packaging signals of MHV and BCoV. The primary sequences corresponding to the stem-loop structures of PEDV and HCoV 229E were also mapped to the multi-alignment result, which overlaps the putative packaging signal of SARS-CoV (Fig 1). This confirmed the hypothesis that the highly variable region of ORF1b was responsible for the genomic RNA packaging. Variety in primary sequences and the conservation of the secondary structures may be the most important trait of the genomic packaging signals of coronaviruses.

Multi-alignment of N protein and M protein Because the packaging signal should be first recognized and bound by the N protein to form a complex, which

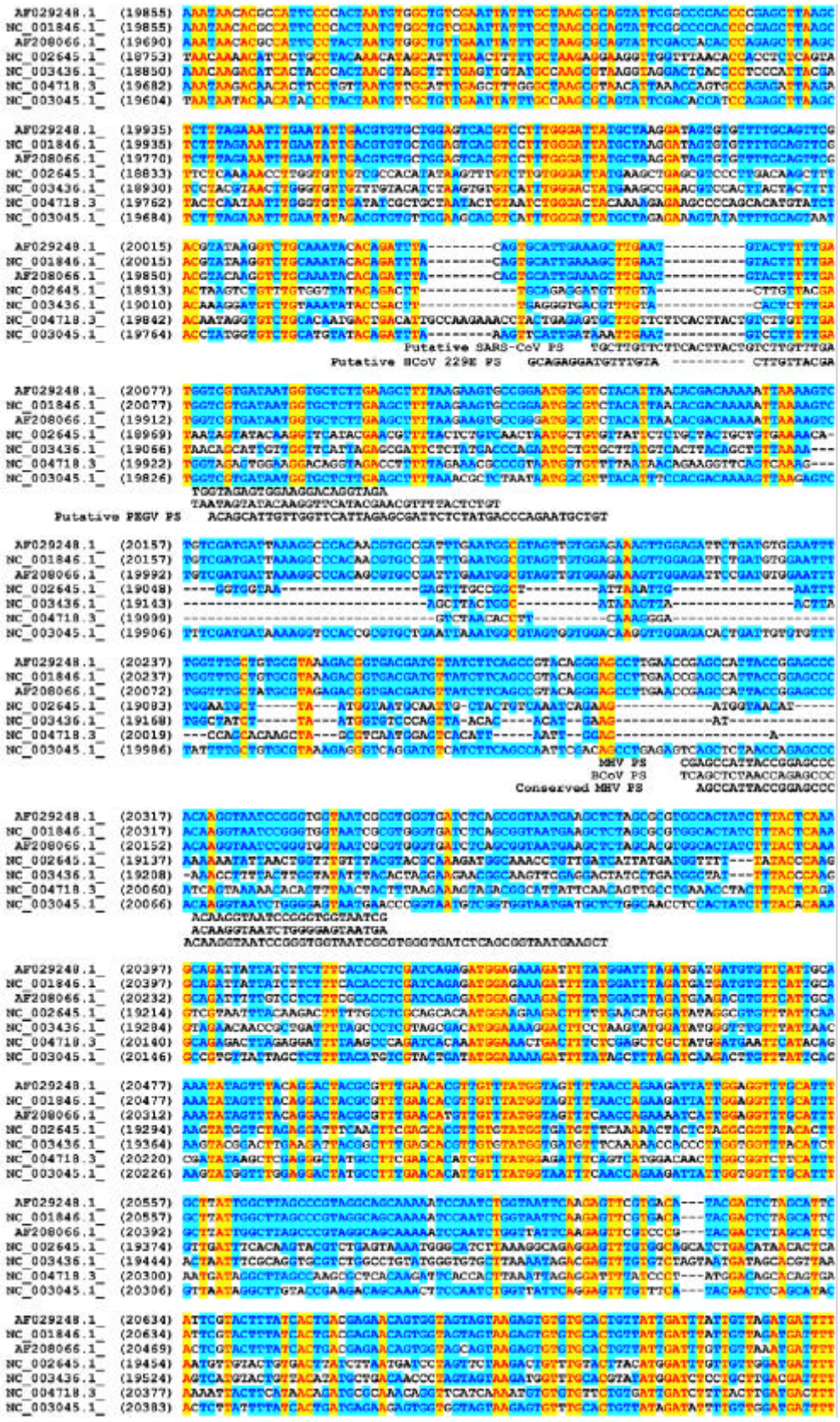

Fig 1. The multi-alignment result of genomic sequences. AF029248, NC_001846 and AF208066 are the genomic sequences of three MHV strains. NC_002745 is the genomic sequence of HCoV 229E. NC_003436 is the genomic sequence of PEDV. Nc_004718 is the genomic sequence of SARS-CoV. NC_003045 is the genomic sequence of BCoV. The packaging signals (PS) of all viruses are mapped in this figure, together with the conserved 69 bp packaging signal of MHV.

