©Freund Publishing House Ltd. International Journal of Nonlinear Sciences and Numerical Simulation 6(2),181-186, 2005

Particular Symmetry in RNA Squence of SARS and the Origin of SARS Coronavirus

Xuan Xiao[1,2 *], Jin-Song Yao[3*], Shi-Huang Shao[1], Zheng-Jun Li[3], Yi-Sheng Zhu[3] and Zheng-De Huang[1]**

[1] *Bioinformatics Research Center, Donghua University, Shanghai 200051, China*

[2] *Computer Department , Jing-De-Zhen Ceramic Institute, Jing-De-Zhen 33300, China*

[3] *Department of Biomedical Engineering, Shanghai Jiaotong University, Shanghai 200030, China*

Abstract

Severe acute respiratory syndrome (SARS) belongs to coronavirus, however, it is dramatically different from all previously known coronaviruses. A peculiar character of RNA sequence is found in SARS, revealing particular symmetry in its sequencing. Comparison of symmetry between SARS and other coronaviruses shows heuristically that SARS coronavirus might come from the avian infectious bronchitis virus or porcine epidemic diarrhea virus.

Keywords: SARS coronavirus; recombination; particular symmetry

I. Introduction

The obtained entire genome sequence of SARS virus shows that SARS coronavirus (SARS-CoV) is sufficiently different from all previously known three groups of coronaviruses [1-2]. Comparison between the predicted amino acid sequences for three well-defined enzymatic proteins and the four major structural proteins of SARS-CoV with those from representative coronaviruses illustrates that SARS-CoV forms a distinct group within the genus of coronaviruses[3]. Marra *et al.* also obtained a similar result based on the analysis of different SARS-CoV isolates [4].

Till now, the origin and evolutionary history of the SARS-CoV remain unclear, and all analyses were based on gene sequence homologous alignment. Tuen *et al.* assumed that the SARS-CoV was evolved from virus relative innocuousness or causing slight symptom to the human, but several mutants that changed virus tropism happened in some animal carrier (e. g. palm civets) and made this virus become deadly to the human [5]. Because highly frequency homologous recombination often takes place when the coronaviruses were replicated [6-7]. In addition, Guan *et al.* found that the virus sequence is homology with SARS-CoV on wild animal up to 99.8% [8]. Many researchers suggested that SARS-CoV has a different coronaviruses recombinant history. For example, Stavrinides *et al.* reported that SARS virus is in fact a mosaic of mammalian and avian-like viruses and the recombination between the parents viruses may have occurred in the host-determining S gene [9]. Zhang *et al.* employed 7 recombination detection technique and conducted phylogenetic analysis, and found that 7 putative recombination regions exist between SARS and other 6 coronaviruses:

* *Xuan Xiao and Jin-Song Yao Contribute equally to this work*

** *Corresponding author Bioinformatics Research Center, Donghua University, Shanghai 200051, China*
Email: shshao@dhu.edu.cn

porcine epidemic diarrhea virus (PEDV), transmissible gastroenteritis virus (TGEV), bovine coronavirus (BCoV), human coronavirus 229E (HCoV), murine hepatitis virus (MHV), and avian infectious bronchitis virus (IBV) [10]. There were also claims that the SARS virus was not a host range mutant of any previously described coronaviruses due to its low sequence identity to known coronaviruses [11-12]. Holmes *et al.* pointed out that the phylogenetic patterns cited as evidence for recombination are more probably caused by a variation in substitution rate among lineages, the recombination can not explain the origin of SARS-CoV[13].

It is not suitable for constructing the phylogenetic trees using sequence alignment when alignment regions are characterized by low consistence or variable length [14-15]. The comparability between SARS-CoV and other coronaviruses is low, it is necessary to use a method to investigate the origin of the SARS-CoV. In this paper, the visual method of particular symmetry is used to analyze all the full-length RNA genomes of known coronavirus strains and we find a new sequence characteristic of SARS-CoV.

2. Particular Symmetry

Analyzing the published 153 SARS-CoVs and other 24 coronavirus (all downloaded from National center for biotechnology information) with our method introduced in the following section, we discover a special characteristic of SARS-CoV. From about 3232 to 5624nt, 5703 to 7195nt, 12128 to 14470nt, 16444 to 19231nt, 19728 to 21803nt in the SARS-CoV genome sequences near 5-terminal, the number of Adenine (A) is almost equal to the number of thymine (T) in the above five sections, and the A are mostly mastered in the 5'-terminal of the segment, T are mostly in the 3'-terminal region. Because A and T are complementary pair in double-helix structure, this kind of characteristic is named as *particular symmetry*. For all the other coronaviruses, this characteristic doesn't exist in the same regions and the number of T is obviously larger than that of A. Only in the PEDV and IBV genome sequences, there exist the similar distribution of A and T in the region of 3232 to 5624nt. The ratio of T/A in the entire SARS-CoV genome sequence is also close to that in the PEDV and IBV.