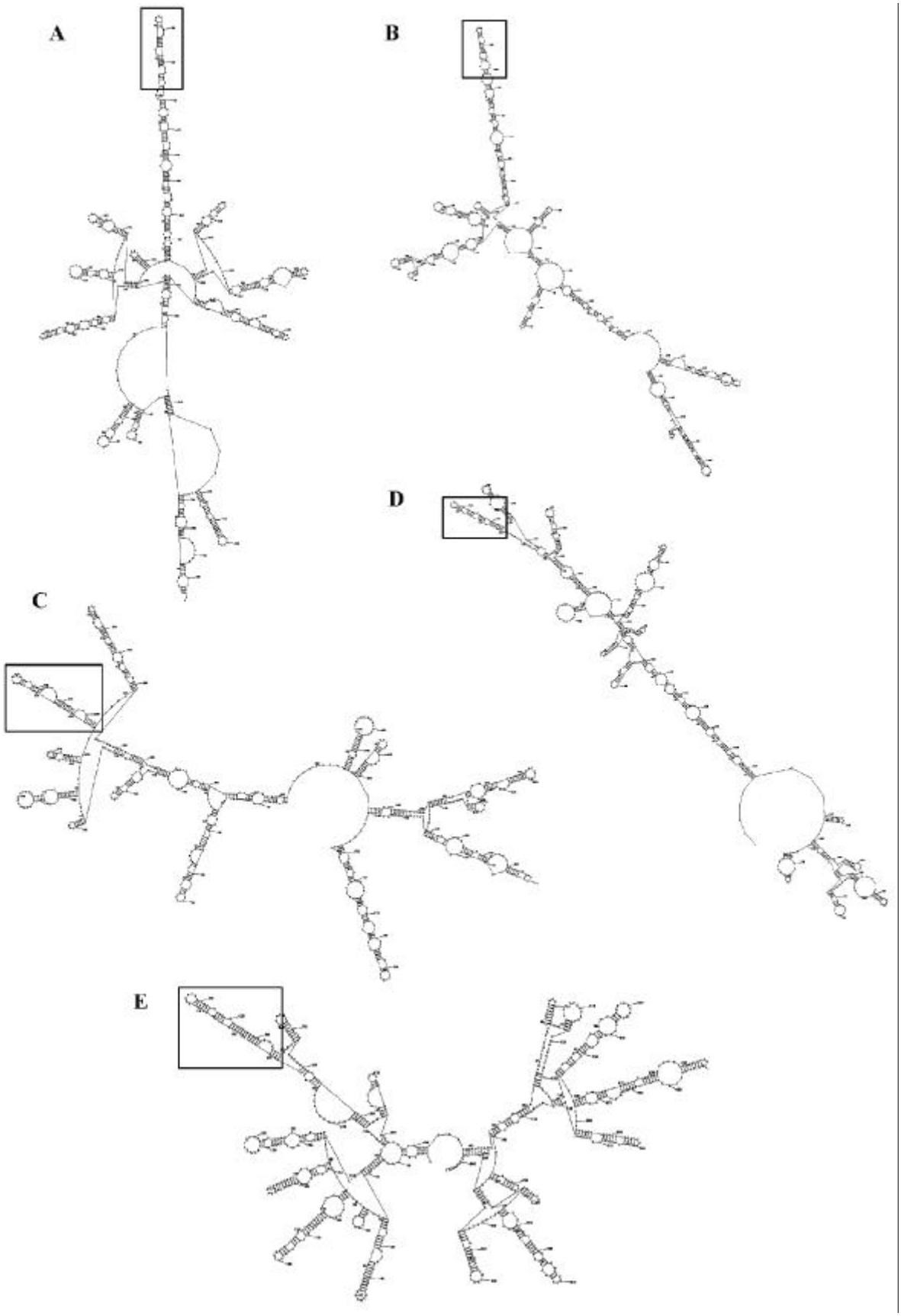

Fig 2. The RNA secondary structures modeled by using RNA structure 3.71. (A)-(E) are the RNA secondary structures of MHV, BCoV, HCoV 229E, PEDV, and SARS-CoV, respectively. The possible key packaging signals are indicated with rectangle on the top of the long stem-loop structures.

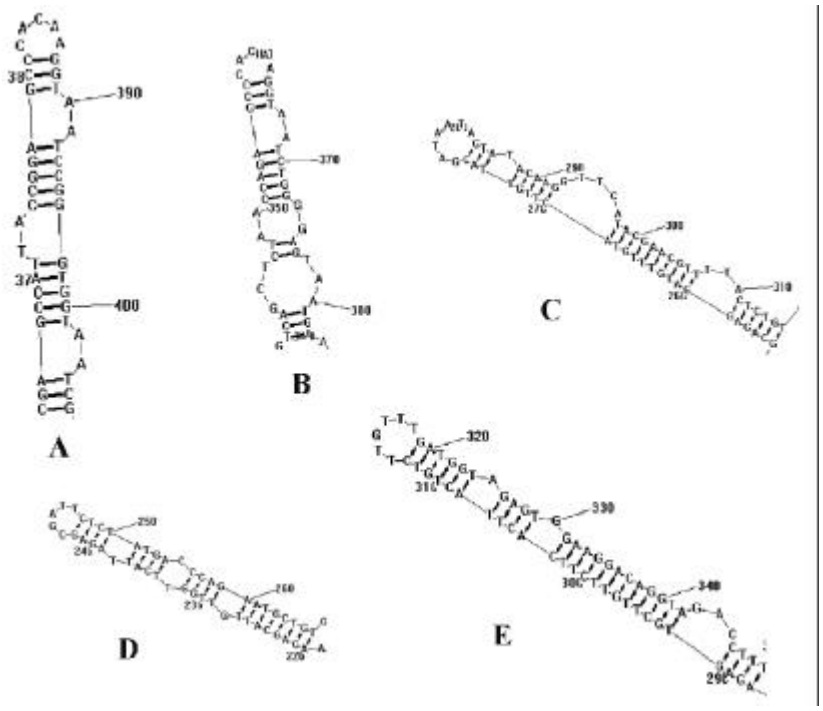

Fig 3. The secondary structures of the possible key packaging signals on the top of the long stem-loop structures (see Fig 2). (A)-(E) are the possible key packaging signals of MHV, BCoV, HCoV 229E, PEDV, and SARS-CoV viruses, respectively.

further interacts with the M protein[5,7], the packaging process can be expressed with a formula: [(Packaging signal ⟷ N protein) ⟷ M protein]. The relatively high mutation ratio in the primary sequences of the genomic packaging signal selectively presses to the N proteins directly, which may finally influence the primary sequences of the N and M proteins. In this study, multi-alignments were also performed on the N and M proteins of the SARS-CoV, MHV, BCoV, PEDV, and HCoV 229E (Fig 4). Muti-alignments indicated that the middle and C-terminal regions of the N proteins were active for mutation (Fig 4A), and the M proteins were more conservative than the N proteins. The mutation frequencies of the packaging signal, N protein, and M protein give an order of packaging signal > N protein > M protein. This relationship is in good agreement with the order of their mutual interactions, providing an additional proof that the packaging signal of SARS-CoV is situated in the highly variable region (Fig 1).