3. Method

DNA sequencing is a procedure of nature selection, J.H. He [16] first endows a quaternary digit (0, 1, 2, 3) for adenine(A), cytosine(C), guanine(G), and thymine(T), respectively, see Tab.1. J.H. He [16] also endows A, C, T, and G with one of the following numbers: 00, 01, 10, and 11(See Tab.2). For example

11010100101010110101010111101

might imply an DNA sequence for a special genome.

Tab.1 Possible quaternary values for A,C,T, and G.[16]

A	C	T	G
0	1	2	3
1	2	3	0
2	3	0	1
3	0	1	2

Tab.2 Possible binary values for A,C,T, and G[16].

A	C	T	G
00	01	10	11
01	10	11	00
10	11	00	01
11	00	01	10

The number of A-T partnership can decide the degree of stability of nucleotide chain; it can also be used to distinguish each species, because the genome sequence varies with the species evolvement. The more closely phylogenetic relationship between two species, the more similar content proportion of AT in genome sequences they should be.

In this paper, the visualization method of AT content is presented. According to this method, a new sequence characteristic of SARS-CoV is found. Our method circumambulates the difficulty of sequence alignment, and avoids any

bias that may be associated with particular genomic regions. A nucleotide sequence is coded as follows:

$$A=-1, \qquad C=0, \qquad G=0, \qquad T=1 \quad (1)$$

Through the above encoding procedure, a gene sequence is transformed to a serial of digital signals. For example, the sequence of "AATGCTGG" can be coded as a discrete digital sequence P ="−1,−1,1,0,0,1,0,0".

The second step is to use the function of sum:

$$S(p,j)=\sum_{i=1}^{i=j} p_i \quad j=1,2,\cdots,n \qquad (2)$$

where p_i is the value for the i^{th} sequence, n is the length of nucleotide chain.

Fig.1 shows the relationship between S and n for several kinds of coronavirus curves. It is obvious that the curve of SARS-CoV is different from those of other groups of coronavirus most distinctively.

4. Discussion and Conclusion

Utilizing the visualization method of AT content mentioned above, we analyze the different parts of these full-length sequence curves in all known 153 SARS-CoV and other 24 coronavirus isolates obtained from the Genbank. Five segments in SARS-CoV curve are relatively flat, other coronavirus curves are almost upwards all the time. According to the figure 1, the SARS-CoV particular symmetry was thus obtained. From about 3232 to 5624nt, 5703 to 7195nt, 12128 to 14470nt, 16444 to 19231nt, 19728 to 21803nt in the SARS-CoV sequence near 5-terminal, the number of A is almost equal to the number of T, the average ratio of T/A is 1.002, 1.005, 1.007, 1.004, 1.001 respectively. In figure 1, these five segments of SARS-CoV curves are the upward concave shape. This indicates that A is rich in the 5'terminal part of the segment because the curve is downward in the front part, and T is rich in latter part because curve is upward. But the number of T is greater than that of A in the same 5 segments of other coronaviruses. The ratio of T/A is mostly in 1.2 nearby, the average is 1.256, 1.300,1.198,1.194,1.221 respectively, other coronaviruses have not the character of SARS-CoV. The statistical average of particular symmetry of all coronaviruses in five segments of sequences are showed in Tab.3. It should be emphasized that there also exist A=T in other areas of the SARS-CoV curve, but these regional length do not exceed 1200nt or do not satisfy the condition that A mostly exist in front part and T mostly in rear part.

Among coronavirus, the PEDV and IBV have the ratio of T/A 1.026, 0.994 respectively in the first interval from 3232nt to 5624nt, it is obviously that these data are closely to the SARS-CoV ratio of T/A in the same region. PEDV also have the closely ratio of T/A 1.096 in the fourth interval between 16444 and 19231nt. From 2408nt to 5794nt in the IBV sequence and 3223 to 6160nt in the PEDV sequence near 5-terminal, the number of A is almost equal to the number of T, and A almost exist in the former part, T in latter part. The IBV and PEDV have the similar particular symmetry. Other coronaviruses have not the above character; it is suggested that SARS-CoV is closer to IBV and PEDV. This result is consistent with other people's reports. For example, Stavrinides and Guttman[9] used Bayesian, neighbor-joining, and split decomposition phylogenetic technique to the SARS replicase, surface spike, matrix and nucleocapsid proteins, and revealed the origin of SARS. The analyses support an avian-like origin for the matrix and nucleocapsid proteins, and a mammalian-avian mosaic origin for the host-determining spike protein. Qi *et al.* compared the entire sequence of 12 SARS-CoV with 12 other coronavirused based on the method of function of degree of disagreement and suggested that SARS-CoV was closely relation with the group 1 of coronavirus[17].