Antisense RNA and RNAi targeting the genomic packaging signal of SARS-CoV Antisense RNA technology has provided a prospective avenue for therapeutic applications since the middle of 1980s. Target mRNA transcripts may specifically interact with their complementary sequences of RNA, DNA, or chemically modified mucleic acids, thereby inhibiting their translation[23]. RNA interference (RNAi) techniques were developed based on the structure and function studies of nucleic acids. In several RNAi models, the designed double strand RNA (dsRNA) activates a normal cellular process, leading to a highly specific RNA degradation and, perhaps more importantly, making a cell-to-cell spreading for the gene silencing effect[24]. As a positive single strand RNA virus, SARS-CoV is especially appropriate to be used as target for therapeutic antisense RNAs and RNAis. For the essential role of the genomic packaging signal in virion assemble and replication, the hypothetic packaging signal sequence and sequences around it addressed in this study may serve as a good template to design antisense RNA or RNAi. The replication of SARS-CoV genome might be effectively restrained or destroyed without genomic packaging signals.

CONCLUSIONS

In conclusion, genomic packaging signals of coronaviruses enable the whole genome being packaged into virion correctly. Though it has been reported that some subgenomic RNA can also be packaged into virions, but its efficiency for subgenomic packaging is very low[5,7,25]. In this research, the primary sequence