Symmetry is an essential character in DNA. In a double-helix DNA strands, there are A=T and G=C, and also have the phenomena of $A \approx T$ and $G \approx C$ inside single strand [18,19]. Seuoka suggested a hypothesis to explain the symmetrical relation: if select pressure and natural mutation of the double strands were equal, there would appear the phenomena of

A≈T and G≈C in single strand after a long time evolvement [21]. It is showed that the appearing symmetry is the direction of evolvement, and the symmetry phenomenon of A≈T and G≈C exists in the most organism have finished sequence, and the longer the sequence is, the higher the precision is. The ratios of A/T and C/G fluctuate between 0.999 and 1.001 in human each chromosome.

It is clear that the SARS-Coves owns more particular symmetry of A ≈ T than other coronaviruses from figure 1, and the ratio of T/A in SARS-CoV entire genome sequence also close to 1 than that of other coronaviruses, above all shows that SARS-CoV evolve from other coronaviruses, and it is most possible that SARS-CoV is closely related to the PEDV and IBV.

We should emphasize that symmetry is the natural character in particle world[20], and in biology [22,23,24] as well. We will further study the particular symmetry in SARS in the future, emphasizing on more mysterious characters of SARS.

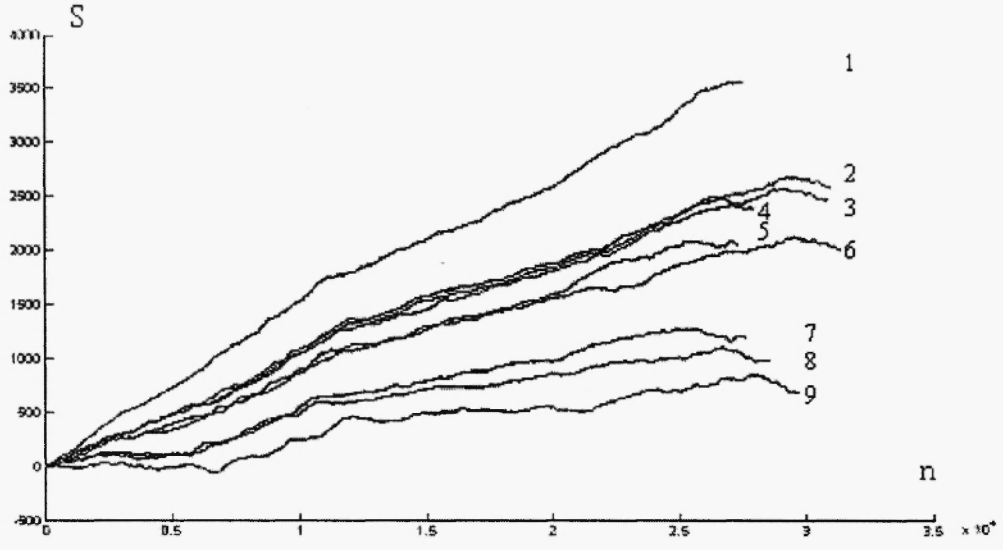

Figure 1. The coronavirus's curves based on visualization method of AT content. The curve 1 is the human coronavirus NL63, 2 the bovines coronavirus, 3 human coronavirus OC43, 4 the porcine epidemic diarrhea virus, 5 human coronavirus 229E, 6 murine hepatitis virus, 7 avian infectious bronchitis virus, 8 transmissible gastroenteritis virus, 9 SARS coronavirus TW1.

Tab.3. Statistical average data of all coronaviruses particular symmetry in five segments of sequences.

coronavirus	3232bp-5624bp			5703bp-7195bp			12128bp-14470bp			16444bp-19231bp			19728bp-21803bp		
	T	A	T/A	T	A	T/A	T	A	T/A	T	A	T/A	T	A	T/A
SARS-CoV	708	706	1.002	467	465	1.005	687	683	1.007	823	820	1.004	638	637	1.001
No SARS-CoV	805	647	1.256	533	413	1.300	783	654	1.198	924	775	1.194	702	575	1.221