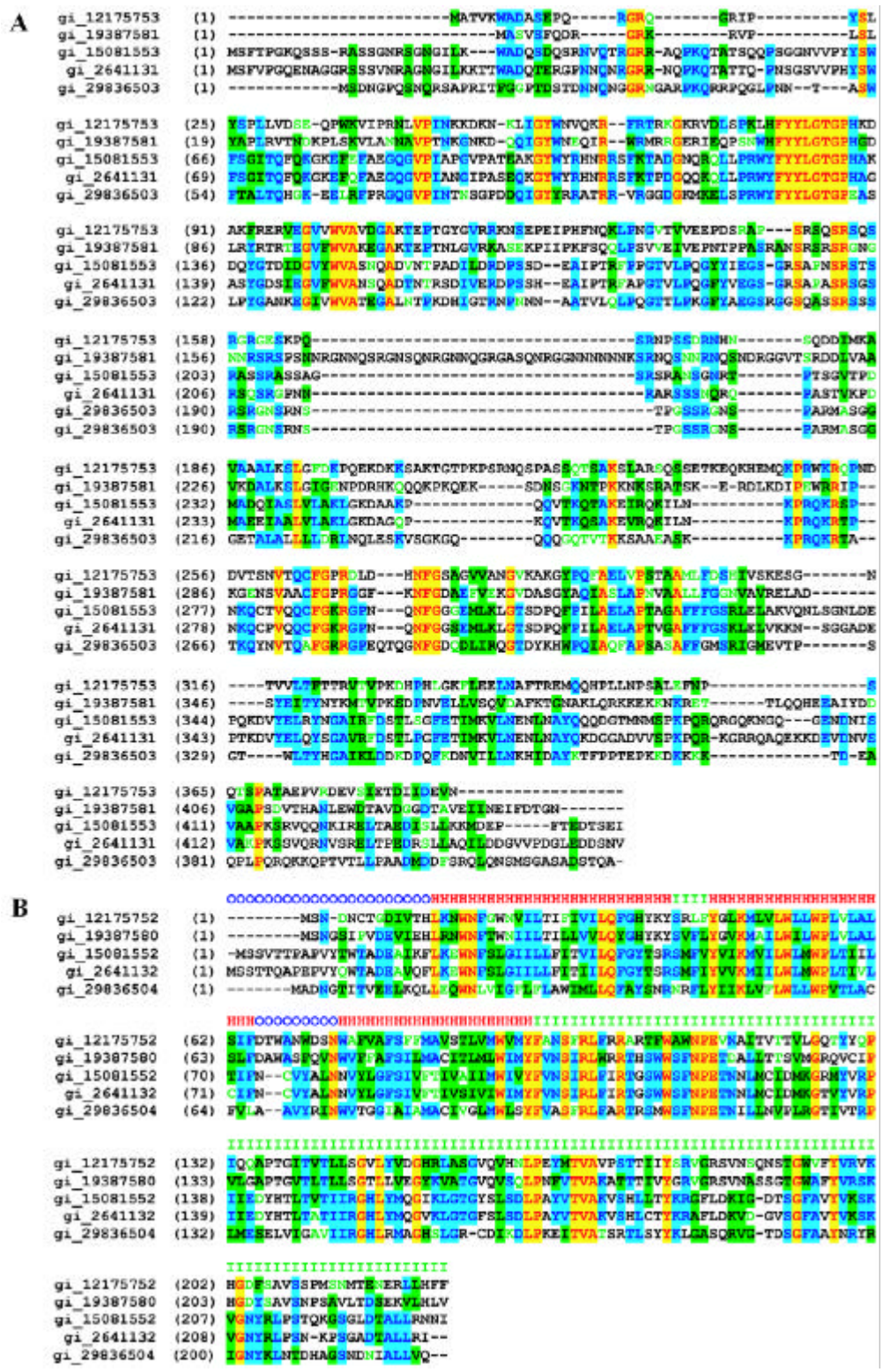

Fig 4. The multi-alignment result of N proteins and M proteins. (A) The multi-alignment result of the N proteins, gi_12175753 is the HCoV 229E N protein; gi_19387581 is the PEDV N protein; gi_15081553 is the BCoV N protein; gi_2641131 is the MHV N protein; gi_29836503 is the SARS-CoV N protein. (B) The multi-alignment result of the M proteins, gi_12175752 is the HCoV 229E M protein; gi_19387580 is the PEDV M protein; gi_15081552 is the BCoV M proetein; gi_2641132 is the MHV M protein; gi_29836504 is the SARS-CoV M protein. The transmembrane helices are indicated with red "H", the out membrane N terminal of the M protein was indicated with blue "O", and the C terminal cytoplasmic sequence is indicated with green "I".

and the secondary structure of the probable packaging signal region of SARS_CoV have been addressed employing various bioinformatics methods. For the regions where the packaging signals located, the primary sequences of SARS-CoV, HCoV, and PEDV show different profiles from that of MHV and BCoV. The secondary structure comparison pointed out that the possible genomic packaging signal of SARS-CoV, similar to MHV and BCoV, located at the top of a long stem-loop structure. Similar results were also obtained for the PEDV and HCoV. Similar to other coronaviruses, the genomic packaging signal of SARS_CoV is also characterized by high homology in secondary structure but high variability in primary sequences. The high mutation ratio in the region of genomic packaging signals selectively press the evolution of N and M proteins. The multi-alignments revealed that the N proteins of coronaviruses may also undergo rounds of mutations, but much weaker than the packaging signal region. Only subtle variations were detected in the M proteins of different coronaviruses. The possible reason is that the structure of the N protein-packaging signal complex is highly conserved. The key sequences of packaging signals proposed in this study provided useful clues in designing antisense RNAs or RNAi. Furthermore, finding analogs similar to the genomic packaging signal is another way for anti-SARS drugs.

REFERENCES

1 Peiris JS, Lai ST, Poon LL, Guan Y, Yam LY, Lim W, *et al*. Coronavirus as a possible cause of severe acute respiratory syndrome.Lancet 2003; 361(9366): 1319-25.

2 Ksiazek TG, Erdman D, Goldsmith CS, Zaki SR, Peret T, *et al*. A novel coronavirus associated with severe acute respiratory syndrome. N Engl J Med 2003; Apr 30 [epub ahead of print].

3 Ahlquist P. RNA-dependent RNA polymerases, viruses and RNA silencing. Science 2002; 196: 1270-3.

4 Johnson-Saliba M, Jans DA. Gene therapy: optimizing DNA delivery to the nucleus. Curr Drug Targets 2001; 2: 371-99.

5 Narayanan K, Maeda A, Maeda J, Makino S. Characterization of the coronavirus M protein and nucleocapsid interaction in infected cells. J Virol 2000; 74 : 8127-34.

6 Stephens EB, Compans RW. Assembly of animal viruses at cellular membranes. Annu Rev Microbiol 1988; 42:489-516.

7 Narayanan K, Makino S. Cooperation of an RNA packaging signal and a viral envelope protein in coronavirus RNA packaging. J Virol 2001; 75: 9059-67.

8 Adam MA, Miller AD. Identification of a signal in a murine retrovirus that is sufficient for packaging of nonretroviral RNA into virions. J Virol 1988; 62: 3802-6.

9 Beck, J, Bartos H, Nassal M. Experimental confirmation of a hepatitis B virus (HBV) epsilon-like bulge-and-loop structure in avian HBV RNA encapsidation signals. Virology 1997; 227: 500-4.

10 Qu F, Morris TJ. Encapsidation of turnip crinkle virus is defined by a specific packaging signal and RNA size. J Virol 1997; 71: 1428-35.

11 Woo K, Joo M, Narayanan K, Kim KH, Makino S. Murine coronavirus packaging signal confers packaging to nonviral RNA. J Virol 1997; 71: 824-7.

12 Makino S, Yokomori K, Lai MM. Analysis of efficiently packaged defective interfering RNAs of murine coronavirus: localization of a possible RNA-packaging signal. J Virol 1990; 64: 6045-53.

13 Van der Most RG, Bredenbeek PJ, Spaan WJ. A domain at the 3' end of the polymerase gene is essential for encapsidation of coronavirus defective interfering RNAs. J Virol 1991; 65: 3219-26.

14 Molenkamp R, Spaan WJ. Identification of a specific interaction between the coronavirus mouse hepatitis virus A59 nucleocapsid protein and packaging signal. Virology 1997; 239 (1): 78-86.

15 Bos EC, Dobbe JC, Luytjes W, Spaan WJ. A subgenomic mRNA transcript of the coronavirus mouse hepatitis virus strain A59 defective interfering (DI) RNA is packaged when it contains the DI packaging signal. J Virol 1997; 71 (7): 5684-7.

16 Fosmire JA, Hwang K, Makino S. Identification and characterization of a coronavirus packaging signal. J Virol 1992; 66 (6): 3522-30.

17 Cologna R, Hogue BG. Identification of a bovine coronavirus packaging signal. J Virol 2000; 74 (1): 580-3.

18 Lapps W, Hogue BG, Brian DA. Sequence analysis of the bovine coronavirus nucleocapsid and matrix protein genes. Virology 1987; 157: 47-57.

19 Rita Shiang. Vector NTI Suite v5.5.1. Biotech Software & Internet Report 2000; 1: 214-9.

20 Mathews DH, Sabina J, Zuker M, Turner DH. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J Mol Biol 1999; 288: 911-40.

21 http://www.mbb.ki.se/tmap/index.html

22 Raymond C, Jeannie FS, Brenda GH. Identification of nucleocapsid binding sites within coronavirus-defective genomes. Virology 2000; 277: 235-49.

23 Dean NM. Functional genomics and target validation approaches using antisense oligonucleotide technology. Curr Opin Biotechnol 2001; 12: 622-5.

24 Winston WM, Molodowitch C, Hunter CP. Systemic RNAi in C. elegans requires the putative transmembrane protein SID-1. Science 2002; 295: 2456-9.

25 Narayanan K, Chen CJ, Maeda J, Makino S. Nucleocapsid-independent specific viral RNA packaging via viral envelope protein and viral RNA signal. J Virol 2003; 77: 2922-7.