References

1. Anand, K., Ziebuhr, J., Wadhwani, P., Mesters, J. R., Hilgenfeld, R., (2003) Coronavirus main proteinase ($3CL^{pro}$) structure: basis for design of anti-SARS drugs. *Science*, 300: 1763–1767
2. Eickmann, M., Becker, S., Klenk, H. D., Doerr, H. W., Stadler, K., Censini, S., Guidotti, S., Masignani, V., Scarselli, M., Mora, M., Donati, C., Han, J. H., Song, H. C., Abrignani, S., Covacci, A., Rappuoli, R., (2003) Phylogeny of the SARS coronavirus. *Science*. 302(5650):1504-1505.
3. Rota, P.A., Oberste, M.S., Monroe, S.S., et al. (2003) Characterization of a novel eoronavirus associated with severs acute respiratory syndrome. *Science*, 300(5624): 1394-1399.
4. Marra, M.A., Jones, S.J., Astell, C.R., et al., (2003) The genome sequence of the SARS-associated coronavirus. *Science,* 300(5624): 1399-1404.
5. Tuen, W. N., Gabriel, T., Antoine, D., (2003) A double epidemic model for the SARS propagation. *BMC Infect Dis* 3: 19
6. Lai, M.M.C., Holmes, K.V.,(2001). Coronaviridac: The viruses and their replication. In Kripe DM, Howley PM, eds. 4th ed. *Fields Virology.* New York:Lippincott Wilkins,1163-1186
7. Sawicki, S.G., Sawicki, D.L. (1998) A new model for coronavirus transcription. Adv Exp Med Biol 440:215-219.
8. Guan, Y., Zheng, B. J., He, Y. Q., Liu, X. L., Zhuang, Z. X., Cheung, C. L., Luo, S. W., Li, P. H., Zhang, L. J., Guan, Y. J., Butt, K. M., Wong, K. L., Chan, K. W., Lim, W., Shortridge, K. F., Yuen, K. Y., Peiris, J. S. M., Poon, L. L. M., (2003) Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. *Science* 302(5643): 276–278
9. Stavrinides, J., Guttman, D.S. (2004) Mosaic evolution of the severe acute respiratory syndrome coronavirus. *J. Virol* 78(1): 76–82
10. Zhang, X.W., Yap, Y.L., Danchin, A,(2004) Testing the hypothesis of a recombinant origin of the SARS-associated coronavirus. *Arch Virol.*2004 Oct 11[Epub ahead of print]
11. Holmes, K. V. (2003). SARS-associated coronavirus. *N. Engl. J. Med.* 348:1948-1951.
12. Ruan, Y. J., Wei, C. L., Ee, L. A., Vega, V. B., Thoreau, H., Yun, S. T. S., Chia, J. M., Ng, P., Chiu, K. P., Lim, L., Tao, Z., Peng, C. K., , Ean, L. L. O., Lee, N. M., Sin, L. Y., Ng, L. F. P., Chee, R. E., Stanton, L. W., Long, P. M. and Liu. E. T. (2003). Comparative full-length genome sequence analysis of 14 SARS coronavirus isolates and common mutations associated with putative origins of infection. *Lancet* 361:1779-1785.
13. Holmes, E.C., Rambaut, A., (2004) Viral evolution and the emergence of SARS coronavirus.*Phil. Trans. R. Soc.Lond.* 359: 1059-1065.
14. John, G., Rob, D., Ward, W., (1993) Alignment-ambiguous nucleotide sites and the exclusion of systematic data. *Mol Phylogenet Evol*, 2: 152-157.
15. Ward, W., John, G., Rob, D..(1995) Elision: A method for accommodating multiple molecular sequence alignments with alignment-ambiguous sites. *Mol Phylogenet Evol,* 4:1-9.
16. He,J.H. (2004) Mysterious Pi and a Possible Link to DNA Sequencing, International Journal of Nonlinear Sciences and Numerical Simulation,5(3), 263-274
17. Qi, Z., Hu, Y., Li, W., Chen, Y. J., et al., (2003) Phylogeny of SARS-CoV as inferred from complete genome comparison. *Chinese Science Bulletion,* 48(12):1175-1178.
18. Lobry, J.R., Lobry, C.,(1999) Evolution of DNA composition under no-strand-bias conditions when the substitution rates are not constant. *Mol Biol Evol.* 16(6):719-723.
19. Rocha, E.P., Danchin, A., (2001) Ongoing evolution of strand composition in bacterial genomes. *Mol Biol Evol.* 18(9):1789-1799.
20. El Naschie,M.S. (2004) Transfinite Electrical Networks, Spinoral Varieties and Gravity Q Bits, International Journal of Nonlinear Sciences and Numerical Simulation, 5(3), 191-198

21. Sueoka, N., (1995)Intrastrand parity rules of DNA base composition and usage biases of synonymous codons. *J.Mol. Evol.*40(3):318-325.
22. Kuikka,J.T. Fractal analysis of medical imaging, *International Journal of Nonlinear Sciences and Numerical Simulation*, 3 (2002), 81-88
23. Kuikka,J.T. Scaling laws in physiology: Relationships between size, function, meta-bolism and life expectancy, *International Journal of Nonlinear Sciences and Numerical Simulation*, 4 (2003), 317-328
24. He,J.H.,Chen,H. Effects of size and pH on metabolic rate, *International Journal of Nonlinear Sciences and Numerical Simulation*, 4 (2003), 429-432